UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ADAPTIVE KALMAN BASED FORECASTING FOR ELECTRIC LOAD AND DISTRIBUTED GENERATION

LUCAS DANTAS XAVIER RIBEIRO

ORIENTADOR: JOÃO PAULO CARVALHO LUSTOSA DA COSTA

DISSERTAÇÃO DE MESTRADO EM
ENGENHARIA ELÉTRICA

PUBLICAÇÃO: 659/2017 DM PPGEE

BRASÍLIA/DF: ABRIL - 2017.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ADAPTIVE KALMAN BASED FORECASTING FOR ELECTRIC LOAD AND DISTRIBUTED GENERATION

LUCAS DANTAS XAVIER RIBEIRO

DISSERTAÇÃO DE MESTRADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

APROVADA POR:

João Paulo Carvalho Lustosa da Costa

Prof. João Paulo Carvalho Lustosa da Costa, Dr.-Ing. (ENE-UnB, Fraunhofer IIS e TU Ilmenau)
(Orientador)

Prof. Rafael Timóteo de Sousa Júnior, Dr. (ENE-UnB)
(Examinador Interno)

Wesley Fernando Usida

Wesley Fernando Usida, Dr. (ANEEL)
(Examinador Externo)

BRASÍLIA/DF, 7 DE ABRIL DE 2017.

FICHA CATALOGRÁFICA

RIBEIRO, LUCAS DANTAS XAVIER
Adaptive Kalman based Forecasting for electric load
and distributed generation. [Distrito Federal] 2017.
xvii, 290p., 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2017)
Dissertação de Mestrado - Universidade de Brasília.
Faculdade de Tecnologia.
Departamento de Engenharia Elétrica.

1. Kalman Filter 2. Principal Component Analysis
3. Load Forecasting 4. Distributed Energy Resources
I. ENC/FT/UnB II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

RIBEIRO., L. D. X. (2017). Adaptive Kalman based Forecasting for electric load and distributed generation. Dissertação de Mestrado em Engenharia Elétrica, Publicação 659/2017 DM PPGEE, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 290p.

CESSÃO DE DIREITOS

NOME DO AUTOR: Lucas Dantas Xavier Ribeiro.

TÍTULO DA DISSERTAÇÃO DE MESTRADO:Adaptive Kalman based Forecasting for electric load and distributed generation.

GRAU / ANO: Mestre / 2017

Lucas Dantas Xavier Ribeiro

Lucas Dantas Xavier Ribeiro
SGAN 911 Módulo F, Bloco H, Apartamento 12, Asa Norte
70790-110 Brasília - DF - Brasil.

DEDICATÓRIA

Alis volat propriis.
Per aspera, ad astra.

AGRADECIMENTOS

Primeiramente à Deus

À minha família por todo o cuidado, carinho e pela educação que me foi dada. Sem vocês jamais teria chegado neste ponto. Agradeço à minha mãe por toda a dedicação e por ser meu maior exemplo de vida. Ao meu pai por ter me ensinado muitos dos valores que me definem como homem. À minha irmã pela amizade e carinho por todos estes anos. À minha querida tia Jailma por ser a melhor tia do mundo. À minha avó Nina por ser uma segunda mãe para mim.

Agradeço ao meu orientador, Prof. João Paulo, por todo o auxílio dispensado na orientação deste trabalho, em particular pela ajuda e estímulo às publicações. Agradeço ao meu amigo Jayme Milanezi Júnior por ser a pessoa que mais me auxiliou nesta atividade de pesquisa. Também agradeço a meus amigos Márcio Venício Pilar de Alcântara e Elton Mário de Lima por terem me auxiliado com opiniões, artigos e com o empréstimo de livros didáticos. Todos vocês contribuíram para a realização desta dissertação.

Aos meus amigos da Agência Nacional de Energia Elétrica (ANEEL), em particular á equipe da Superintendência de Pesquisa e Desenvolvimento e Eficiência Energética (SPE). Agradeço à agência por ter me concedido licença para cursar disciplinas no horário da manhã. Ao Superintendente, Máximo Luiz Pompermayer agradeço a confiança em mim depositada. Igualmente, agradeço a Aurélio Calheiros de Melo Júnior, a Fábio Stacke Silva, a Carmen Silvia Sanches e a Sheyla Maria das Neves Damasceno. Amigos que sempre me ajudaram a conciliar o trabalho com este curso de pós-graduação. Aos demais amigos da SPE e da ANEEL, o meu agradecimento pela amizade e companheirismo, graças a vocês dou muito valor ao nosso ambiente de trabalho.

Agradeço ao Prof. José Luiz Cerda-Arias e ao colega de pós-graduação Marcos Vinícius Vieira, que foram responsáveis pela obtenção dos dados de demanda elétrica de Leipzig e de Brasília, respectivamente. Ao Prof. Edison Pignaton pelas sugestões e pela atenção dispensada em uma fase crucial deste trabalho. Ao colega Ricardo Kehrle Miranda pela ajuda e parceria dispensada.

RESUMO

PREVISÃO ADAPTATIVA DE CARGA E DE GERAÇÃO DISTRIBUÍDA BASEADA EM FILTROS DE KALMAN

Autor: Lucas Dantas Xavier Ribeiro
Orientador: João Paulo Carvalho Lustosa da Costa
Programa de Pós-graduação em Engenharia Elétrica
Brasília, abril de 2017

O desenvolvimento econômico está relacionado à disponibilidade de energia elétrica, especialmente em virtude da dependência quase total que a maioria das indústrias e dos serviços essenciais têm de seu uso. A disponibilidade de energia perene, barata e confiável é de primordial importância econômica.

Dado que o conjunto de requerimentos encontrados pelas companhias de distribuição constitui um cenário complexo, ferramentas robustas de previsão de demanda são necessárias para implementar planos de expansão e operações eficientes e razoáveis.

A inserção de geração distribuída adiciona um novo nível de complexidade a esta tarefa, pois não somente a geração descentralizada diminui a carga de modo aleatório e intermitente, como também inevitavelmente produz alterações nas séries históricas de carga usadas para fazer as previsões. Ambos os efeitos agem no sentido de aumentar os erros de predição no curto e no longo prazo, ameaçando a eficiência operacional e, no pior caso, a estabilidade do sistema.

Este trabalho apresenta a previsão de carga e geração como um problema de estimação dinâmica de estado via filtros adaptativos de Kalman. As variáveis a serem estimadas são das demandas de base, média e de pico, assim como a geração fotovoltaica. Como medições e observações, são utilizadas previsões de tempo, datas e eventos de calendário, tarifas de eletricidade, índices e estimativas econômicas e demográficas. Combinações preprocessadas destas medições são usadas como as variáveis de entrada para a previsão.

A metodologia proposta foi comparada com outras técnicas do estado da arte, sendo os desempenhos avaliados com base nos critérios de Erro Médio Quadrático (MSE), Raiz do Erro Médio Quadrático (RMSE), Coeficiente de correlação, Erro Médio Percentual (MAPE), Erro Médio Absoluto (MAE), Erro Médio de Tendência (MBE), Erro Máximo Absoluto (MXE) e Erro Máximo Percentual (MPE). Na maioria dos cenários analisados, o sistema de predição adaptativo proposto superou as técnicas de referência baseadas em redes neurais e espaço de estados.

ABSTRACT

ADAPTIVE KALMAN BASED FORECASTING FOR ELECTRIC LOAD AND DISTRIBUTED GENERATION

Author: Lucas Dantas Xavier Ribeiro
Supervisor: João Paulo Carvalho Lustosa da Costa
Programa de Pós-graduação em Engenharia Elétrica
Brasília, abril de 2017

Economic development is related to the availability of electricity, especially because most industries and basic services depend almost entirely on its use. The availability of a source of continuous, cheap, and reliable energy is of foremost economic importance.

Since the set of requirements faced by power distribution utilities assemble a complex scenario, robust load forecasting tools are needed to implement efficient and reasonable expansion and operation plans.

The introduction of distributed generation adds a new level of complexity to this task, as not only the decentralized generation reduces load in a random and intermittent way, but also inevitably embeds in the historic loads used to forecast. Both effects act to increase prediction errors in short and long term, jeopardizing operational efficiency and, in worst case, system reliability.

This work presents the load and generation forecasting as a dynamic state estimation problem by means of Kalman adaptive filters. The variables to be estimated are daily base, average and peak electric load, as well as PV generation. As measurements and observations, this work uses weather forecasts, calendar dates and events, energy tariffs, economical and demographic indexes and estimatives. Preprocessed combinations of these measurements are the input variables employed for forecasting.

The proposed methodology is compared with other state-of-art techniques, the performances evaluated with base in error performance criteria such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Correlation coefficient, Mean Average Percentual Error (MAPE), Mean Absolute Error (MAE), Mean Bias Error (MAE), Maximum Absolute Error (MXE) and Maximum Percentual Error (MPE). In most evaluated scenarios, the proposed adaptive prediction system outperforms the benchmark techniques, based on state space and neural networks.

TABLE OF CONTENTS

1 INTRODUÇÃO 1

1.1 Formulação do problema 8

1.2 Organização da dissertação 11

2 INTRODUCTION 1

2.1 Problem formulation 8

2.2 Organization of this dissertation 10

3 LOAD AND GENERATION MODELING 12

3.1 Overview 12

3.2 Generalized Additive Models, Neural Networks and Linear Regression . 14

3.3 Electric load dynamics 21

3.4 Factors influencing electric load 23

3.4.1 Weather Variables 24

3.4.2 Socioeconomic variables 47

3.4.3 Electricity tariffs 49

3.4.4 Calendar and Weather events 51

3.4.5 The Load Model 53

3.5 Photovoltaic Generation Model 58

3.5.1 The Solar irradiation model 62

3.5.2 PV panel model 67

3.5.3 State space representation 71

4 LOAD AND GENERATION FORECASTING 74

4.1 Load Forecasting 74

4.1.1 Preprocessing and Feature selection 78

4.1.2 Grey model forecasting 79

4.1.3 Proposed Kalman based adaptive prediction scheme 82

4.2 Photovoltaic Generation Forecasting 91

4.2.1 Grey box model for PV 93

4.2.2 Proposed Kalman based adaptive PV prediction scheme 94

5 RESULTS 96

5.1 Electric load forecasting . . . 96

5.1.1 First forecasting scenario - Leipzig 2001-2003 . . . 97

5.1.2 Second load forecasting scenario - Brasilia 2001-2003 . . . 138

5.1.3 Third load forecasting scenario - Brasilia 2004-2010 . . . 144

5.2 Photovoltaic generation forecasting . . . 150

5.2.1 Site A1 - Oss region, Netherlands . . . 153

5.2.2 Site A2 - Queensland, Australia . . . 161

5.2.3 Site A3 - South Australia, Australia . . . 165

5.2.4 Site A4 - Utrecht region, Netherlands . . . 169

5.2.5 Site A5 - Amsterdam region, Netherlands . . . 172

5.2.6 Site A6 - Apeldoorn region, Netherlands . . . 175

5.2.7 Site A7 - California, USA . . . 180

6 CONCLUSIONS 184

6.1 General conclusions . . . 184

6.2 Load forecasting conclusions . . . 186

6.3 PV Generation forecasting conclusions . . . 188

6.4 Directions for future research . . . 189

BIBLIOGRAPHY 191

APPENDICES 201

A The Kalman filter 202

A.1 Overview . . . 202

A.2 State space representation . . . 203

A.2.1 Obtaining a discrete state space representation from a difference equation . . . 206

A.3 Filter derivation . . . 208

A.4 Variance tracking . . . 213

B The SPCTRL2 Radiative Transfer Model and SEDES2 Cloud Cover Modifier 215

B.1 Introduction . . . 215

B.2 Key concepts . . . 221

B.2.1 Local solar position . . . 225

B.3 Direct Normal Irradiance . . . 227

B.3.1 Rayleigh Scattering . . . 229

B.3.2 Aerosol Scattering and Absorption 229
B.3.3 Water Vapor, Ozone and Uniformly Mixed Gas Absorption . . . 230
B.4 Diffuse Irradiance 231
B.4.1 Rayleigh scattering term 232
B.4.2 Aerosol scattering term 232
B.4.3 Ground and sky reflectance term 233
B.5 Global irradiance on tilted surfaces 235
B.6 Cloud cover modifiers 235

C Principal Component Analysis 240
C.1 Data standardization 240
C.2 Singular Value Decomposition 241
C.3 Projection Matrix 243

D Artificial Neural Networks 245
D.1 Multilayer perceptron 246
D.2 Supervised learning 247
D.3 Backpropagation algorithm 248

LIST OF TABLES

1.1 Suprimento de Eletricidade Doméstico, PIB e população de 2000 a 2015 - Mundo e Alemanha 2

2.1 Worldwide and Domestic German Energy Supply (DES), German Gross Domestic Product (GDP) and Population from 2000 to 2015 2

3.1 Common link functions of the Exponential family 17
3.2 List of METAR's measurements and observations 26
3.3 Approximate calculation of the heating degree-days 31
3.4 Approximate calculation of the cooling degree-days 32
3.5 Approximate calculation of the heating degree-days 37
3.6 Photometric and corresponding radiometric units 44
3.7 List of Socioeconomic variables obtained per Case study 48
3.8 List of event variables 52
3.9 Inputs and outputs per PV forecasting step 62

4.1 Listing of the candidate input variables 77
4.2 Inputs and outputs per predicting phase 84

5.1 Error metrics for Base load, Substation S1 101
5.2 Error metrics for Average load, Substation S1 102
5.3 Error metrics for Average load, Substation S1 (continuation) 103
5.4 Error metrics for Peak load, Substation S1 104
5.5 Error metrics for Base load, Substation S2 106
5.6 Error metrics for Average load, Substation S2 107
5.7 Error metrics for Peak load, Substation S2 108
5.8 Error metrics for Peak load, Substation S2 (continuation) 109
5.9 Error metrics for Base load, Substation S3 110
5.10 Error metrics for Average load, Substation S3 111
5.11 Error metrics for Peak load, Substation S3 112
5.12 Error metrics for Peak load, Substation S3 (continuation) 113
5.13 Error metrics for Base load, Substation S4 114

5.14 Error metrics for Average load, Substation S4 115
5.15 Error metrics for Peak load, Substation S4 116
5.16 Error metrics for Peak load, Substation S4 (continuation) 117
5.17 Error metrics for Base load, Substation S5 118
5.18 Error metrics for Average load, Substation S5 119
5.19 Error metrics for Peak load, Substation S5 120
5.20 Error metrics for Peak load, Substation S5 (continuation) 121
5.21 Error metrics for Base load, Substation S6 122
5.22 Error metrics for Average load, Substation S6 123
5.23 Error metrics for Peak load, Substation S6 124
5.24 Error metrics for Peak load, Substation S6 (continuation) 125
5.25 Error metrics for Base load, Substation S7 126
5.26 Error metrics for Average load, Substation S7 127
5.27 Error metrics for Peak load, Substation S7 128
5.28 Error metrics for Peak load, Substation S7 (continuation) 129
5.29 Error metrics for Base load, Substation S8 130
5.30 Error metrics for Average load, Substation S8 131
5.31 Error metrics for Peak load, Substation S8 132
5.32 Error metrics for Peak load, Substation S8 (continuation) 133
5.33 Error metrics for Base load, Substation S9 134
5.34 Error metrics for Average load, Substation S9 135
5.35 Error metrics for Peak load, Substation S9 136
5.36 Error metrics for Peak load, Substation S9 (continuation) 137
5.37 Error metrics for Base load, Brasilia first period 141
5.38 Error metrics for Average load, Brasilia first period 142
5.39 Error metrics for Peak load, Brasilia first period 143
5.40 Error metrics for Base load, Brasilia second period 146
5.41 Error metrics for Average load, Brasilia second period 147
5.42 Error metrics for Peak load, Brasilia second period 148
5.43 Error metrics for Peak load, Brasilia second period (continuation) . . . 149
5.44 PV Systems in site A1 153
5.45 Results for Systems A1a and A1b, single station 155
5.46 Results for Systems A1c and A1d, single station 155
5.47 Results for Systems A1a and A1b, multiple stations 157
5.48 Results for Systems A1c and A1d, multiple stations 157
5.49 PV Systems in site A2 161
5.50 Results for Systems A2a and A2b, single station 162

5.51 Results for Systems A2a and A2b, multiple stations 163
5.52 PV Systems in site A3 165
5.53 Results for Systems A3a, A3b and A3c, single station 166
5.54 PV Systems in site A4 169
5.55 Results for System A4a, single station 170
5.56 Results for System A4a, multiple stations 171
5.57 PV Systems in site A5 172
5.58 Results for System A5a, single station 173
5.59 Results for System A5a, multiple stations 174
5.60 PV Systems in site A6 175
5.61 Results for Systems A6a and A6b,single station 177
5.62 Results for Systems A6a and A6b, multiple stations 178
5.63 PV Systems in site A7 180
5.64 Results for Systems A7a and A7b, single station 181
5.65 Results for Systems A7a and A7b, multiple stations 182

6.1 ANN and Kalman filter processing time ratio. 185
6.2 Summary of results - Leipzig scenario. Best methods per substation and load type. 186
6.3 Summary of results - Leipzig scenario. Best input sets per substation and load type. 186
6.4 Summary of results - Brasilia scenarios. Best methods per substation and load type. 187
6.5 Summary of results - Brasilia scenarios. Best input sets per substation and load type. 187
6.6 Summary of results - PV forecasting scenarios. Best methods per system.188

B.1 Coefficients for the Equation of Time and for the Sun declination . . . 227
B.2 Typical sample values of Albedo for different surfaces/enviroments . . . 234
B.3 SEDES2 Coefficients by wavelenght (1st part) 237
B.4 SEDES2 Coefficients by wavelenght (2nd part) 238
B.5 SEDES2 Coefficients by wavelenght (3rd part) 239

LIST OF FIGURES

1.1 Oferta de eletricidade e geração renovável alemã em TWh, entre 2001 e 2015. Fonte: [110], licença Creative Commons by SA 4.0. 4

1.2 Esquema simplificado de um sistema elétrico com geração distribuída . 5

1.3 Percentual de energia renovável no consumo líquido de eletricidade na Alemanha, de 2005 a 2015. Em 2015 as fontes renováveis supriram 38 % do consumo. Fonte: [105] 6

1.4 Evolução trimestral do número de unidades de geração distribuída conectadas às redes de distribuição brasileiras. Fonte: ANEEL 7

1.5 Capacidade instalada (esquerda) e número de unidades conectadas (direita) no Brasil, divididas por fonte primária de energia. Fonte: ANEEL 7

1.6 Custo médio líquido do sistema FV para o consumidor, considerando sistemas para instalação em telhados com potência nominal entre 10 kWp e 100 kWp. Fonte: [105] 8

2.1 German and European photovoltaic (PV) generated power in MW, between 2001 and 2015. Source: [110], Creative Commons license by SA 4.0. . . 3

2.2 Simplified schematic of a electric power system with Distributed Energy Resources 5

2.3 Percentage of renewable energy in Germany's net electricity consumption, from 2005 to 2015. In 2015 the renewable sources accounted for 38 % of the consumption. Source: [105] 6

2.4 Quarterly evolution of the number of DG units connected to the Brazilian grid. Source: ANEEL. 6

2.5 Installed capacity (left) and number of connections (right) of DG units in Brazil, by energy source. Source: ANEEL. 7

2.6 Average net system price to customer, for rooftop systems with nominal power from 10 kWp to 100 kWp. Source: [105] 7

3.1 From top left to bottom left: Load as an electric device, a building, a distribution feeder, a distribution substation, a citywide distribution grid and as national load centers. In specific contexts, load can refer to any of this six levels of aggregation. 13

3.2 Comparison of the forecasting error among five scenarios. The intrisic noise, which is oftenly related to measurement errors, is independent of the forecasting algorithm and is constant among all cases. 20

3.3 Peak load in Brasilia from 2001 to 2010. Weakly variation is visible in y axis (Day of the week), while the demand growth trend superposed to the seasonal variation is visible in the x axis (Week). 21

3.4 Factors and their main effects over the electricity demand. 24

3.5 This chart is a summary of the human comfort zone as a function of ambient conditions (weather and climate). Modified from the original in Wikimedia Commons, license CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0), acessible at https://commons.wikimedia.org/wiki/File%3APsychrometricChart.Sea

3.6 Scatter plot of mean temperature and mean demand in Megawatts (MW). Five polynomial best fitting showcases the large residuals and diverse possibilities from nonlinear behavior. 29

3.7 Logarithm of the absolute temperature, as measured in Leipzig from 2001 to 2003 30

3.8 Heating Degree Days at 18 Celsius reference, as measured in Leipzig from the 2001 to 2003. The heating peaks are measured in the winter season. 31

3.9 Scatter plot of HDD and mean demand in MW, as measured in Leipzig from the 2001 to 2003. Five polynomial best fitting curves showcase the more straightforward dependence, yet nonlinear. 32

3.10 Scatter plot of CDD and peak demand in MW, as measured in Brasília from 2001 to 2003. Because the large number of noncorrelated points stays at zero degree-days, they do not affect the determination of the CDD coefficient. 33

3.11 Maximum water vapor content of air as a function of temperature. This is the reference to calculation of Relative Humidity. 35

3.12 Natural illumination model. The five “window” surfaces are used to approximate the sunlight incident into buildings and open spaces. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licens sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/co Altitude_schematic.svg 41

3.13 (a) Cross section through a human eye. (b) Schematic view of the retina and its photoreceptors (adapted from Encyclopedia Brittannica, 1994 edition) 42

3.14 Normalized spectral sensivity of rod and cone cells. By Maxim Razin [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://commons.wikimedia.org/wiki/File:Cone-response.svg 43

3.15 Eye sensivity function. The values of CIE 1978 $V(\lambda)$ are shown in the left-hand ordenate, while the correspondent luminous efficacy (conversion factor for Watts to lumens) are shown in the right-hand ordenate. Both are maximum in 555 nm wavelenght. By Jordanwesthoff (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://commons.wikimedia.org/wiki/File% 45

3.16 Natural solar irradiation. The roof and four wall surfaces are used to approximate the heat absorption by irradiation into buildings. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licens sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/co Altitude_schematic.svg 46

3.17 60 day moving average of electricity tariffs in Brasília, in Brazilian Reais (BRL) per MWh, by consumer class 51

3.18 Global cumulative growth of PV capacity. Source: reproduced from IEA Solar photovoltaic roadmap 2014 [51] 58

3.19 Apparent difference between module types. From left to right. Polycrystaline Silicon module, Monocrystaline Silicon and Thin Film module. 59

3.20 Box diagram of the solar photovoltaic simulational framework 61

3.21 Sun azimuth angle and elevation angle. Azimuth reference is the geographical north pole. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/commons/f/f7 Altitude_schematic.svg 63

3.22 Direct Normal Irradiation as a function of time and wavelength 64

3.23 Schematic of a crystalline silicon solar panel 68

3.24 Typical spectral response of the polycrystalline silicon cell (blue) and glass reflectivity per wavelenght (red) 69

3.25 Example of Extraterrestrial Solar irradiation (AM0), ASTM Standard Spectrum (AM1,5) in cloudless sky and Absorbed Spectrum by a typical poly-Si cell (AM1,5) 70

4.1 Proposed data model for Load forecasting 76
4.2 Correlation between several candidate variables and peak demand, which corresponds to the projection in the planes $x = 0$ or $y = 0$ 79
4.3 Correlation between the 126 selected variables and peak demand, which corresponds to the projection in the planes $x = 0$ or $y = 0$ 80
4.4 Box Diagram of the proposed Kalman based predicting scheme 84
4.5 Proposed data model for Load forecasting 93

5.1 Leipzig population and population growth rate from 1990 to 2015. The 1999 growth peak is due to the incorporation of surrounding towns. Credits: EUROSTATs 98
5.2 Locations of the eight substations in Leipzig. Credits: Jayme Milanezi Jr. [73] 98
5.3 Evolution of electric load in Substation S1, from 2001 to 2004. Base load is plotted in black, Average load in blue and Peak load in red. . . 99
5.4 Sum of the Total Squared Error for the 8 substations as a function of Model Order. The minimum is achieved when the Order is set to 7. . . 100
5.5 Prediction (red line) plotted against the measured base load in Substation S1 (blue line) over 360 days of observation. 102
5.6 Prediction (red line) plotted against the measured average load in Substation S1 (blue line) over 360 days of observation. 103
5.7 Prediction (red line) plotted against the measured peak load in Substation S1 (blue line) over 360 days of observation. 105
5.8 Prediction (red line) plotted against the measured base load in Substation S2 (blue line) over 360 days of observation. 107
5.9 Prediction (red line) plotted against the measured average load in Substation S2 (blue line) over 360 days of observation. 108
5.10 Prediction (red line) plotted against the measured peak load in Substation S2 (blue line) over 360 days of observation. 109
5.11 Prediction (red line) plotted against the measured base load in Substation S3 (blue line) over 360 days of observation. 111
5.12 Prediction (red line) plotted against the measured average load in Substation S3 (blue line) over 360 days of observation. 112

5.13 Prediction (red line) plotted against the measured peak load in Substation S3 (blue line) over 360 days of observation. 113
5.14 Prediction (red line) plotted against the measured base load in Substation S4 (blue line) over 360 days of observation. 115
5.15 Prediction (red line) plotted against the measured average load in Substation S4 (blue line) over 360 days of observation. 116
5.16 Prediction (red line) plotted against the measured peak load in Substation S4 (blue line) over 360 days of observation. 117
5.17 Prediction (red line) plotted against the measured base load in Substation S5 (blue line) over 360 days of observation. 119
5.18 Prediction (red line) plotted against the measured average load in Substation S5 (blue line) over 360 days of observation. 120
5.19 Prediction (red line) plotted against the measured peak load in Substation S5 (blue line) over 360 days of observation. 121
5.20 Prediction (red line) plotted against the measured base load in Substation S6 (blue line) over 360 days of observation. 123
5.21 Prediction (red line) plotted against the measured average load in Substation S6 (blue line) over 360 days of observation. 124
5.22 Prediction (red line) plotted against the measured peak load in Substation S6 (blue line) over 360 days of observation. 125
5.23 Prediction (red line) plotted against the measured base load in Substation S7 (blue line) over 360 days of observation. 127
5.24 Prediction (red line) plotted against the measured average load in Substation S7 (blue line) over 360 days of observation. 128
5.25 Prediction (red line) plotted against the measured peak load in Substation S7 (blue line) over 360 days of observation. 129
5.26 Prediction (red line) plotted against the measured base load in Substation S8 (blue line) over 360 days of observation. 131
5.27 Prediction (red line) plotted against the measured average load in Substation S8 (blue line) over 360 days of observation. 132
5.28 Prediction (red line) plotted against the measured peak load in Substation S8 (blue line) over 360 days of observation. 133
5.29 Prediction (red line) plotted against the measured base load in Substation S9 (blue line) over 360 days of observation. 135
5.30 Prediction (red line) plotted against the measured average load in Substation S9 (blue line) over 360 days of observation. 136

5.31 Prediction (red line) plotted against the measured peak load in Substation S9 (blue line) over 360 days of observation. 137
5.32 Evolution of Brasilia's population and GDP between 1999 and 2014. Credits: CODEPLAN 138
5.33 Location of Juscelino Kubistchek International Airport relative to Brasilia and the Federal District. Credits: Google Maps 139
5.34 Evolution of electric load in Brasilia, from July 2001 to December 2003. Base load is plotted in black, Average load in blue and Peak load in red. 139
5.35 Total Squared Error for the second scenario, as a function of Model Order. The minimum is achieved when the Order is set to 8. 140
5.36 Prediction (red line) plotted against the measured base load in Brasilia (2001-2003 period) over 360 days of observation. 141
5.37 Prediction (red line) plotted against the measured average load (blue line) in Brasilia (2001-2003 period) over 360 days of observation. . . . 143
5.38 Prediction (red line) plotted against the measured peak load (blue line) in Brasilia (2001-2003 period) over 360 days of observation. 144
5.39 Evolution of electric load in Brasilia, from February 2004 to June 2010. Base load is plotted in black, Average load in blue and Peak load in red. Two outliers in the Peak load are not visible in this graph. 145
5.40 Total Squared Error for the second scenario, as a function of Model Order. The minimum is achieved when the Order is set to 8. 145
5.41 Prediction (red line) plotted against the measured base load (blue line) in Brasilia (2004-2010 period) over 360 days of observation. 147
5.42 Prediction (red line) plotted against the measured average load (blue line) in Brasilia (2004-2010 period) over 360 days of observation. . . . 148
5.43 Prediction (red line) plotted against the measured peak load (blue line) in Brasilia (2004-2010 period) over 360 days of observation. 149
5.44 The European sites selected for the forecast. Credits: Google Earth. . . 150
5.45 Location of the Australian sites selected for the forecast. Credits: Google Earth 152
5.46 Location of the North American generation site selected for the forecast. Credits: Google Earth 153
5.47 Site A1 and the six airport weather stations used for the forecast. Credits: Google Earth 154
5.48 Error graphs for forecasts in Systems A1a and A1b, from left to right. Single station. 156

5.49 Error graphs for forecasts in Systems A1c and A1d, from left to right. Single station 156

5.50 Error graphs for forecasts in Systems A1a and A1b, from left to right. Multiple stations. 158

5.51 Error graphs for forecasts in Systems A1c and A1d, from left to right. Multiple stations. 158

5.52 Prediction (red line) plotted against the measured PV Generation (blue line) in site A1a over 360 days of observation. 159

5.53 Prediction (red line) plotted against the measured PV Generation (blue line) in site A1b over 360 days of observation. 159

5.54 Prediction (red line) plotted against the measured PV Generation (blue line) in site A1c over 360 days of observation. 160

5.55 Prediction (red line) plotted against the measured PV Generation (blue line) in site A1d over 360 days of observation. 160

5.56 Site A2 and the two airport weather stations used for the forecast. Credits: Google Earth 161

5.57 Error graphs for forecasts in Systems A2a and A2b, from left to right. Single station. 162

5.58 Error graphs for forecasts in Systems A2a and A2b, from left to right. Multiple stations. 163

5.59 Prediction (red line) plotted against the measured PV Generation (blue line) in site A2a over 360 days of observation. 164

5.60 Prediction (red line) plotted against the measured PV Generation (blue line) in site A2b over 360 days of observation. 164

5.61 Site A3 and the airport weather station used for the forecast. Credits: Google Earth 165

5.62 Error graphs for forecasts in Systems A3a (top left), A3b (top right) and A3c (bottom). Single station. 167

5.63 Prediction (red line) plotted against the measured PV Generation (blue line) in site A3a over 360 days of observation. 168

5.64 Prediction (red line) plotted against the measured PV Generation (blue line) in site A3b over 360 days of observation. 168

5.65 Prediction (red line) plotted against the measured PV Generation (blue line) in site A3c over 360 days of observation. 169

5.66 Site A4 and the six airport weather stations used for the forecast. Credits: Google Earth 170

5.67 Error graphs for forecasts in System A4a. Single station. 171

5.68 Error graphs for forecasts in System A4a. Multiple stations. 171
5.69 Prediction (red line) plotted against the measured PV Generation (blue line) in site A4a over 360 days of observation. 172
5.70 Site A5 and the six airport weather stations used for the forecast. Credits: Google Earth 173
5.71 Error graphs for forecasts in System A5a. Single station. 174
5.72 Error graphs for forecasts in System A5a. Multiple stations. 174
5.73 Prediction (red line) plotted against the measured PV Generation (blue line) in site A5a over 360 days of observation. 175
5.74 Site A6 and the six airport weather stations used for the forecast. Credits: Google Earth 176
5.75 Error graphs for forecasts in Systems A6a and A6b, from left to right. Single station. 177
5.76 Error graphs for forecasts in Systems A6a and A6b, from left to right. Multiple stations. 178
5.77 Prediction (red line) plotted against the measured PV Generation (blue line) in site A6a over 360 days of observation. 179
5.78 Prediction (red line) plotted against the measured PV Generation (blue line) in site A6b over 360 days of observation. 179
5.79 Site A7 and the two airport weather stations used for the forecast. Credits: Google Earth 180
5.80 Error graphs for forecasts in Systems A6a and A6b, from left to right. Single stations. 181
5.81 Error graphs for forecasts in Systems A6a and A6b, from left to right. Multiple stations. 182
5.82 Prediction (red line) plotted against the measured PV Generation (blue line) in site A7a over 360 days of observation. 183
5.83 Prediction (red line) plotted against the measured PV Generation (blue line) in site A7b over 360 days of observation. 183

6.1 Scatter plot of the ANN to Kalman filter processing time ratio, as a function of the input size. 185
6.2 Principal components horizontally sorted in decreasing order of singular value. There are noticeable discontinuities around the first, the seventh and eighth component. Model order in this case has been selected as 7. 190
6.3 Principal components horizontally sorted in decreasing order of singular value. There are noticeable discontinuities around the first and the tenth component. Model order in this case has been selected as 10. 190

B.1 Electromagnetic wave propagating from left to right. The electric field is in a vertical plane and the magnetic field in a horizontal plane. The electric and magnetic fields are always in phase and at 90 degrees to each other. 216

B.2 Electromagnetic spectrum expressed in terms of energy and wavelength. In detail, the visible spectrum perceived by the human eyes as colors. . 216

B.3 Atmospheric Opacity as a function of the EM radiation wavelength. Note that the atmosphere is highly transparent to the visible spectrum. 217

B.4 Solar irradiance at space (yellow) and at sea level (red). For comparison, the gray line corresponds to the blackbody spectrum at 5778 K. 218

B.5 Dates for seasons, apoapsis and periapsis of Earth's orbit. The elliptical form is exagerated. 219

B.6 Analemma plotted as seen at noon GMT from the Royal Observatory, Greenwich (latitude 51.48° north, longitude 0.0015° west). 220

B.7 CERES-Aqua 2003-2004 mean annual clear sky and total sky albedo. Clear sky albedo is the fraction of the incoming solar radiation that is reflected back into space by regions of the Earth on cloud-free days. Total sky albedo include cloudy days. Data source: http://daac.gsfc.nasa.gov/giovanni/ 221

B.8 Solar Azimuth and Elevation (complement of Zenith Angle) angles. Panel Azimuth and tilt angles. Azimuth reference is the geographical south pole. 223

B.9 Burning a dry leaf with a magnifier lens. The bright spot over the smoking leaf is concentrated Direct irradiation of the Sun. The daylit ground is illuminated by the Global irradiance. The shadowed areas are dimly illuminated by the Diffuse irradiation component. Photography credits go to Dave Gough, available at https://www.flickr.com/photos/spacepleb/1505372433 (CC BY 2.0 license) 224

B.10 SPCTRL2 Extraterrestial Solar Radiation. 228

D.1 Perceptron with 3 inputs and bias. From left to right: inputs, weights, summation block, activation function and output. 246

D.2 Schematic of Multilayer Perceptron. From left to right, the input layer (light blue), the hidden layer (yellow) and output layer (green). 247

LIST OF SYMBOLS, NOMENCLATURES AND ACRONYMS

ANEEL: Acronym for Agência Nacional de Energia Elétrica, the Brazilian electricity regulatory agency.

AM: Air Mass.

ANN: Artificial Neural Network.

AR: Auto-Regressive model.

ARMA: Auto-Regressive Moving Average model.

ARMAX: Auto-Regressive Moving Average with eXogenous inputs model.

ARX: Auto-Regressive with eXogenous inputs model.

BP: Back-Propagation, an ANN training method.

CDD: Cooling Degree-Days.

CIE: Commission Internationale de l'Éclairage (International Commission on Illumination).

$cov(X, Y)$: Covariance of X and Y.

DER: Distributed Energy Resources. Refers to energy supplies, storage and power sources positioned closer to demand centers, frequently installed in customer sites.

DG: Distributed Generation. Refers to power sources positioned closer to demand centers, frequently installed in customer sites. Unlike DER, does not refers to storage technologies.

ELD: Enthalpy Latent Days.

GMT: Greenwich Mean Time, the mean solar time at the Royal Observatory in Greenwich, London.

HDD: Heating Degree-Days.

I_λ: Irradiance at wavelength λ, the power irradiated over a surface by a light source in the λ wavelength per unity of area.

KF: Kalman Filter.

λ: Wavelength of electromagnetic irradiation.

MAE: Mean Absolute Error metric.

MAPE: Mean Average Percentual Error metric.

MBE: Mean Bias Error metric.

METAR: METeorological Aerodrome Report. Acronym to a format for reporting weather information used by airports and pilots worldwide.

MLP: MultiLayer Perceptron, an ANN architecture.

MPE: Maximum Percentual Error metric.

MSE: Mean Squared Error metric.

MXE: MaXimum absolute Error metric.

PCA: Principal Component Analysis.

PV: PhotoVoltaic. Physical property that enables direct conversion of light into electricity using semiconducting materials.

ϕ_λ: Illuminance at wavelength λ, the luminous flux over a surface by a light source in the λ wavelength per unity of area.

$E(f), \overline{f}, \mu_f$: Expected value of the stochastic function $f[k]$.

RMSE: Acronym for the Root Mean Squared Error metric.

σ_f: Standard deviation of stochastic function $f[k]$.

SVD: Singular Value Decomposition.

$Var(f)$: Variance of stochastic function $f[k]$.

z: Notation for a scalar z.

$z[k]$: Value of the scalar function z at discrete time step k.

$\hat{z}[k]$: Prediction for the value of scalar function z at discrete time step k.

Z: Notation for a vector Z.

$Z[k]$: Value of the vectorial function Z at discrete time step k.

$\hat{Z}[k]$: Prediction for the value of vectorial function Z at discrete time step k.

$\mathbf{Z}$: Notation for a matrix $\mathbf{Z}$.

$\mathbf{Z}[k]$: Value of the matricial function $\mathbf{Z}$ at discrete time step k.

$\hat{\mathbf{Z}}[k]$: Prediction for the value of matricial function $\mathbf{Z}$ at discrete time step k.

1 INTRODUÇÃO

Mundialmente, o desenvolvimento econômico depende diretamente da disponibilidade de energia elétrica, especialmente em virtude da dependência quase total que a maioria das industrias e dos serviços essenciais têm de seu uso. A disponibilidade de uma fonte de energia perene, barata e confiável é de primordial importância econômica.

Grandes montantes do suprimento energético são mundialmente destinados a setores energeticamente intensivos, como o tratamento de água, irrigação, industria de transformação e transportes. Em particular, os países mais ricos tem as maiores demandas energéticas por habitante, uma vez que o Produto Interno Bruto (PIB) é altamente correlacionado com a utilização de energia.

Esta dependência pode ser linearmente modelada ao se considerar dados de 2003 a 2007 [2]. A relação causal entre crescimento econômico, caracterizado em diversos indicadores, e o consumo de eletricidade é investigado em inúmeros artigos. O estudo apresentado em [22] conclui que a causalidade é mais forte em paises desenvolvidos da OECD. Várias variáveis são utilizadas para indicar as dependências entre consumo de energia e atividades econômicas: Produto Interno Bruto (PIB), população e índices de preços [7]. Em [11], testes de Granger indicam relação de causalidade do consumo energético para a renda na Índia e na Indonésia, ao passo que o mesmo teste aponta para uma relação bidirecional para a Tailândia e as Filipinas. Esta dependência bidirecional aponta para um sistema retroalimentado, no qual a disponibilidade de um suprimento barato de energia promove o crescimento econômico, e então a atividade econômica aquecida demanda um consumo maior de eletricidade e/ou melhoria da eficiência energética. Deste ponto de vista, as demandas energéticas devem ser analisadas não somente como um serviço essencial, mas também como um indicador econômico.

Tabela 1.1: Suprimento de Eletricidade Doméstico, PIB e população de 2000 a 2015 - Mundo e Alemanha

Ano	DES	DES	PIB	Pop.	DES / Capita
	Mundo	Alemanha	Alemanha	Alemanha	Mundo — Alemanha
	(TWh)	(TWh)	10^9 US$	10^6Hab.	TWh /10^6Hab.
2000	15406,03	579,6	1949,95	82,21	2,52 — 7,05
2001	15638,45	585,1	1950,65	82,35	2,52 — 7,11
2002	16190,43	587,4	2079,14	82,49	2,58 — 7,12
2003	16793,16	600,7	2505,73	82,53	2,64 — 7,28
2004	17572,76	610,2	2819,25	82,52	2,73 — 7,39
2005	18333,46	614,1	2861,41	82,47	2,81 — 7,45
2006	19030,16	619,8	3002,45	82,38	2,89 — 7,52
2007	19922,93	621,5	3439,95	82,27	2,98 — 7,55
2008	20283,94	618,2	3752,37	82,11	3,00 — 7,53
2009	20123,69	581,4	3418,01	81,90	2,94 — 7,10
2010	21404,5	615,0	3417,30	81,78	3,09 — 7,52
2011	22050,91	606,1	3757,46	81,80	3,15 — 7,41
2012	22504,33	605,7	3543,98	80,43	3,17 — 7,53
2013	23092,66	603,8	3752,51	82,13	3,22 — 7,35
2014	24240,89	591,1	3879,28	80,98	3,34 — 7,30
2015	25893,62	595,1	3363,45	81,41	3,52 — 7,31

Os dados na Tabela 1.1 extraídos de [4, 3] mostram a evolução de três indicadores relacionados às economias mundial e alemã no período de 2000 a 2012. As primeiras duas colunas na Tabela 1.1 correspondem ao Suprimento de Eletricidade Doméstico no mundo e na Alemanha (DES), do inglês Domestic Energy Supply. A terceira e quarta coluna apresentam o PIB e a população em milhões de habitantes. A última coluna na Tabela 1.1 apresenta o suprimento de eletricidade per capita (DES/capita) para o mundo e para a Alemanha. É importante observar que a população alemã manteve-se praticamente constante, embora o montante de energia suprida tenha crescido.

O consumo por habitante na Tabela 1.1 segue uma curva ascendente no período entre 2000-2015, e isto indica a necessidade de contínuos investimentos na rede elétrica. A previsão de carga mostra-se, portanto, como uma ferramenta essencial para as companhias de distribuição de eletricidade. Devido às regulamentações de monopólio natural aprovadas na maioria dos países, estas empresas são obrigadas a cumprir variados padrões contratuais relacionados à confiabilidade, eficiência, segurança e outros aspectos da qualidade de energia. Além disto, as companhias devem igualmente levar em consideração a escassez e a flutuação de preços dos recursos energéticos, e também ações de responsabilidade ambiental como controles de emissão de CO_2 [5]. Por fim, as companhias devem também monitorar o crescimento da Geração Distribuída (GD) no lado da demanda, principalmente no que diz respeito à geração fotovoltaica, que está em rápida expansão no mundo [73].

Essa geração distribuída é tipicamente composta de unidades de geração com capacidade nominal variando de frações de kW a até 5 MW, interconectadas ao sistema de distribuição e instaladas juntamente com a carga do consumidor ou diretamente conectadas ao sistema elétrico, utilizando a rede para prover energia a uma unidade consumidora remota. Sistemas solares fotovoltaicos (FV) transformam a energia do Sol em eletricidade. Semicondutores que exibem o efeito fotovoltaico, por exemplo as células solares de silício tipo N ou tipo P, convertem a radiação solar em corrente elétrica contínua (DC). Inversores de frequência então são usados para converter a geração DC em corrente alternada (AC), a qual é injetada no sistema de potência.

Conforme exposto na Figura 1.1, ocorreu um crescimento exponencial na capacidade de renováveis na Alemanha, em particular de paineis fotovoltaicos [3]. Até 2010, cerca de metade de toda a energia FV gerada na Europa foi produzida na Alemanha, mas em virtude dos crescentes aumentos nos preços da energia e de políticas de incentivo à geração fotovoltaica adotadas por outros estados da União Européia, este percentual foi ligeiramente diminuido nos anos seguintes. Em 2015, as fontes renováveis supriram mais de 30 % do consumo de eletricidade na Alemanha.

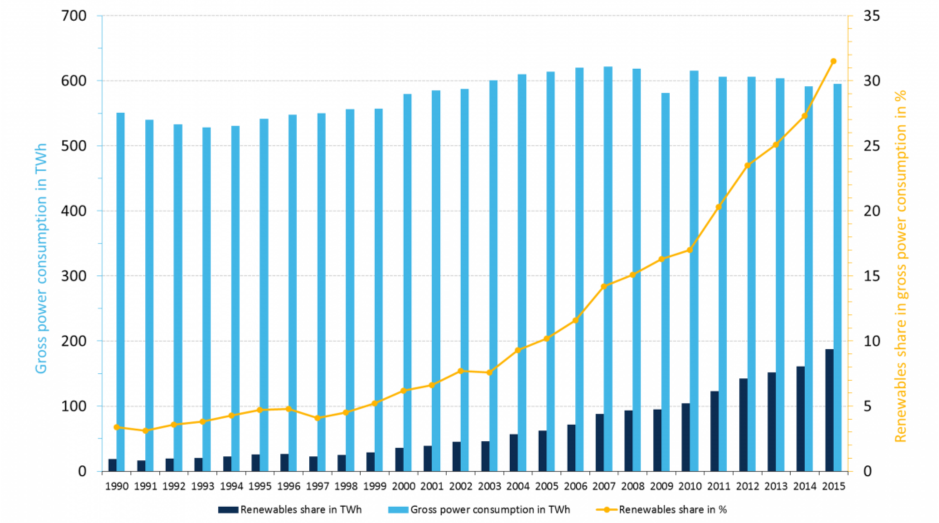


Figura 1.1: Oferta de eletricidade e geração renovável alemã em TWh, entre 2001 e 2015. Fonte: [110], licença Creative Commons by SA 4.0.

De acordo com [52], no ano de 2014, a geração de eletricidade foi responsável por 23.815 TWh ou 18 % do consumo mundial de energia, partindo de 6.287 TWh ou 9.4 % em 1974. Combustíveis fósseis permanecem como a principal fonte primária da eletricidade, uma vez que óleo, carvão e gás natural são responsáveis por 66,7 % da geração, menor que os 75,2 % em 1974. Hidroelétrica é a maior fonte primária renovável, suprindo 16,4 % da geração em 2013, decaindo de 20,9 % em 1974. A participação da fissão nuclear triplicou entre 1974 e 2014, indo de 3,3 % para 10,6 % da geração. Todas as outras fontes combinadas, incluido solar e eólicas, foram em 2013 responsáveis por 6,3 % da geração.

Um sistema elétrico é usualmente composto de três subsistemas: geração, transmissão e distribuição. Geração representa a etapa de conversão da fonte primária de energia em eletricidade, usualmente realizada em grandes usinas localizadas a uma distância física considerável até os centros de carga. A transmissão é composta por linhas de alta tensão, projetadas para transportar eficientemente grandes blocos de eletricidade da geração até os sistemas de distribuição. As redes de distribuição são o último elo com os consumidores no setor elétrico, sendo este subsistema responsável por reduzir a tensão para os níveis padronizados de consumo para fins industriais e residenciais, distribuindo eletricidade para um grande número de consumidores e garantindo que os padrões de qualidade de energia são atendidos.

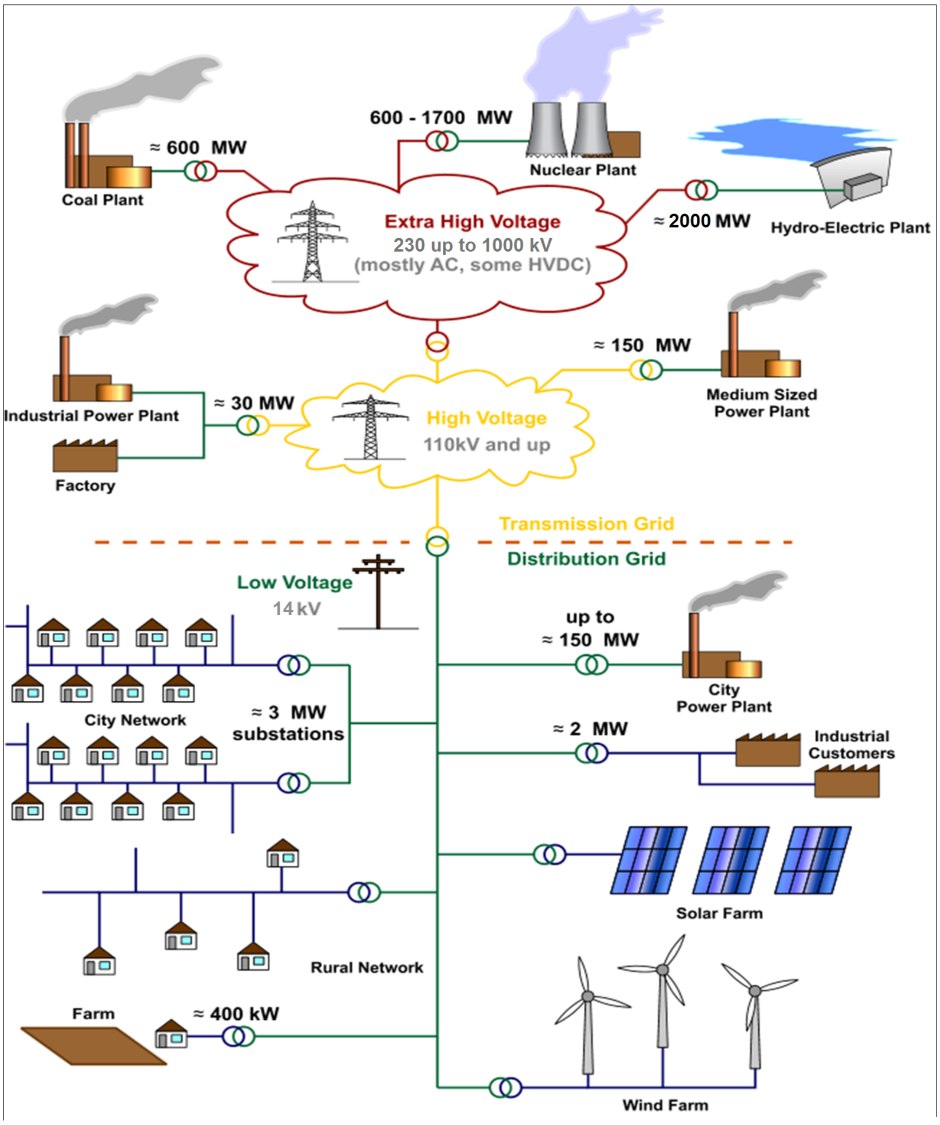


Figura 1.2: Esquema simplificado de um sistema elétrico com geração distribuída

Dado que o conjunto de requerimentos encontrados pelas companhias de distribuição remontam a um cenário complexo, ferramentas robustas de previsão de demanda são necessárias para implementar planos de expansão e operações eficientes e razoáveis. Os sistemas elétricos atuais requerem um permanente equilibrio entre geração e carga, pois sistemas de armazenamento de energia em larga escala ainda não atingiram viabilidade econômica para a maioria das redes elétricas. Na ocorrência de um desequilibrio entre geração e demanda de energia, a frequência do sistema passa a oscilar e as unidades geradoras devem rapidamente aumentar ou diminuir a geração para se restabelecer o equilibrio e restaurar a estabilidade de frequência do sistema. A reserva girante de geração empregada para manter a estabilidade no presente é resultado do planejamento e da previsão de carga realizados no passado. Os planos de operação que determinam quando cada gerador permanece em modo de espera ou em geração nominal são também oriundos de estudos de previsão de demanda.

A inserção de geração distribuída adiciona um novo nível de complexidade a esta tarefa, pois não somente a geração descentralizada reduz a carga de modo aleatório e intermitente, como também inevitavelmente produz alterações nas séries históricas de carga usadas para fazer as previsões. Ambos efeitos agem no sentido de aumentar os erros de predição no curto e no longo prazo, ameaçando a eficiência operacional e, no pior caso, a estabilidade do sistema [26].

Ao passo que todas as fontes de geração distribuída tem visto crescimento na sua capacidade instalada, a fonte solar fotovoltaica tem visto a maior taxa de implantação nos últimos anos. Nos Estados Unidos, Fotovoltaicas constituem de 80 a 90% da capacidade instalada dentre as instalações de GD com até 2 MW. Na Alemanha, de acordo com [105], a energia fotovoltaica gerada somou 38,5 TWh e supriu aproximadamente 7,5 % do consumo líquido de eletricidade da Alemanha em 2015, conforme ilustrado na figura 1.3. Em dias úteis ensolarados, a energia fotovoltaica pode atender 35 % da demanda instantânea, valor que sobe a 50 % em feriados e fins de semana. Ao fim de 2015, a capacidade nominal FV instalada na Alemanha foi de cerca de 40 GW distribuidos em 1,5 milhão de unidades geradoras. Com este nível de grandeza, a capacidade instalada em FV excede a de todas as demais fontes na Alemanha.

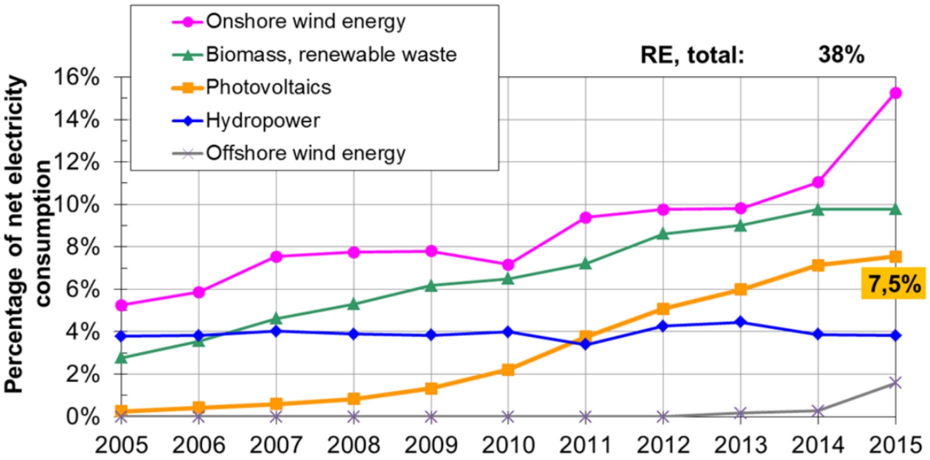


Figura 1.3: Percentual de energia renovável no consumo líquido de eletricidade na Alemanha, de 2005 a 2015. Em 2015 as fontes renováveis supriram 38 % do consumo. Fonte: [105]

Países em desenvolvimento e emergentes tem também experimentado tendências semelhantes. De acordo com a Agência Nacional de Energia Elétrica (ANEEL), desde a publicação da Resolução Normativa 482/2012, tem havido um constante crescimento no número de novas unidades de geração distribuída conectadas à rede de distribuição, conforme exposto na Figura 2.4. Esta Resolução Normativa regulamenta a conexão de geração distribuída às redes de distribuição, estabelecendo procedimentos e as obrigações das empresas de distribuição e dos consumidores.

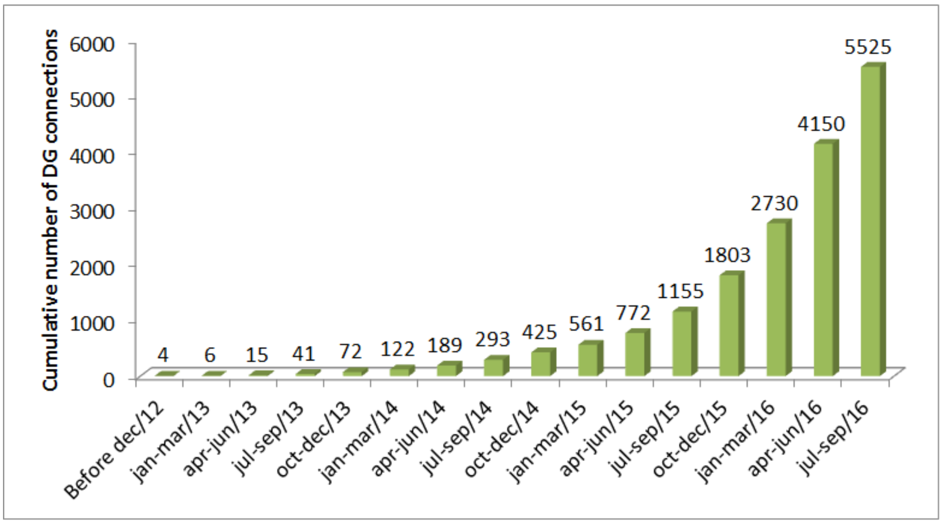


Figura 1.4: Evolução trimestral do número de unidades de geração distribuída conectadas às redes de distribuição brasileiras. Fonte: ANEEL

Semelhantemente ao observado na Alemanha e nos Estados Unidos, a geração solar fotovoltaica é a maior fonte de geração distribuída no Brasil, não somente em capacidade instalada como particularmente no número de unidades conectadas. Esta predominância é mostrada na figura 1.5.

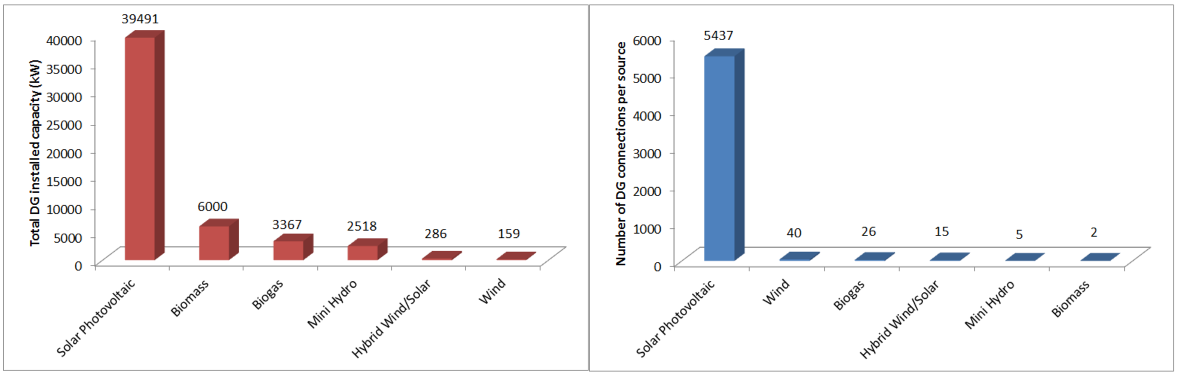


Figura 1.5: Capacidade instalada (esquerda) e número de unidades conectadas (direita) no Brasil, divididas por fonte primária de energia. Fonte: ANEEL

Este crescimento mundial da geração fotovoltaica é uma consequência da curva de aprendizado tecnológica e dos custos decrescentes, ilustrados na figura 2.6, que apresenta os preços de mercado na Alemanha. A fonte fotovoltaica tem experimentado rápido desenvolvimento tanto em custo quanto em performance. Nos Estados Unidos, foi observado que os custos diminuiram 31 % de 2010 para 2014 [26], enquanto que na Alemanha os custos caíram em quase 75 % desde 2006.

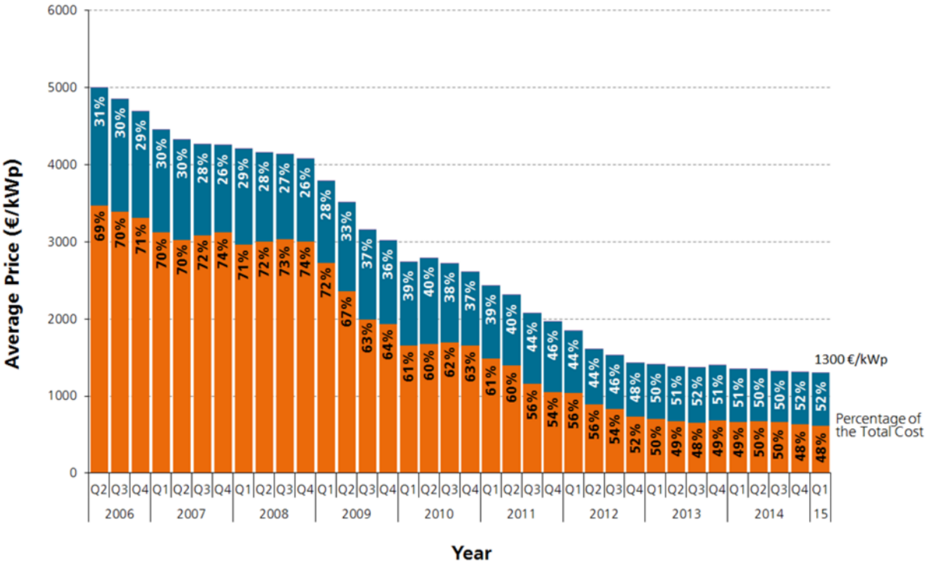


Figura 1.6: Custo médio líquido do sistema FV para o consumidor, considerando sistemas para instalação em telhados com potência nominal entre 10 kWp e 100 kWp. Fonte: [105]

Portanto, os operadores de sistemas de potência devem empregar ferramentas adaptativas que não somente são efetivas para prever a demanda, mas que também estão aptas a rastrear a mudança no comportamento da demanda ocasionado pela crescente presença de GD. Por outro lado, os fatores relevantes que governam a geração fotovoltaica, como a irradiação solar e a temperatura ambiente, são também correlacionados com o consumo de energia, embora de modo não linear. Aprimorando as metodologias de previsão para modelar e adaptar à presença de GD pode também melhorar ainda mais o desempenho destes métodos quando efetuando previsões em sistemas convencionais.

1.1 Formulação do problema

A previsão de carga é uma importante ferramenta empregada para assegurar que a energia suprida pelas distribuidoras está em equilíbrio com as cargas e com as perdas de energia inerentes ao sistema elétrico. A previsão de carga é sempre definida como a ciência ou arte de prever a carga futura em um dado sistema, por um período de tempo determinado. Estas predições podem prever a carga para as horas e minutos seguintes, com o objetivo de auxiliar a operação, ou predizer a demanda a 20 anos para fins de planejamento da expansão. A crescente capacidade instalada de recursos energéticos distribuídos levanta novas questões para este campo de pesquisa, pois é necessário prever não apenas o crescimento da capacidade como também a geração intermitente associada à GD.

Com relação às escalas de tempo e ao alcance das predições, previsão de demanda pode ser categorizada em três áreas [90]:

1. Previsão de longo prazo, que é utilizada para predizer as cargas para até 50 anos no futuro, de modo que a suportar o planejamento para a expansão;
2. Previsão de médio prazo, que é utilizada para prever cargas semanais, mensais e anuais a até 10 anos no futuro, permitindo o planejamento eficiente das operações do sistema;
3. Previsão de curto prazo, que é empregada para prever cargas até uma semana no futuro, de modo a minimizar custos das operações diárias e despachos de geração.

Nas três categorias precedentes, um modelo acurado é necessário para representar matematicamente a relação entre a carga e variáveis influentes, como datas, clima, fatores econômicos, entre outros. A relação precisa entre a carga e estas variáveis é usualmente determinada pelo seu papel no modelo de carga. Uma vez que este é construído, os parâmetros do modelo são determinados por meio de técnicas de estimação. Existem cinco componentes fundamentais em um problema de estimação [90]:

1. As variáveis a serem estimadas
2. As medições ou observações disponíveis
3. O modelo matemático que descreve como as medições estão relacionadas com a variável de interesse
4. O modelo matemático das incertezas de medição e estimação
5. O critério de avaliação de desempenho que irá julgar qual algoritmo é o “melhor”.

Nos últimos 50 anos, os algoritmos de estimação de parâmetros usados na previsão de demanda limitaram-se à múltipla regressão baseada no critério de minimização de erro dos mínimos quadrados [47]. Estes métodos evoluiram para os de séries temporais estocásticas, a exemplo dos modelos Autorregressivos (AR) e de Médias Móveis (MA). Atualmente, o estado da arte reside em modelos de espaço de estados finamente ajustados e em sistemas especialistas, baseados em técnicas de aprendizado de máquina. Além disso, as Redes Neurais Artificiais (RNA) têm mostrado sucesso como a base de sistemas especialistas para a previsão de curto prazo. Contudo, o sistema especialista utilizado por uma empresa não necessariamente é adequado para o uso em um sistema elétrico diferente, no mínimo requerendo novo treinamento e algumas vezes a troca de variáveis ou do próprio modelo matemático para tornar-se útil para uma empresa diferente.

Este trabalho apresenta a previsão de carga e geração como um problema de estimação dinâmica de estado. As variáveis a serem estimadas são as demandas base, média e de pico, assim como a geração fotovoltaica. Como medições e observações, este trabalho utiliza previsões de tempo, datas e eventos de calendário, tarifas de energia, índices e estimativas econômicas e demográficas. Combinações preprocessadas destas medições são usadas como as variáveis de entrada para a previsão. O modelo matemático é a representação em espaço de estados, e as matrizes de covariância do filtro de Kalman modelam as incertezas. Os critérios de performance incluem Erro Médio Quadrático (MSE), Erro Médio Percentual (MAPE) e Erro Máximo Percentual (MPE). Além destes, diferentes abordagens empregadas para solucionar este problema de estimação dinâmica de estados são também discutidas, assim como são realizadas comparações entre a solução propostas e outras soluções do estado da arte.

A presente dissertação contribui para este tópico de pesquisa ao propor e validar ferramentas de análise para produzir, aprimorar e selecionar conjuntos relevantes de variáveis de entrada, o que melhora a capacidade dos algoritmos de predição para prever carga e geração fotovoltaica. Um esquema de predição híbrido baseado em filtros de Kalman e modelos Grey é apresentado para com confiabilidade e precisão realizar a predição de carga e geração distribuída. Como um resultado secundário, a modelagem de carga adotada neste trabalho pode ser empregada para sintetizar cargas estocásticas e geradores distribuídos em sistemas elétricos simulados.

1.2 Organização da dissertação

Esta dissertação está dividida em seis capítulos, bibliografia e quatro apêndices. Os capítulos apresentam os tópicos principais e as conclusões obtidas nesta pesquisa, enquanto que a base matemática e os conceitos úteis que suportam as proposições deste trabalho estão organizados em apêndices.

O Capítulo 2 contém uma versão desta introdução em inglês, ao passo que este Capítulo 1 apresenta o mesmo conteúdo em lingua portuguesa.

O Capítulo 3 trata da modelagem da carga elétrica e da geração fotovoltaica que é desenvolvida neste trabalho. Premissas, suposições e conceitos aplicados ao longo deste trabalho para construir modelos razoáveis para cada tipo de carga e de gerador fotovoltaico podem ser encontrados neste capítulo. Também existe, para cada tipo de carga e gerador, uma discussão e os passos de preprocessamento necessários para produzir os fatores relevantes que devem ser incluídos como variáveis de entrada no algoritmo de predição.

O Capítulo 4 propõe um esquema adaptativo de predição para planejamento e operação de redes elétricas baseado em filtros de Kalman. Além do algorítmo de predição, o capítulo também destaca o modelo de dados empregado, explicando a natureza e o preprocessamento das variáveis exógenas que são selecionadas como dados de entrada, assim como aborda o estado da arte em técnicas de predição de carga e de geração fotovoltaica.

O Capítulo 5 apresenta os resultados das previsões em diversos cenários. O algorítmo de previsão de demanda é utilizado para predizer as cargas de base, média e de pico em dois sistemas elétricos distintos, um na Alemanha e outro no Brasil. A previsão de geração fotovoltaica é realizada em dois locais, na Holanda e na Nova Zelândia. Um comparativo é feito entre os algoritmos propostos e métodos do atual estado da arte.

O Capítulo 6 sumariza as realizações e constatações obtidas por esta pesquisa, realizando uma conclusão objetiva a respeito dos algorítmos de predição, os comparativos e as direções a serem tomadas em trabalhos subsequentes.

A bibliografia contém a listagem das referências bibliográficas que são citadas neste trabalho, organizadas em ordem alfabética.

O Apêndice A apresenta uma introdução para representações em espaço de estados e filtros de Kalman, também demonstrando o modelo específico e a estrutura de filtro empregadas no método de previsão proposto. Apêndice B trata da modelagem da irradiação solar e simulação empregada tanto para prever a geração fotovoltaica quanto para aprimorar as previsões de demanda. Apêndice C fornece um sucinto fundamento a respeito de Análise de Componentes Principais, a principal ferramenta empregada para selecionar as variáveis de entrada para o algorítmo proposto. O Apêndice D apresenta um resumo das técnicas de redes neurais artificiais usadas nesta pesquisa.

2 INTRODUCTION

Economic development, throughout the world, depends directly on the availability of electric energy, especially because most industries and basic services depend almost entirely on its use. The availability of a source of continuous, cheap, and reliable energy is of foremost economic importance.

Large amounts of energy supply are set apart worldwide to energetically intensive sectors such as water treatment, irrigation, transformation industry and transport. In particular, the richest countries have the highest energetic demands per inhabitant since Gross Domestic Product (GDP) is highly correlated with the energy demands.

This dependence can be linearly modeled considering data from 2003 to 2007 [2]. The causal relationship between economic growth, characterized in diverse indicators, and the electricity consumption is investigated in numerous of papers. The study presented in [22] conclude that causality is stronger in developed OECD countries than in developing countries. Several variables are used to assert the dependencies between energy consumption and economic activities: Gross Domestic Product (GDP), population and price indexes [7]. In [11], Granger tests indicate short-run causality from energy consumption to income for India and Indonesia, while the test points to bidirectional relationship for Thailand and the Philippines. This bidirectional dependence points towards a feedback loop, where the availability of cheap energy supply promotes economic growth, and then the increased economic activity demands a even greater energy consumption and/or improved energy efficiency. From this standpoint, energy demands should be approached not only as a essential service, but also as an economic issue.

Table 2.1: Worldwide and Domestic German Energy Supply (DES), German Gross Domestic Product (GDP) and Population from 2000 to 2015

Year	DES	DES	GDP	Pop.	DES / Capita
	World	Germany	Germany	Germany	World — Germany
	(TWh)	(TWh)	10^9 US$	10^6Inhab.	TWh /10^6Inhab.
2000	15406,03	579,6	1949,95	82,21	2,52 — 7,05
2001	15638,45	585,1	1950,65	82,35	2,52 — 7,11
2002	16190,43	587,4	2079,14	82,49	2,58 — 7,12
2003	16793,16	600,7	2505,73	82,53	2,64 — 7,28
2004	17572,76	610,2	2819,25	82,52	2,73 — 7,39
2005	18333,46	614,1	2861,41	82,47	2,81 — 7,45
2006	19030,16	619,8	3002,45	82,38	2,89 — 7,52
2007	19922,93	621,5	3439,95	82,27	2,98 — 7,55
2008	20283,94	618,2	3752,37	82,11	3,00 — 7,53
2009	20123,69	581,4	3418,01	81,90	2,94 — 7,10
2010	21404,5	615,0	3417,30	81,78	3,09 — 7,52
2011	22050,91	606,1	3757,46	81,80	3,15 — 7,41
2012	22504,33	605,7	3543,98	80,43	3,17 — 7,53
2013	23092,66	603,8	3752,51	82,13	3,22 — 7,35
2014	24240,89	591,1	3879,28	80,98	3,34 — 7,30
2015	25893,62	595,1	3363,45	81,41	3,52 — 7,31

The data in Table 2.1 extracted from [4, 3] shows the evolution of three indicators related to the world and the German economies over the period from 2000 to 2012. The first two columns in Table 2.1 correspond to the Domestic Energy Supply (DES) in the world and in Germany, respectively, while the third and fourth columns are the GDP and the population in millions of inhabitants in Germay. Finally, the last column in Table I presents DES/Capita for the World and Germany. Note that German population is practically constant although the amount of energy supplied has increased.

The energy consumption per inhabitant (DES/Capita) in Table 2.1follows an ascendant curve within 2000-2015, and that indicates the need for continuous investments in the electric grid. Load forecasting comes up, therefore, as an essential tool for the electricity distribution companies. Due to natural monopoly regulations enacted on most countries, these companies must comply with several contractual standards related to reliability, efficiency, safety and other power quality aspects. Moreover, the companies should equally take into account the scarcity and fluctuation of energy resources as much as environmental care such as CO_2 emissions control [5]. Furthermore, the companies should also take heed of the increase on distributed generation on the demand side, mainly concerning photovoltaic generation, which is in rapid expansion throughout the world.

These are typically comprised of generation units rated from fractional kW and up to 5 MW in nameplate capacity, interconnected to the distribution system and installed either behind the consumer's load or directly connected to the system, using the grid to provide power to a remote consumer unit. Solar photovoltaic systems transform solar energy into electric power. Semiconductors that exhibit the photovoltaic effect, such as silicon-N or silicon-P solar cells, convert solar radiation into Direct Current (DC) electricity. Solid state inverters then converts the DC generation into Alternate Current (AC), which is injected into the power grid.

As depicted in Fig. 2.1, there has been an exponential growth in the installed capacity of renewable sources in Germany, photovoltaic panels in particular [3]. Until 2010, over half of the entire PV generated power in Europe came out from Germany, but due to growing energy prices and PV friendly policies adopted by other EU states, this percentange has slightly decreased in the following years. In 2015, the renewable sources supplied more than 30 % of Germany's electricity consumption.

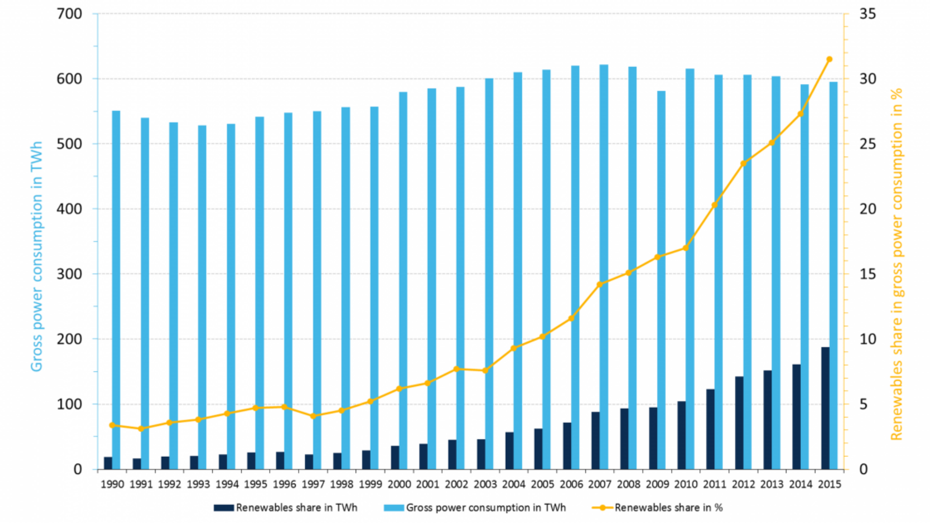


Figure 2.1: German and European photovoltaic (PV) generated power in MW, between 2001 and 2015. Source: [110], Creative Commons license by SA 4.0.

According to the [52], in year 2014, the electricity generation accounted for 23,815 TWh or 18 % of the World total energy consumption, up from 6,287 TWh or 9.4 % in 1974. Fossil fuels are still the main primary source for electricity, as oil, coal and natural gas accounts for 66.7 % of the generation, down from 75.2 %. Hydropower is the main renewable source, supplying 16,4 % of the electricity generation in 2014, down from 20.9 %. Nuclear fission share has increased threefold between 1974 and 2014, from 3.3% to 10.6 % of the generation. The other sources combined, including solar and wind power, account for 6.3 % of the electricity production.

An electric system is usually composed of three subsystems: generation, transmission and distribution. Generation represents the conversion of a primary energy source into electricity, usually performed in large scale facilities at a considerable physical distance from the consumption centers. Transmission is comprised of high voltage power lines, designed to efficiently transport large blocks of electricity from generation to distribution facilities. Distribution grids are the last link to the consumers in the electric system, responsible to decreasing voltage for industrial and residential consumption and distributing the electricity to several consumers while ensuring that power quality standards are met.

Since the set of requirements faced by electricity distribution companies assemble a complex scenario, robust load forecasting tools are needed to implement efficient and reasonable expansion and operation plans. Current electric power systems require a permanent balance between generation and load, since large scale energy storage has not achieved economical feasibility in most power grids. At the onset of an unbalance between load and generation, the system frequency starts to oscillate and generation units must be quickly stepped up or down in order to reobtain balace and to restore the system's frequency stability. The generation spinning reserve used to keep stability today is the result of planning and forecast performed several years prior. The operational plans that determine when each generator stays at stand by or at full power is also a product of load forecasts.

The introduction of distributed generation adds a new level of complexity to this task, as not only the decentralized generation reduces load in a random and intermittent way, but also inevitably embeds in the historic loads used to forecast. Both effects act

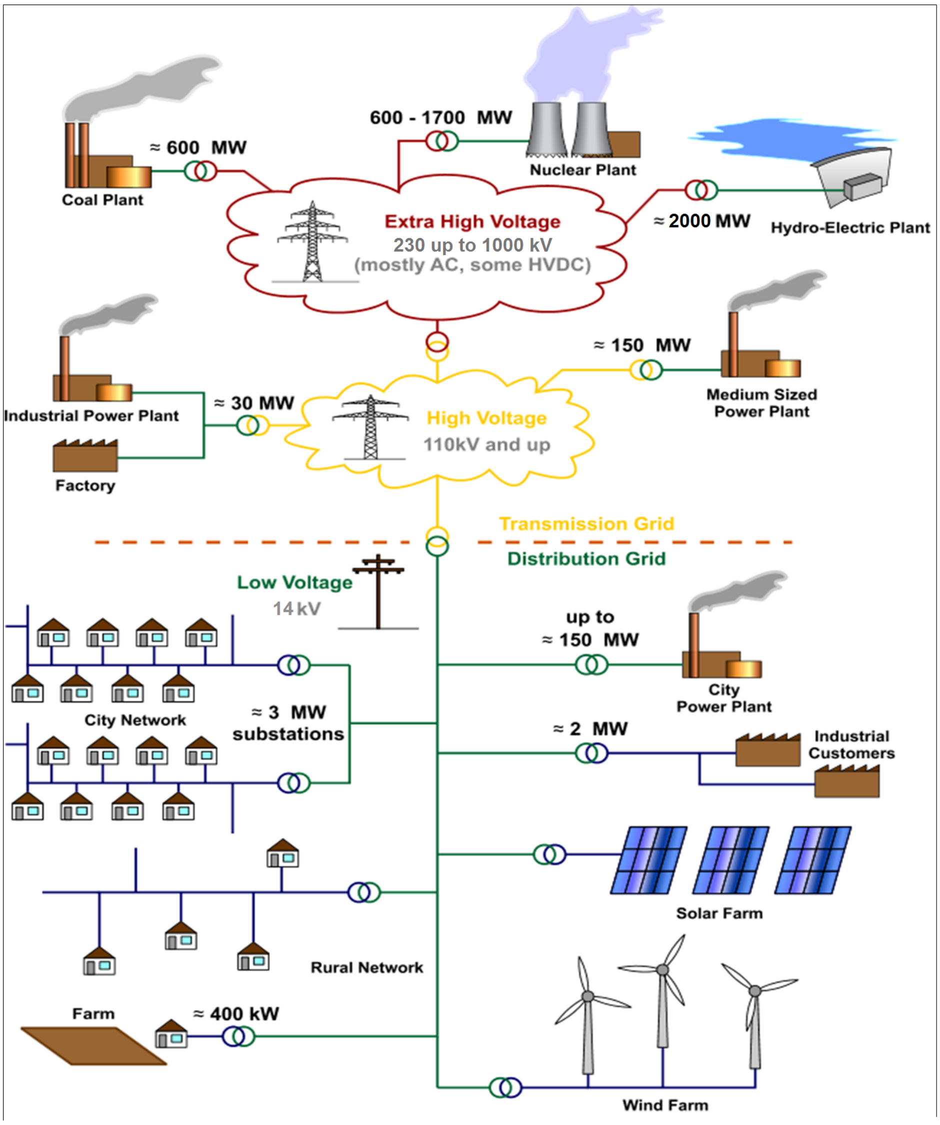


Figure 2.2: Simplified schematic of a electric power system with Distributed Energy Resources

to increase prediction errors in short and long term, jeopardizing operational efficiency and, in worst case, system reliability [26].

While all DERs have seen growth in installed capacity, photovoltaic solar has seen the largest adoption in recent years. In the US, photovoltaic constitutes 80 to 90 % of the total installed capacity among DER installations two megawatts or less. In Germany, according to [105], photovoltaic generated power amounted to 38.5 TWh and covered approximately 7.5 % of Germany's net electricity consumption in 2015, as depicted in Figure 2.3. On sunny weekdays, PV power can cover 35 percent of the momentary electricity demand, while on weekends and holidays the coverage rate of PV can reach 50 percent. At the end of 2015, the total nominal PV power installed in Germany was circa 40 GW, distributed over 1.5 million power plants. With this figure, the installed

PV capacity exceeds that of all other types of power plants in Germany.

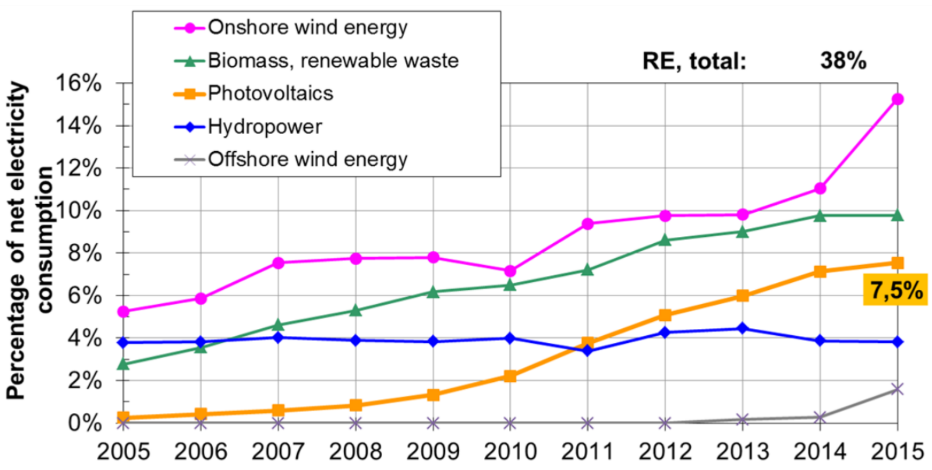


Figure 2.3: Percentage of renewable energy in Germany's net electricity consumption, from 2005 to 2015. In 2015 the renewable sources accounted for 38 % of the consumption. Source: [105]

Developing and emergent countries are also experiencing similar trends. According to the Brazilian Electricity Regulatory Agency (ANEEL), since the the normative resolution 482/2012 was enacted, there has been a steady growth in the number of DG units connected to the distribution grid, as shown in Figure 2.4. This normative resolution regulates the connection of DER to the distribution grids, establishing the procedures and obligations for the utilities and consumers.

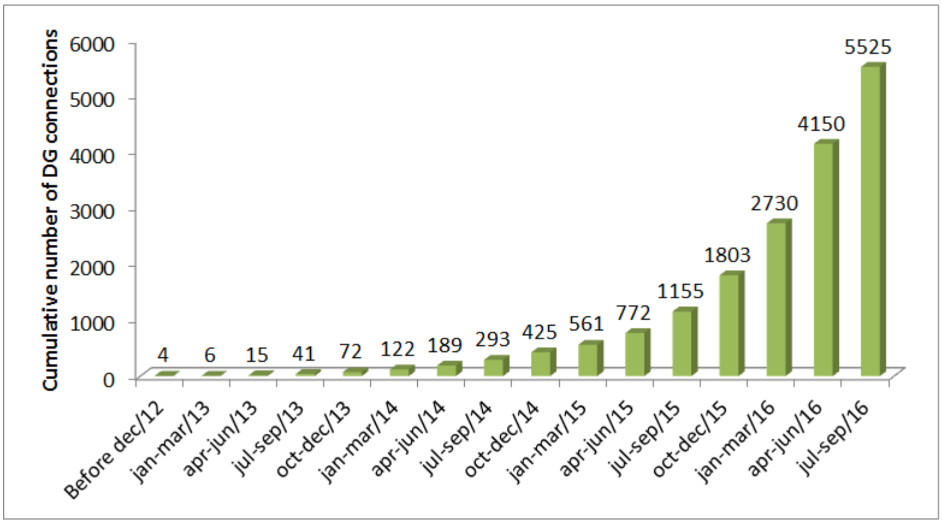


Figure 2.4: Quarterly evolution of the number of DG units connected to the Brazilian grid. Source: ANEEL.

Likewise Germany and United States, Solar photovoltaic is the largest distributed generation source in Brazil, not only in installed capacity but particularly in the number of connections. This predominance is illustrated in Figure 2.5.

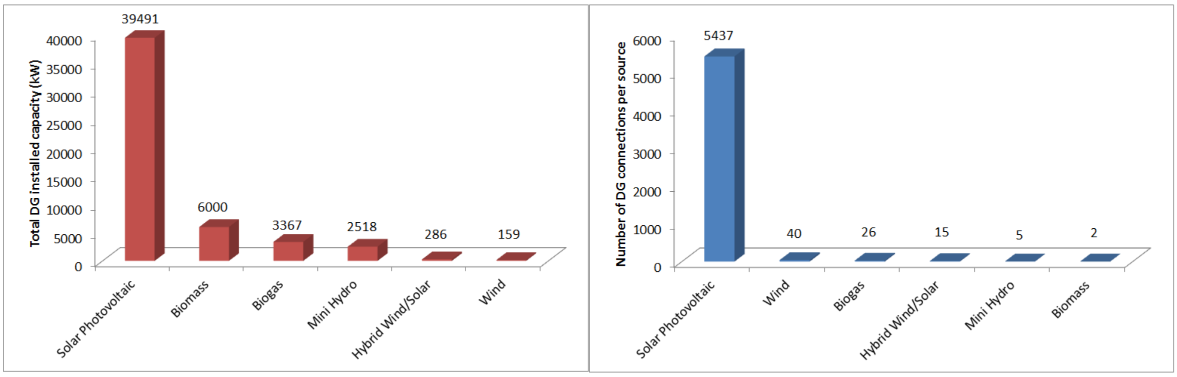


Figure 2.5: Installed capacity (left) and number of connections (right) of DG units in Brazil, by energy source. Source: ANEEL.

This worldwide growth of PV is a consequence of the technological learning curve and its decreasing costs, illustrated in figure 2.6, which depicts Germany's market prices. PV has experienced rapid development in terms of both cost and performance. In the United States, it has been reported that costs decreased by 31 % from 2010 to 2014 [26], while in Germany costs have dropped by almost 75 % since 2006.

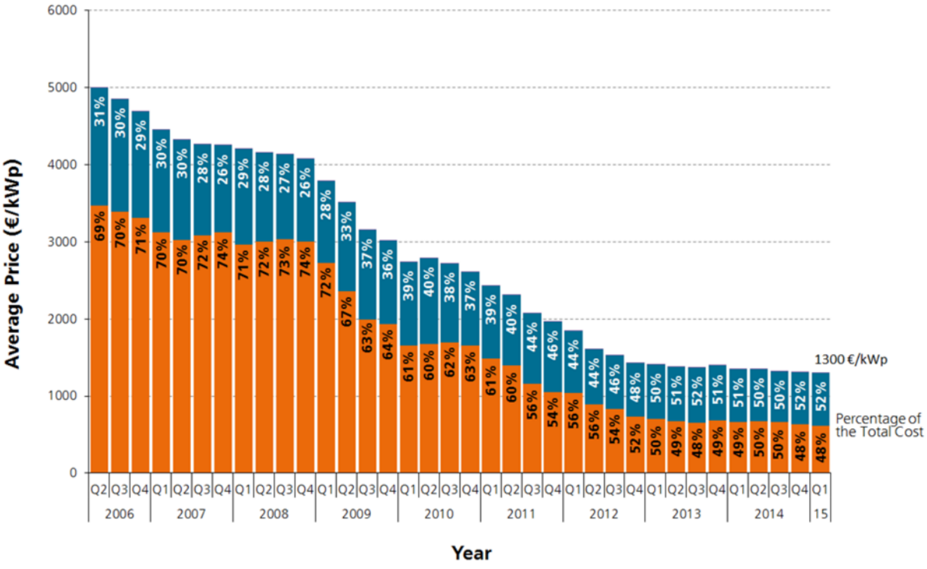


Figure 2.6: Average net system price to customer, for rooftop systems with nominal power from 10 kWp to 100 kWp. Source: [105]

Therefore, power system operators must employ adaptive tools that not only can reliably predict load, but also be able to track the change in the demand behavior caused by the growing presence of Distributed Energy Resources (DER). On the other hand, the relevant factors that drive PV generation such as solar irradiation and ambient temperature are also correlated to the electric load, albeit indirectly or nonlinearly. Improving the forecasting methodologies to model and adapt to distributed PV generation could also further enhance the performance of such methods when predicting conventional systems.

2.1 Problem formulation

Electrical load forecasting is an important tool used to ensure that the energy supplied by utilities meets the load plus the energy lost in the system. Load forecasting is always defined as basically the science or art of predicting the future load on a given system, for a specified period of time ahead. These predictions may be just for a fraction of an hour ahead for operation purposes, or as much as 20 years into the future for planning purposes. The growing installed capacity of distributed energy resources raises new questions to this research field, as not only the growth rate but also the intermittent power generation must be predicted.

Regarding the time scales and prediction range, load forecasting can be categorized into three subject areas, namely [90]:

1. Long term forecasting, which is used to predict loads as distant as 50 years ahead so that expansion planning can be facilitated;
2. Medium term forecasting, which is used to predict weekly, monthly, and yearly peak loads up to 10 years ahead so that efficient operational planning can be carried out;
3. Short term forecasting, which is used to predict loads up to a week ahead so that daily operations and dispatching costs can be minimized.

In the preceding three categories, an accurate load model is required to mathematically represent the relationship between the load and influential variables such as time, weather, economic factors, and so on. The precise relationship between the load and these variables is usually determined by their role in the load model. After the mathematical model is constructed, the model parameters are determined through the use of estimation techniques. There are five fundamental components of an estimation problem [90]:

1. The variables to be estimated
2. The measurements or observations available - weather forecasts.
3. The mathematical model describing how the measurements are related to the variable of interest.
4. The mathematical model of the uncertainties present.
5. The performance evaluation criterion to judge which estimation algorithms are “best”.

Over the past 50 years, the parameter estimation algorithms used in load forecasting have been limited to multiple variable regression based on the least error squares minimization criterion [47]. These have evolved to stochastic time series approaches, such as Autorregressive (AR) and Moving Average (MA). Currently, the state-of-art resides in finely tuned Space state models and Expert systems, which are based in machine learning techniques. Furthermore, the artificial neural network (ANN) had showed success as the basis of expert systems for short term forecasting . However, the expert system used by a utility is not necessarily suitable for a different power system, at least requiring retraining and sometimes a change of variables or mathematical model to be useful for other electric utility company.

This work presents the load and generation forecasting as a dynamic state estimation problem. The variables to be estimated are daily base, average and peak electric load, as well as PV generation. As measurements and observations, this work uses weather forecasts, calendar dates and events, energy tariffs, economical and demographic indexes and estimatives. Preprocessed combinations of these measurements are the input variables employed for forecasting. The mathematical model is a State space representation, and the Kalman filter covariance matrices model the uncertainties. The performance criteria encompasses Mean Square Error (MSE), Mean Average Percentual Error (MAPE) and Maximum Percentual Error (MPE). Furthermore, the different approaches used to solve this dynamic estimation problem are also discussed, as well as comparisons are performed between the proposed solution and other state-of-art approaches.

The present work contributes to this research subject by proposing and testing analysis tools to produce, enhance and select a relevant set of input variables, enhancing the predicting algorithms ability to forecast load and PV generation. A hybrid Kalman based predicting scheme is presented in order to realiably and accuratelly forecast the electric load and photovoltaic generation. As a side result, the load modelling adopted in this work can be employed to synthesize stochastic electric loads and distributed generators on simulated electrical systems.

2.2 Organization of this dissertation

This dissertation is divided in six chapters, bibliography and four appendices. The chapters present the core topics and key findings of this research, while the mathematical background and the useful concepts that support the proposals of this work are organized in four appendices.

Chapter 1 portrays a version of this introduction in Portuguese, while this second chapter has the same content in English language.

Chapter 3 comprises the load and PV generation modelling that is developed in this work. Premises, assumptions and concepts envisaged in this research to build reasonable models for each type of electric load and PV generator will be found in this chapter. There is also, for each type of load and generator, a discussion and preprocessing steps needed to produce the relevant factors that must be included as input variables in the forecasting algorithm.

Chapter 4 proposes the Kalman based adaptive prediction scheme for electric grid planning and operation. Besides the forecasting algorithm, the chapter also features the data model used, explaining the nature and pre-processing of the exogeneous variables which are adopted as inputs, as well as approaching the state-of-art in electric load and PV generation forecasting.

Chapter 5 presents the forecasting results in several scenarios. The load forecasting algorithm is used to predict base, average and peak load in two different power systems, one in Germany and the other in Brazil. The PV generation forecasting is also performed in two locations, in Netherlands and in New Zealand. A comparison is made between the proposed algorithm and current state-of-art methods.

Chapter 6 summarizes the achievements and findings obtained by this research, accomplishing an objetive conclusion with respect to the forecasting algorithms, the comparisons and directions for future work.

The bibliography contains the listing of the bibliographic references that are cited in this work, sorted in alphabetical order.

Appendix A presents an introduction to State Space representations and Kalman filters, also demonstrating the specific representation and filter structure employed in the proposed forecasting method. Appendix B deals with the solar irradiation modelling and simulation used both to predict PV generation and to enhance load forecasting. Appendix C gives a succint primer about Principal Component Analysis, the main tool for input variable selection featured in the proposed forecasting algorithm. Appendix D presents an overview of the artificial neural network techniques used in this research.

3 LOAD AND GENERATION MODELING

The objective of this chapter is to present the common mathematical foundations that encompass most of the methods currently employed to forecast electrical load and PV generation, as well as provide a detailed description of the exogenous variables that are employed as inputs in thepredictors proposed in this dissertation. Section 3.1provides an introduction, while Section 3.2 gives a succinct description of Generalized Additive Models and their relationship with linear, time series, state space and Artificial Neural Network approaches. Section 3.3 provides a description of the most important drivers and variables related to the electric demand. Section 3.5 deals with photovoltaic generation modelling.

3.1 Overview

Accurate load models in conjunction with efficient predictors are basic requirements for the optimum economic operation of power systems. A prerequisite to the development of an accurate load-forecasting model is an understanding of the characteristics of the load to be modeled. This knowledge of load behavior is gained from experience with the load and thorough statistical analysis of past demand time series. Utility companies with similar cultural, climatic and economic contexts usually experience similar load behavior, thus allowing load models developed for one utility to suit another company with slight modifications.

The term “load” is a wide conception, assuming different meanings in the context of power systems. In the strictest sense, load is the electrical device connected to a power system that consumes energy. In the wider sense, it represents the total power (active and/or reactive) consumed by all devices connected to a power system. In-between these two meanings, load can also designate a portion of the system that is not explicitly represented in a system model, but rather is treated as if it were a single power-consuming device connected to a bus in the system model. This single device can represent the electric devices in a building floor, an entire building, a feeder bus or even an distribution network. In this dissertation, the term “load” refers to the electric demand as measured in a distribution substation.

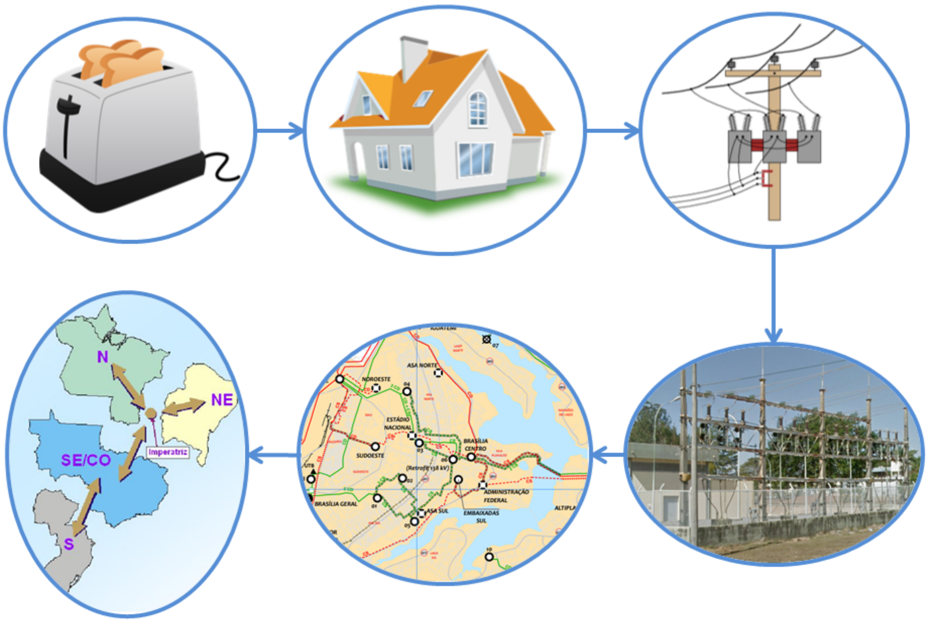


Figure 3.1: From top left to bottom left: Load as an electric device, a building, a distribution feeder, a distribution substation, a citywide distribution grid and as national load centers. In specific contexts, load can refer to any of this six levels of aggregation.

A load model in this matter is a mathematical representation of the relationship between power and exogenous variables causally related to the load, where the active power is the output from the model and the exogenous variables are its inputs. The system load is a random and non-stationary process composed of a very large number of individual components. The load behavior is influenced by a number of variables, such as weather, day of the week, the season, social, demographic and economic factors, as well as other relevant inputs. A number of papers discussing load modeling can be found in literature, presenting several techniques [47].

Linear models are widely adopted for the load forecasting problem, which include linear regression models, stochastic process models, exponential smoothing and ARMA models [12, 28, 73]. These methods model the load as a linear combination of its own past values and the exogenous input variables. They are relatively simple and when properly parametrized offer reasonable forecasting performance and interpretability for its parameters, giving the operators insight about the load behavior. However, the simplicity comes with a price, and as several studies report, without modifications these techniques usually display poor adaptability to changing conditions and unreliable performances when there unknown nonlinear dependencies.

ARMA load forecasting models can be converted to State space models and vice versa.

One difference between the two methods is that the state space formulation often allows a more concise presentation and manipulation. The state space load forecasting method has many variations, but they all model the load as a function of state variables. These models can be employed as the base of online and adaptive predictors, such as the Kalman filter, which adds robustness through the application of an internal noise model. Despite these advantages, state space models are not as common as ARMA models for load forecasting, probably because ARMA requires fewer explanatory variables and parameters [40], such as the difficult to estimate noise covariance matrices Q and R [74].

In the last decade, Artificial Neural Networks (ANNs) have received substantial attentions in load forecasting, with good performance reported in several papers [49, 61, 8, 9]. These techniques have the ability not only to learn the load series but also to model unspecified nonlinear relationship between load and the exogenous variables, being particulary effective at modeling weather effects [47, 29]. Recently, machine learning techniques and fuzzy logic approaches have also been used for load forecasting and achieved relatively good performances [32, 109].

In common, linear models, state space and most ANN and machine learning approaches share the generalized additive model as their mathematical base. It is important to highlight the origins of this model family in order to better understand the relationship between the different approaches usually employed in load forecasting.

3.2 Generalized Additive Models, Neural Networks and Linear Regression

According to the Kolmogorov Superposition Theorem (KST) [63], every continuous function f of n variables $x_1, x_2, ..., x_n$ can be represented as a superposition of continuous functions of one variable and the additive operation:

$$f\left(x_1, ..., x_n\right) = \sum_{q=0}^{2n} g_q \left(\sum_{p=1}^{n} \psi_{pq}\left(x_p\right) \right) \tag{3.1}$$

where g_q and ψ_{pq} are continuous univariate functions on $\mathbb{R}$ and every ψ_{pq} is independent of f.

KST has applications in various fields, such as non-linear control circuit and system theory, statistical pattern recognition, neural networks, image and multidimensional signal processing [64, 79, 67]. Unfortunately, though the theorem asserts the existence of this superposition form, it gives no tools for its construction. Certain constructive proofs exist [37], but they tend to require highly complicated functions, which are not suitable for modeling approaches due either to complexity or lack of interpretability.

An important subclass derived from the KST is the Generalized Additive Model (GAM) [45]. Dropping the outer sum in Eq. (3.1), at the cost of universal generality the model can be approximated by the simpler relationship shown in (3.2):

$$f\left(x_1, ..., x_n\right) = g\left(\sum_{p=1}^{n} \psi_p\left(x_p\right)\right) \tag{3.2}$$

Equivalently, (3.2) can also be written in the form shown in (3.3):

$$g^{-1}\left(f\left(x_1, ..., x_n\right)\right) = \sum_{p=1}^{n} \psi_p\left(x_p\right) \tag{3.3}$$

Though not every phenomenon could be approximated by a GAM, every phenomenon can be well approximated by a sum of GAMs. The choice of (3.2) or (3.3) is dictated by the existence of previous knowledge about the link function g or its inverse g^{-1} and whether it is easier to transform the raw variables x_p or the projections $f\left(x_1, ..., x_n\right)$.

The basic feed-forward Artificial Neural Network with one hidden layer can be obtained from (3.3) by means of a variable substitution and a fixed choice of g^{-1} and ψ_p. Making g^{-1} the identity function, choosing a logistic function for ψ_p and changing x_p by the linear combination of all raw variables plus a constant bias ν_p, substituting in (3.3) gives:

$$f\left(x_1, ..., x_n\right) = a_0 + \sum_{p=1}^{n} \kappa_p S_p\left(W_p X + \nu_p\right) \tag{3.4}$$

where κ_p is a constant and $W_p X = \left[\begin{array}{cccc} w_{p1} & w_{p2} & \cdots & w_{pn} \end{array}\right]\left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}\right]^T$. S_p represents the sigmoid function in the form $S\left(\delta\right) = \left(1 + \exp\left(-\delta\right)\right)^{-1}$.

When the projection represents the expectation of some observed quantity Y, equation (3.3) can be rewritten as (3.5), which is the standard formulation for the GAM:

$$g^{-1}\left(E\left(y\right)\right) = a_0 + \psi_1\left(x_1\right) + ... + \psi_n\left(x_n\right) \tag{3.5}$$

The backfitting algorithm is used to fit additive models [45]. It allows one to use an arbitrary smoother (e.g., spline, Logistic functions, Loess, kernel) to estimate the ψ_p and then find the optimum parameters to minimize modeling error. Due to its parametric nature, Neural networks can be trained by the backpropagation algorithm, an iterative fitting technique. It is similar to backfitting, albeit faster because smoothing is not required.

In the special case when ψ_p is linear, the resulting construction is called Generalized Linear Model (GLM), as written in (3.6).

$$g^{-1}\left(E\left(y\right)\right) = a_0 + a_1 x_1 + ... + a_n x_n \tag{3.6}$$

In the GLM, the sum of functions present in the GAM is substituted by a linear combination of the raw variables. The link function g^{-1} allows the response variable Y to have error distribution model different from the normal distribution. The GLM is a flexible generalization of ordinary least squares regression, allowing the linear model to be related to the response variable via a link function g. When modeling stochastic processes, the link function provides the relationship between the linear predictor $\theta = a_0 + a_1 x_1 + ... + a_n x_n$ and the mean $E(y)$ of the distribution function, usually from the exponential family. There are many commonly used link functions, and their choice is informed by several considerations [43]. The most common link functions are presented in Table 3.1:

Table 3.1: Common link functions of the Exponential family

Distribution	Inverse $g^{-1}\left(E(y)\right)$	Link $g\left(\theta\right)$	Link name
Normal	$\theta = E(y)$	$E(y) = \theta$	Identity
Gamma	$\theta = \frac{1}{E(y)}$	$E(y) = \frac{1}{\theta}$	Inverse
Inverse Normal	$\theta = \left(\frac{1}{E(y)}\right)^2$	$E(y) = \frac{1}{\sqrt{\theta}}$	Inverse squared
Poisson	$\theta = \ln\left(E(y)\right)$	$E(y) = \exp\left(\theta\right)$	Log
Bernoulli	$\theta = \ln\left(\frac{E(y)}{1-E(y)}\right)$	$E(y) = \frac{\exp(\theta)}{1+\exp(\theta)}$	Logit
Binomial (k terms)	$\theta = \ln\left(\frac{E(y)}{k-E(y)}\right)$	$E(y) = \frac{k\exp(\theta)}{1+\exp(\theta)}$	Logit
Geometric	$\theta = \ln\left(\frac{E(y)}{1+E(y)}\right)$	$E(y) = \frac{1}{1-\exp(-\theta)}$	Sigmoid

In the special case when g^{-1} is the identity function, the GLM turns into the simple multivariate linear regression, denoted in equation (3.7):

$$E\left(y\right) = a_0 + a_1 x_1 + ... + a_n x_n \tag{3.7}$$

This link function is optimum when the raw variables present gaussian distribution, and in this case the least squares estimator is also the optimum maximum likelihood estimator. In the real world, the true regression function is hardly ever linear, and thus the linear regression will always produce an linear approximation bias ϵ, even with an infinite amount of training data. This is the main drawback of the linear model. However, the main advantage of linear regression is that it reliably converges as more data is obtained. According to the derivations presented in [104], the rate of convergence for the estimation of a linear model with $n+1$ parameters as a function of the number of data points k is given by equation (3.8):

$$MSE = \sigma^2 + \epsilon^2 + O\left(k^{-1}\right) \tag{3.8}$$

where MSE denotes the mean square error of the regression, σ^2 is the intrinsic noise around the true regression function, ϵ^2 is the squared approximation bias, and $O\left(k^{-1}\right)$ is the estimation variance.

The estimation variance is inversely proportional to the number of data points, and as such tends to zero as k becomes very large. It is important to notice that the rate at which the estimation variance shrinks does not depend on the number of parameters. One can conclude that the MSE of the linear model can never be smaller than $\sigma^2 + \epsilon^2$, but will converge to it with sufficiently large k. In the rare case where the true regression function is linear, ϵ^2 is equal to zero and the MSE will converge to the intrinsic noise σ^2, which can be related in real world to the measurement errors.

Comparing the linear model with the convergence rate of a very general sum of GAM, important insights can be derived from the strengths and limitations of the more general models. Picking the kernel regression as an example, for this method the limiting approximation bias is actually zero, provided that a reasonable regression function is chosen. However, the fitting algorithm converge more slowly, because the data points must be used to optimize both the coefficients of a parametric model and the sheer shape of the regression function. Again according to [104], the rate of convergence for these models is given by (3.9):

$$MSE = \sigma^2 + O\left(k^{-4/(p+4)}\right) \tag{3.9}$$

There two differences between (3.8) and (3.9). Provided a reasonable kernel is selected, approximation bias in equation (3.9) can be equal to zero. This is the main advantage of the Kernel regression over the simpler linear model. However, the rate of convergence of the estimation variance is a function of the number of parameters present in Kernel's functions, in contrast with the parameter independent rate of convergence of the linear case. As the number of parameters increase, the nonparametric rate gets slower, and consequently the fully nonparametric estimate becomes imprecise for the same amount of computing effort, yielding the infamous curse of dimensionality.

The GAM offer a trade-off between these two extremes. Not every regression function is additive, so they generally have an approximation bias. But each ψ_p can be estimated by a simple one-dimensional smoothing, which converges at $O\left(k^{-4/5}\right)$, which is almost as fast as the linear case.

$$MSE = \sigma^2 + \left(\epsilon_{GAM}\right)^2 + O\left(k^{-4/5}\right) \tag{3.10}$$

Since linear models are a sub-class of additive models, $\epsilon_{GAM} \leq \epsilon$. Henceforth, GAM models are preferred when k is large, as in this condition the difference between $O\left(k^{-1}\right)$

and $O\left(k^{-4/5}\right)$ becomes smaller than the difference between ϵ_{GAM} and ϵ. In the condition where k is small, the decreased prediction bias does not offset the increased estimation variance the GAM has over the linear model. This is important for system of a very large number of input variables, as in this case the number of data points k must be much bigger in order to fit a reasonable model to the large number of inputs. Thus, linear models also have a slight advantage when the number of raw variables is large.

Linear regression is the one of most widely used statistical technique for electric load forecasting. Proposed methods of this type are usually used to model the relationship of load consumption and other factors such as weather, day type, and customer class. However, the presence of periodical load components, autocorrelation between consecutive days, nonlinearities and trends deserves special attention, as these effects and phenomenons must also be taken into account in the electric load model. It is possible to decrease the model bias of a linear regressor by means of extracting part of the nonlinearities from the model, however the increase in the number of input variables tend to increase the

The main objective of this work is to find a compromise between the simple linear model and a complex sum of GAM, such as an artificial neural network. Starting from the linear model baseline illustrated in Figure 3.2 column (a), the forecasting algorithm begins with a relatively small parameter estimation error $O(k)$ but a large model bias error due to the nonlinearities. In contrast, the baseline ANN has a small model bias error but a larger parameter estimation error, usually not large enough to offset the advantage over the linear regression as shown in Figure3.2 (e). It is important to note that estimation error is directly proportional to the number of model parameters and inversely proportional to the amount of training data, and it can be reduced by decreasing the number of parameters or increasing the training period in terms of number of time steps. The first option has the drawback of increasing model bias error, while the latter is not always feasible due to insufficiency in the gathered data or the time variant nature of the system.

In this work, the main strategy is to reduce both the linear bias error and the parameter estimation error by means of a selection of nonlinearly generated input variables and feature selection through principal component analysis.

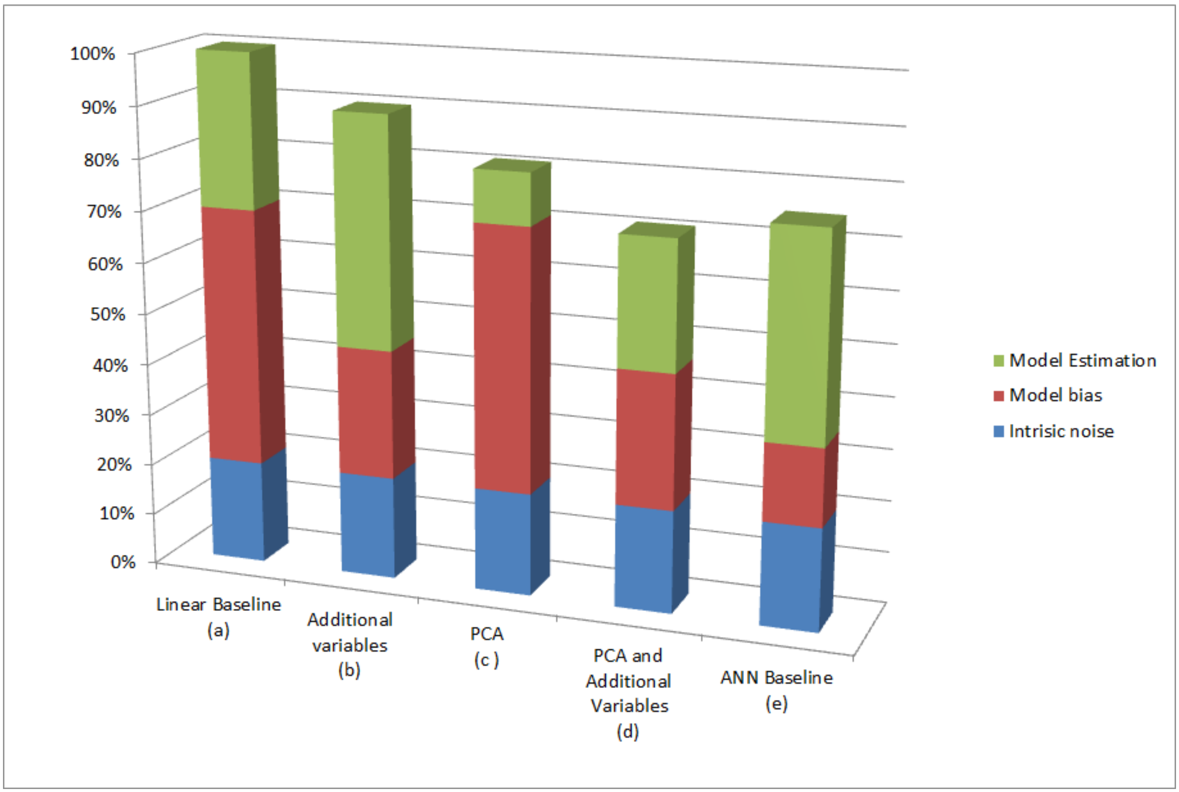


Figure 3.2: Comparison of the forecasting error among five scenarios. The intrisic noise, which is oftenly related to measurement errors, is independent of the forecasting algorithm and is constant among all cases.

Using the ample knowledge and data collected over decades about the nonlinear dependencies between electric load and PV generation to certain variables, the model bias error can be decreased by means of decoupling nonlinearities from the linear estimator, expanding Equation (3.7) to the form shown in Equation (3.11):

$$E\left(y\right) = a_0 + a_1 x_1 + ... + a_n x_n + ... + \psi_1\left(x_1\right) + ... + \psi_m\left(x_m\right) \tag{3.11}$$

However, as the number of parameters increases from n to $n+m$, the decrease in model bias is obtained at the cost of increased estimation error, as illustrated in Figure 3.2 column (b).

Principal component analysis is a tool for extracting features from a given input set, reducing dimensionality while maintaining a large fraction of the set's variance. The decreased number of dimensions reduces the number of input variables, which results in a lower parameter estimation error. However, some information is discarded in the process, slightly increasing the model bias error as shown in Figure 3.2 column (c).

The goal of this research is to carefully combine both approaches as illustrated in 3.2 column (d), employing additional variables and PCA feature selection to obtain a simple linear forecasting system that is competitive with more ellaborate model, such as ANN and support vector machines.

3.3 Electric load dynamics

Electric load time series can show several patterns accordingly to the types of customers connected the system. Residential, business, industrial and public energy consumers displays typical load patterns over a day, a week and a year. Also, external factors such as weather, demographics and economic output do influence the consumption of electricity. These patterns and dependencies have been documented for some locations [54, 21], and can usually be recognized by their load pattern over a day. For residential and commercial customers, load series show a strong seasonal behavior as well as dependence on local weather conditions. On the other hand, load series with an industrial profile are more irregular because the energy consumption is determined by operational decisions in a production or manufacturing facility. It is not unusual to have large industrial customers supplied by dedicated substations. To produce accurate forecasts for such industrial substations, it may be necessary to monitor have information regarding operational decisions taken by plant managers.

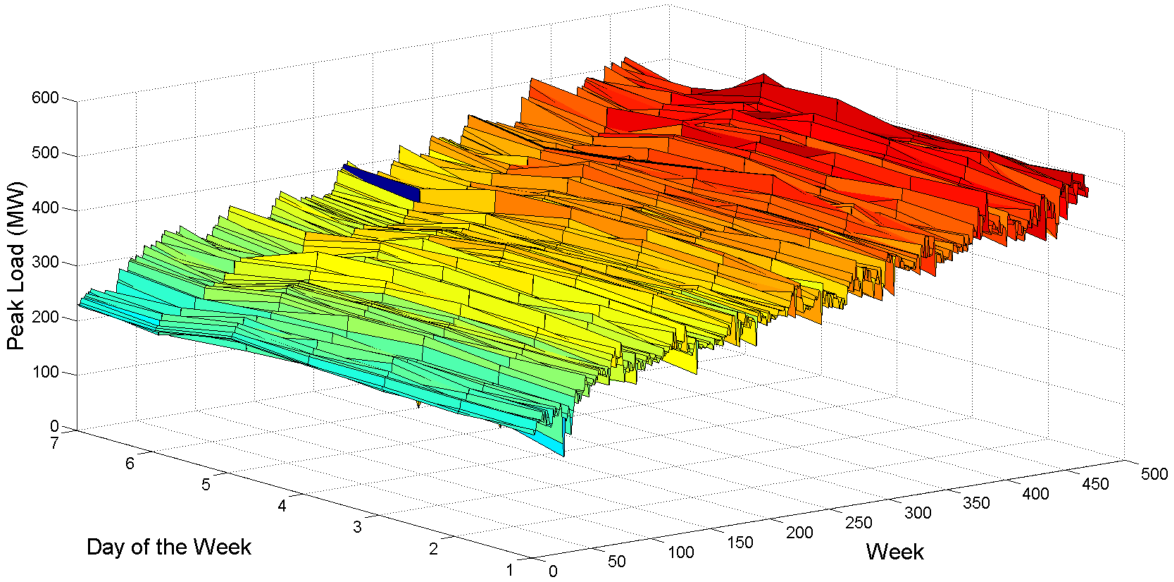


Figure 3.3: Peak load in Brasilia from 2001 to 2010. Weakly variation is visible in y axis (Day of the week), while the demand growth trend superposed to the seasonal variation is visible in the x axis (Week).

As well documented as these load patterns may be, however, the exact composition of residential, commercial, and industrial customers connected to the system is always changing and mostly unknown. For the type of load series under study, building a model for load forecasting must take into account trends and seasonal patterns at multiple levels. A growth trend, a winter-summer pattern and a weekly pattern are shown in figure 3.3. These patterns also interact with external variables that affect the load, such as weather fluctuations and tariff changes. When the weather is cold, there is a requirement for heating, which translates into an increase in energy demand. Hot days in summer trigger the use of air conditioning equipment, also increasing the demand. Power consumption behavior in holidays is markedly different from the workday patterns. The load on different weekdays also can behave differently. For example, Mondays and Fridays being adjacent to weekends, may have structurally different loads than Tuesday through Thursday. This is particularly true during the summer time. Holidays are more difficult to forecast than non-holidays because of their relative infrequent occurrence.

In order to work under these circumstances, linear time series based methods such as Box Jenkins models are based on the assumption that the data have an internal structure, such as autocorrelation, trend, or seasonal variation. These forecasting methods detect and explore such a structure. Box Jenkins approaches have been used for decades in the load forecasting field, in particular ARMAX (autoregressive moving average with exogenous variables) and ARIMAX (autoregressive integrated moving average with exogenous variables) are the most often used classical time series methods [33]. ARMA models are usually used for stationary processes while ARIMA is an extension of ARMA to nonstationary processes. ARMA and ARIMA use the time and load as the only input parameters. Since load generally depends on the weather and time of the day, ARIMAX is the most natural tool for load forecasting among the classical time series models.

If adaptability to changing conditions or recursive formulation is needed, an ARMA model can be converted to a State Space model in conjunction with a predicting technique, such as the Kalman filter [100, 99]. The state-space model provides a flexible approach to time series analysis, especially for ease in estimation and in handling missing values.

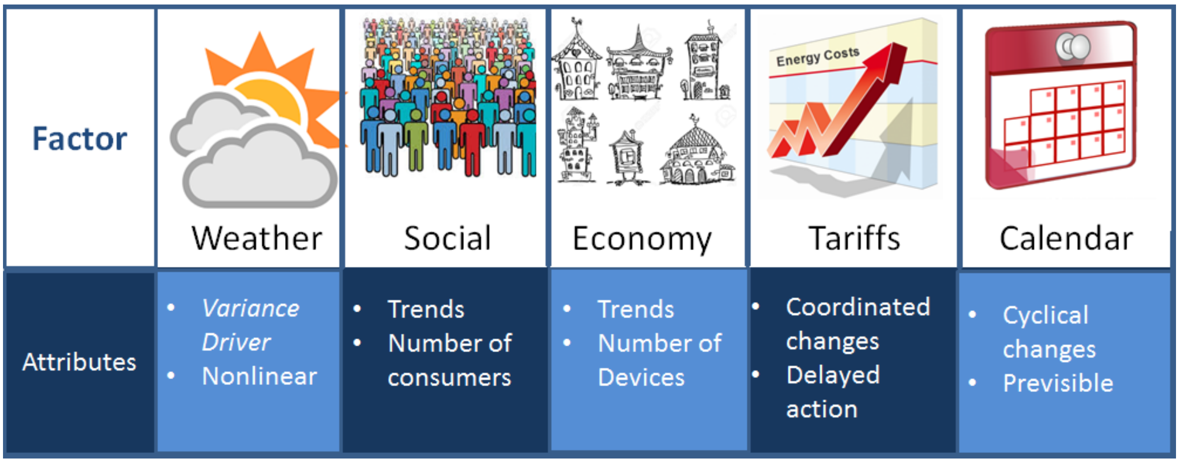
The use of Artificial Neural Networks has been a widely studied electric load forecasting technique since 1990 [84]. Neural networks are essentially nonlinear circuits that have the demonstrated capability to do nonlinear curve fitting. In applying a neural network to electric load forecasting, one must select one of a number of architectures (e.g. Hopfield, Multilayer Perceptron, Boltzmann machine), the number and connectivity of layers and elements, use of bi-directional or uni-directional links, and the number format (e.g. binary or continuous) to be used by inputs and outputs. Thus, Neural Networks are well suited to provide electric load forecasting with none or little modifications in their basic formulation. The most popular artificial neural network architecture for electric load forecasting is Multilayer Perceptron (MLP), whose formulation is a linear combination of the artificial neurons described in equation (3.4).

Back propagation is a supervised training algorithm for MLP neural networks. The learning step is a phase where the actual numerical parameters assigned to element inputs are determined by matching historical data (such as time and weather) to desired outputs (such as historical electric loads) in a pre-operational training session. In general, ANN offer great adaptability and native nonlinear fitting support, at the cost of higher computational complexity than a linear model with the same number of parameters and higher susceptibility to the curse of dimensionality, an effect that limits the ANN precision if a large number inputs are needed. Also, even in the simpler ANNs, their internal parameters can sum up to a very large number, thus giving this kind of model a Black Box characteristic, with very little system insight or interpretability for the user.

Despite the advantages ANN have over linear and State Space approaches, carefully designed linear models still have an advantage in forecasting performance [47], mainly due to the large amount of data used and precise adjustment formulas built in the models for discovering nonlinearity patterns [31]. Variants of linear models are still being perfected and employed by system operators in order to forecast the electric load [20]. These advanced linear approaches mostly feature artificially produced input variables based on weather, tariffs and calendar. Knowledge about the variables and factors that influence the electric system load are thus essential to build a robust and precise linear forecasting method.

3.4 Factors influencing electric load

Several factors are known to affect energy demand: temperature, climate events, electricity tariffs, demographic indicators, economic indexes, social conventions and cultural traditions. A succint illustration of these main factors and their main effects over electricity demand are shown in Figure 3.4.



Factor	Weather	Social	Economy	Tariffs	Calendar
Attributes	• *Variance Driver* • Nonlinear	• Trends • Number of consumers	• Trends • Number of Devices	• Coordinated changes • Delayed action	• Cyclical changes • Previsible

Figure 3.4: Factors and their main effects over the electricity demand.

Worldwide, with the ample access to information technology, a diversity of data can be collected and substantial volumes of time series related to electricity demand can be processed. In Subsection 3.4.1, weather forecasts are presented, while in Subsection 3.4.2, socioeconomic variables are shown. In Subsection 3.4.3 and 3.4.4, the variables related to energy tariffs and calendar-weather events are described, respectively. In Subsection 3.4.5, the derivations discussed in the previous subsections are joined in for the load model proposed in this dissertation.

3.4.1 Weather Variables

The influence of weather on electricity consumption is a topic of research since the first half of the 20th century. In the discussion presented in [35], the impact of changing weather conditions over the South East England power system was presented, stressing the effects of decreased temperature over mean and peak load.

The effects of weather on load are usually modeled by expressing the load as a regression of explanatory meteorological factors such as temperature, wind speed, humidity, and others [83]. Although it is recognized that an extremely wide variety of explanatory weather variables is required to totally represent the effects of weather, studies have shown that a few basic meteorological factors usually account for most of the weather-dependent load. Furthermore, temperature do affect the electrical properties and efficiency of semiconductor devices, meaning that it is also a relevant factor when the energy source is connected to the grid through inverters, such as small wind power; or when it is entirely based on semiconductor devices, such as a solar photovoltaic system.

The specific weather variables that are normally used to model weather-dependent load are temperature, wind speed, humidity, and daylight illumination. The latter is usually the least significant of these weather variables, and because its metering is difficult and costly, it is usually omitted from most models [91, 74]. In this dissertation, an alternative methodology to estimate sunlight and natural illumination is presented. For forecasting photovoltaic generation, the main driver is the amount of global solar irradiation arriving at the panel, with temperature being a second order factor.

The history of weather data can be collected from the METeorological Aerodrome Reports (METAR), as measured in airports located near the load centers. METAR is the primary observation code used in the United States to satisfy requirements for reporting surface meteorological data [80]. It has worldwide adoption and most aerodrome stations provide their reports online at specialized internet sites. A METAR contains a report of wind, visibility, runway visual range, present weather, sky condition, temperature, dew point, and altimeter setting collectively referred to as “the body of the report”. The report presents hourly information about the weather variables, as measured or observed in the surface. The reports also feature measurements of cloud cover and indicative codes for weather phenomena such as fog, rain, thunderstorms and snow. Table 3.2 lists the 22 standard measurements and observations presented in a METAR.

Table 3.2: List of METAR's measurements and observations

ID	Measurement/Observation	Unit
1	Maximum Temperature	Celsius
2	Mean Temperature	Celsius
3	Minimum Temperature	Celsius
4	Maximum Dew Point	Celsius
5	Mean Dew Point	Celsius
6	Min Dew Point	Celsius
7	Maximum Relative Humidity	Adimensional (%)
8	Mean Relative Humidity	Adimensional (%)
9	Minimum Relative Humidity	Adimensional (%)
10	Maximum Pressure at Sea Level	hPa
11	Mean Pressure at Sea Level	hPa
12	Minimum Pressure at Sea Level	hPa
13	Maximum Visibility	km
14	Mean Visibility	km
15	Minimum Visibility	km
16	Maximum Wind Speed	km/h
17	Mean Wind Speed	km/h
18	Maximum Wind Shear	km/h
19	Precipitation	mm
20	Cloud Cover	Octas
21	Events	(Fog, Rain, Snow, Thunderstorm)
22	Wind Direction	Degrees

The effects of temperature, humidity, wind speed, solar irradiation and weather events over electric load are different and as such require to be accounted in specific ways. The remainder of this subsection is further divided in subsubsections aimed at discussing the effects of each type of weather variable. Subsubsection 3.4.1.1 explains how temperature affect the electricity demand, as well as explain the concept of degree days. Subsubsection 3.4.1.2 deals with the effects of humidity and the concept of latent heat load, which is parametrized by Enthalpy degree days. Subsubsection 3.4.1.3 presents a formulation for the effect of wind speed and direction in the heat flow over buildings walls. Subsubsection 3.4.1.4 presents the modeling employed to estimate the natural illumination over buildings and open spaces, while subsubsection 3.4.1.5 deals with the heating effects over buildings that is caused by the Sun‘s irradiance.

3.4.1.1 Temperature, Log-Temperature, Heating and Cooling Degree Days

There are several types of electrical devices whose power consumption is temperature dependent. Common examples comprise refrigerators, heaters/ovens and Heating, Ventilation and Air Conditioning equipment (HVAC).

Refrigerators and freezers are ubiquitious at both residences and industries. Their load is dependent of the indoor temperature, and usually they have thermal controls that keep the refrigerator interior just above 0 Celsius, while the freezer set point is below freezing point, around negative 25 Celsius. The power needed to keep this temperature generally adds to the base load of a electric system. Each time the appliance door is opened, the loss of cold air to the environment temporally increases the power consumption, adding an user dependent and random component to the device's demand.

Heaters and ovens, similar to refrigerators, have their energy consumption linked to the interior temperature. Usually they are employed to add heat to a medium or recipient until a preset temperature is reached. Examples include electric boilers, showers and ovens. In Brazil, water heating by electric showers is prevalent. This load behavior require higher reference temperatures, as water has a far higher thermal conductivity than air and requires to be warmer to ensure comfort.

HVAC devices are employed to keep a comfortable thermal environment inside a building. The average human being feels thermal comfort in a narrow range of temperature and humidity, as shown in figure 3.5. Through it is possible to keep a building in the comfort zone by means of energy efficient architecture and special operational strategies, at temperature and humidity extremes the use of conventional heating, air conditioning and humidity controls is required. The building insulation quality dictates the amount of power needed to keep comfortable conditions once the set points are reached. Human controlled devices tend to be activated while out of comfort zone, then deactivated only when exiting the comfort envelope at the opposite extreme.

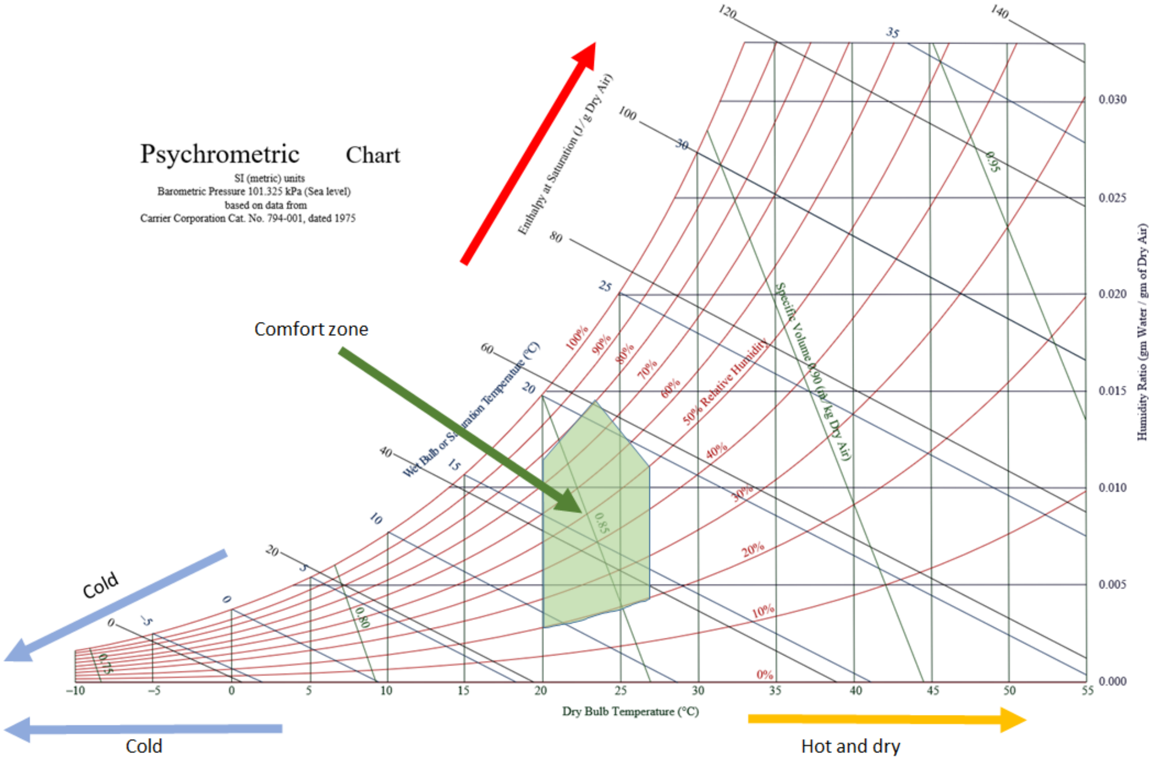


Figure 3.5: This chart is a summary of the human comfort zone as a function of ambient conditions (weather and climate). Modified from the original in Wikimedia Commons, license CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0), acessible at https://commons.wikimedia.org/wiki/File%3APsychrometricChart.SeaLevel.SI.svg.

Considering the temperature measurements present in a METAR, a first attempt at modeling the temperature driven component y_T of the load can be constructed by the simple expression shown in (3.12):

$$y_T = b_{T1}T_{Avg} + b_{T2}T_{Min} + b_{T3}T_{Max} \tag{3.12}$$

where T_{Avg}, T_{Min} and T_{Max} denotes average, minimum and maximum temperature, respectively.

However, the load demand response to temperature is known to behave nonlinearly, specially at cold and hot extremes [72, 68]. This can be seen in Fig. 3.6, a scatter plot of mean daily temperature and power demand. Dispersion increases at lower temperatures, while in this example there is a possible inflection point above 20 degrees Celsius. In order to transform raw weather variables such as temperature and humidity in more meaningful inputs for a forecasting algorithm, some pre-processing is needed. In this dissertation, both logarithm and degree days parametrizations are employed.

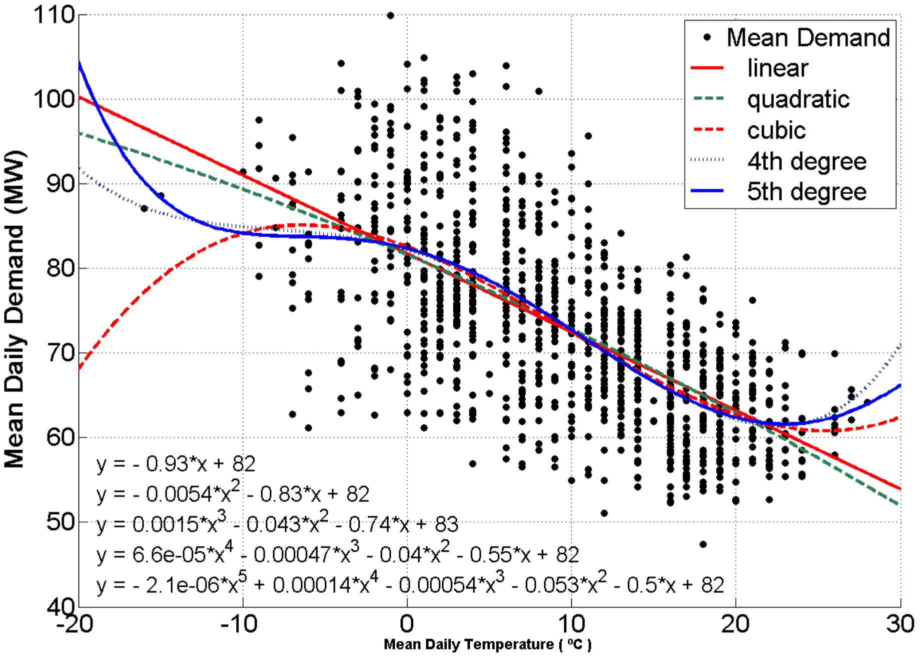


Figure 3.6: Scatter plot of mean temperature and mean demand in Megawatts (MW). Five polynomial best fitting showcases the large residuals and diverse possibilities from nonlinear behavior.

The logarithm of the mean temperature can be calculated by simply converting the value in Celsius to Kelvin, and then calculating the natural logarithm of the Kelvin value. The logarithm transformation model the transient response of the human skin to a rapid change in temperature. A common finding in many studies of thermal sensation thresholds is that despite the variability in thresholds across the body, all regions are more sensitive to cold than to warmth. In general, the threshold for detecting a decrease in temperature (cold) is half that of detecting an increase in skin temperature (warmth), and the better a site is at detecting cold, the better it is at detecting warmth [94]. The logarithm function mimics part of this sensory characteristic by nonlinearly reducing the amplitude variation at the warmer extremes. Fig. 3.7 depicts the sequence of daily results for 1095 data points of average temperature collected in Leipzig.

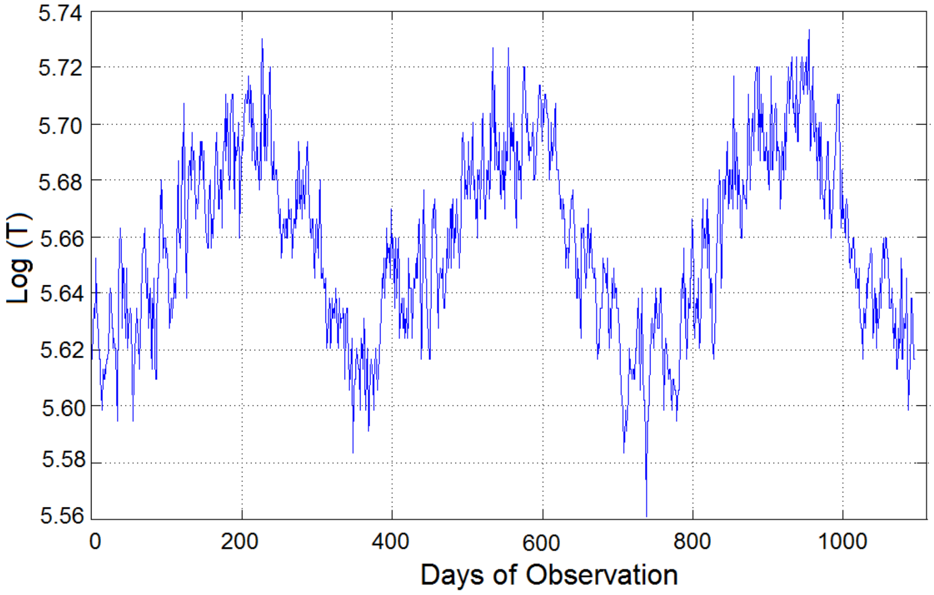


Figure 3.7: Logarithm of the absolute temperature, as measured in Leipzig from 2001 to 2003

Updating the temperature-driven load model gives (3.13):

$$y_T = b_{T1}T_{Avg} + b_{T2}T_{Min} + b_{T3}T_{Max} + b_{T4}\ln\left(T_0 + T_{Avg}\right) \tag{3.13}$$

where T_0 denotes the absolute zero at the employed temperature scale.

The concept of degree-day relates to the necessity of finding a variable that measures the amount of energy needed to heat or cool a building to a comfortable temperature, given the external temperature. Since the last quarter of the 20th century, degree-days are used as tool for energy consumption forecast, as showcased in [93, 27]. Currently, heating (HDD) and cooling degree-days (CDD) have been featured in several load forecast methods, such as [72, 29].

The Heating Degree-Days (HDD) is a measure of the severity and duration of cold weather, which relates to the heating requirements. HDD is defined as the integral sum of the subtraction between a given reference heating temperature (not much different than choosing a temperature set point for a heater) and the ambient temperature over time. Since this continuous time definition is not compatible with the daily temperature time series collected in METARs, the HDD values are estimated by the United Kingdom Meteorological Office (MET Office) method [102], that is simpler yet reasonably accurate [23] and only requires minimum and maximum temperatures. Table 3.3 presents the Estimation Formulas for each condition.

Table 3.3: Approximate calculation of the heating degree-days

Condition	Estimation Formula
$T_{min} > T_{ref}$	$HDD = 0$
$\frac{T_{Max}+T_{Min}}{2} > T_{ref}$	$HDD = \frac{T_{Ref}-T_{Min}}{4}$
$T_{Max} \geq T_{ref}$	$HDD = \frac{T_{Ref}-T_{Min}}{2} - \frac{T_{Max}-T_{Ref}}{4}$
$T_{Max} < T_{ref}$	$HDD = T_{ref} - \frac{T_{Max}+T_{Min}}{2}$

Figure 3.8 shows the resultant daily values for the 1095 days of years 2001-2003. The value of 18°C is chosen in this example as the heating reference temperature.

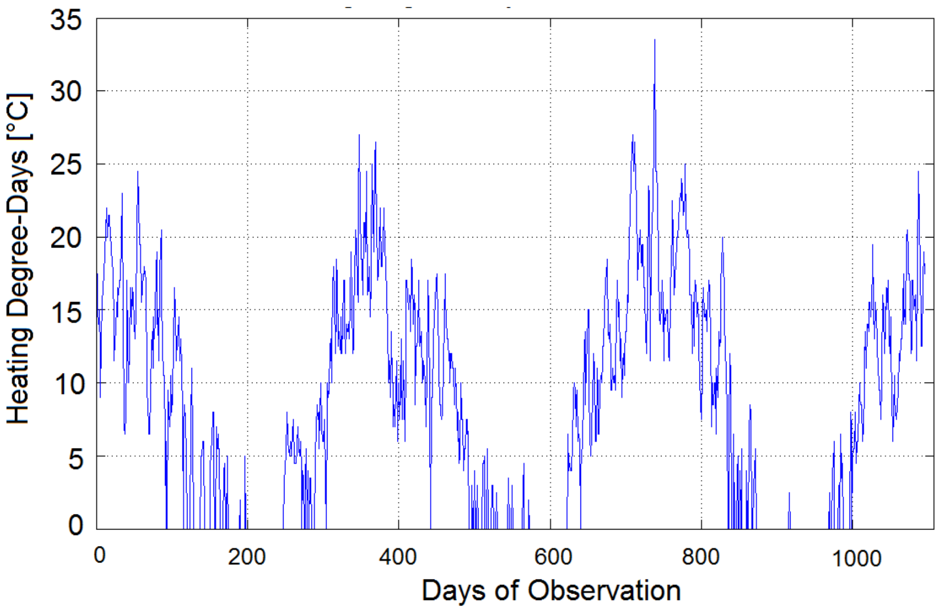


Figure 3.8: Heating Degree Days at 18 Celsius reference, as measured in Leipzig from the 2001 to 2003. The heating peaks are measured in the winter season.

The relationship between the HDD and the electric demand is stronger than the unprocessed temperature value. Fig. 3.9 presents a scatter plot that illustrates the correlation between the two variables.

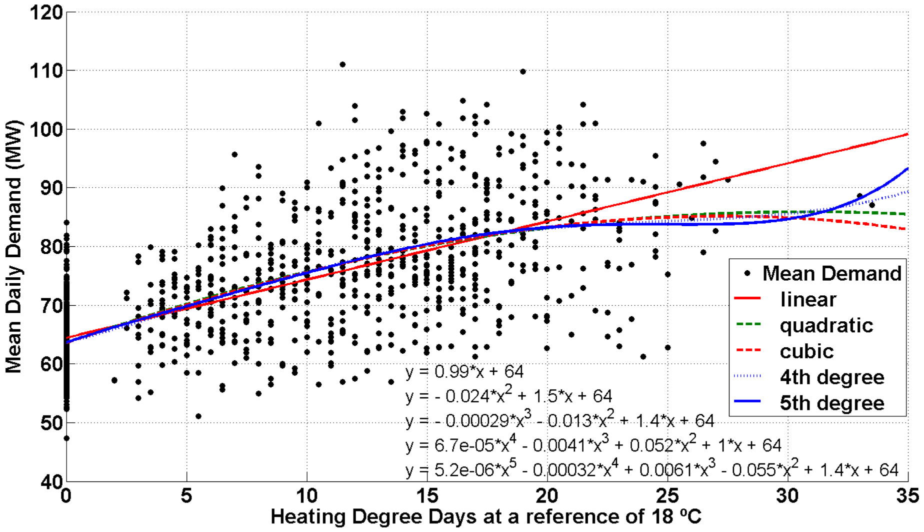


Figure 3.9: Scatter plot of HDD and mean demand in MW, as measured in Leipzig from the 2001 to 2003. Five polynomial best fitting curves showcase the more straightforward dependence, yet nonlinear.

However, in hot climates such as Brasilia's, HDD requires a different parametrization than those performed to model building heating demand. In Brazil, water heating by electric showers is prevalent. Compared to ambient heating, which is prevalent in Leipzig, water heating require higher reference temperatures, as water has a far higher thermal condutivity than air and must be kept closer to the human body temperature to not cause discomfort. Subtropical and mildly temperate climates may need HDD parametrizations for both ambient and water heating (WHDD).

Similarly, the Cooling Degree-Days (CDD) is a measure of the severity and duration of hot weather, which relates to the cooling requirements. CDD was also estimated by the United Kingdom Meteorological Office method [102], as depicted in Table 3.4.

Table 3.4: Approximate calculation of the cooling degree-days

Condition	Estimation Formula
$T_{max} < T_{ref}$	$CDD = 0$
$\frac{T_{Max}+T_{Min}}{2} < T_{ref}$	$CDD = \frac{T_{Max}-T_{Ref}}{4}$
$T_{Min} \leq T_{ref}$	$CDD = \frac{T_{Max}-T_{Ref}}{2} - \frac{T_{ref}-T_{Min}}{4}$
$T_{Min} > T_{ref}$	$CDD = \frac{T_{Max}+T_{Min}}{2} - T_{ref}$

The relationship between the CDD and the electric demand is also stronger than the unprocessed temperature value. Fig. 3.10 presents a scatter plot that illustrates the correlation between the two variables.

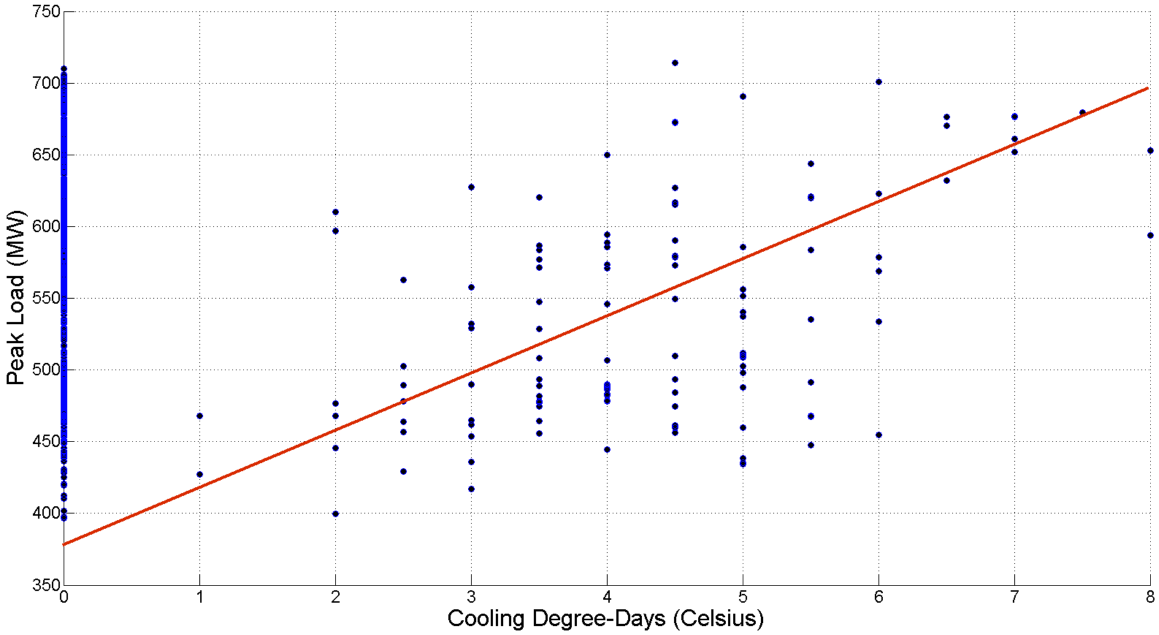


Figure 3.10: Scatter plot of CDD and peak demand in MW, as measured in Brasília from 2001 to 2003. Because the large number of noncorrelated points stays at zero degree-days, they do not affect the determination of the CDD coefficient.

In a load forecasting system, multiple CDD and HDD variables can be created, each one with different reference temperatures. In combination, these multiple variables model the nonlinear relationship of cooling and heating requirements to the electric demand as a piecewise linear function. Adding this variables to the regression presented in (3.13) gives expression (3.14):

$$\begin{aligned} y_T = b_{T1}T_{Avg} + b_{T2}T_{Min} + b_{T3}T_{Max} + b_{T4}\ln\left(T_0 + T_{Avg}\right) \\ + b_{T5}\vartheta_{CDD} + b_{T6}\vartheta_{HDD} + b_{T7}\vartheta_{WHDD} \end{aligned} \tag{3.14}$$

$$y_T = \begin{bmatrix} b_{T1} & b_{T2} & b_{T3} & b_{T4} & b_{T5} & b_{T6} & b_{T7} \end{bmatrix} \begin{bmatrix} T_{Avg} \\ T_{Min} \\ T_{Max} \\ \ln\left(T_0 + T_{Avg}\right) \\ \vartheta_{CDD} \\ \vartheta_{HDD} \\ \vartheta_{WHDD} \end{bmatrix} \tag{3.15}$$

$$y_T = B_T U_T \tag{3.16}$$

where ϑ_{CDD} denotes the cooling degree days variables, ϑ_{HDD} the heating degree days and ϑ_{WHDD} the water heating degree days. Not shown in (3.14), several CDD or HDD variables can be employed, requiring only different reference temperatures. Also, some of these variables can be discarded if the climate is very warm or cold.

3.4.1.2 Humidity and Enthalpy Degree Days

Atmospheric humidity is a measure of water held in the air as a gas. Water can be solid (ice), liquid (water) or a gas (vapor). The vapour component makes up about 99% of all water held in the atmosphere. Relative Humidity (RH) is the most common measure of humidity. It measures how close the air is to being saturated - that is how much water vapor there is in the air compared to how much there could be at that temperature. Figure 3.11 shows the maximum water vapor content the air can carry accordingly to its temperature. Warmer air can hold more water vapor because there is more energy available. If the RH of the air is 100% then it is fully saturated.

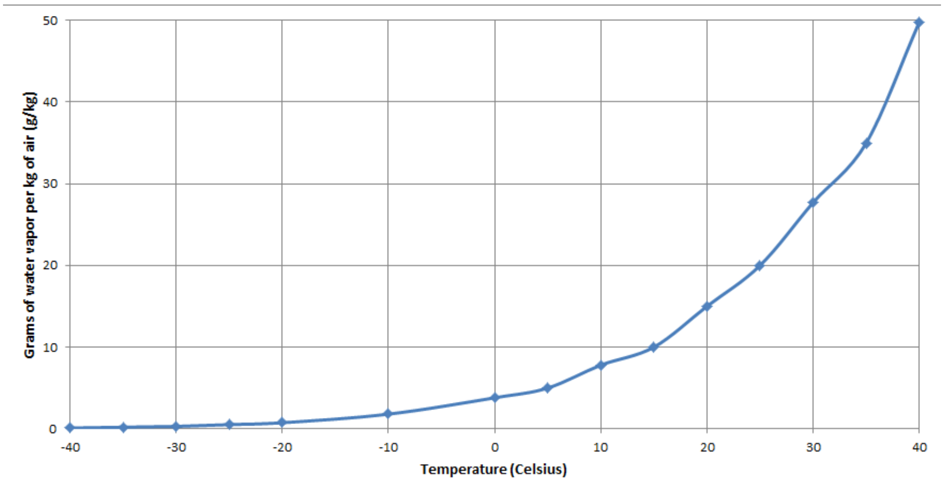


Figure 3.11: Maximum water vapor content of air as a function of temperature. This is the reference to calculation of Relative Humidity.

In warm temperatures, air with very high RH is very uncomfortable, as the saturated air affects the human body cooling mechanism. The air cannot easily absorb more water vapor and so cannot effectively evaporate the sweat from the skin. In cool temperatures, air with very high RH can make humans feel cooler. This is because there is more water vapor in contact with skin and since vapor is a much better heat conductor than dry air, there is greater heat flux from the body to the atmosphere, giving the cold sensation. As many HVAC loads are human operated, the change in sensation caused by humidity can lead to activation or deactivation of electric devices. Considering the humidity measurements present in a METAR, a first attempt at modeling the moisture driven component y_H of the load is given by (3.17):

$$y_H = b_{H1}Hum_{Avg} + b_{H2}Hum_{Min} + b_{H3}Hum_{Max} \tag{3.17}$$

where Hum_{Avg}, Hum_{Min} and Hum_{Max} denotes average, minimum and maximum relative humidity, respectively.

Added to the change of human perception, the latent heat present in humid environment is greater than in dry air, increasing energy requirements to either cool or heat this environment. Moist air is a mixture of dry air and water vapor. Consequently, the enthalpy of humid air includes the sensible enthalpy of the dry air and the latent enthalpy of the evaporated water. The total enthalpy - sensible and latent - is used when calculating cooling and heating processes. If there is no non-expansion work on the system and the pressure is still constant, then the change in enthalpy is equal to the heat consumed or released by the system.

Enthalpy latent days (ELD) indicates the amount of energy required to remove excessive moisture from the outdoor air without reducing the indoor air temperature, but lowering the indoor humidity to an acceptable level [88]. Eq. (3.18) defines Enthalpy latent days as the summation of positive enthalpy differences between the outdoor air enthalpy h_0 with relative humidity x_0, and enthalpy h_b with indoor reference relative humidity x_b. Both enthalpies at the outdoor air temperature θ_0. The reference humidity is set in this load forecasting system as 50%, which is standard in most air conditioning equipment:

$$\mathrm{ELD} = \sum_{t=1}^{24} \left[h_0 \left(\theta_0, x_0[t] \right) - h_b \left(\theta_0, x_b \right) \right] \tag{3.18}$$

Where h_0 and h_b are calculated by the expressions [71]:

$$h_0 = 1,007\theta_0 + M_0 \left(2501 + 1,84\theta_0 \right) \tag{3.19}$$

$$h_b = 1,007\theta_0 + M_b \left(2501 + 1,84\theta_0 \right) \tag{3.20}$$

M_0 and M_b are the water vapor concentration (kg/kg) at relative humidity x_0 and x_b, respectively. They are calculated from the saturated water vapor concentration in air M_s and from humidity degree days $HuDD$.

$$M_0 = M_s \frac{HuDD}{100} \tag{3.21}$$

$$M_b = M_s \frac{x_b}{100} \tag{3.22}$$

Humidity degree days are calculated with the MET office method, using the reference x_b, minimum x_{Min}, and maximum relative humidity x_{Max}, as shown in Table 3.5:

Table 3.5: Approximate calculation of the heating degree-days

Condition	Estimation Formula
$x_{min} > x_b$	$HuDD = 0$
$\frac{x_{Max}+x_{Min}}{2} > x_b$	$HuDD = \frac{x_b - x_{Min}}{4}$
$x_{Max} \geq x_b$	$HuDD = \frac{x_b - x_{Min}}{2} - \frac{x_{Max} - x_b}{4}$
$x_{Max} < x_b$	$HuDD = x_b - \frac{x_{Max}+x_{Min}}{2}$

The saturated water vapor concentration in air $M_s$is obtained from air pressure P and water vapor partial pressure P_S at temperature θ_0:

$$M_s = 0.62198 \frac{P_S}{P - P_S} \tag{3.23}$$

The partial pressure is calculated from the temperature θ_0:

$$PS = 610.78 \exp\left(\frac{17,2694\theta_0}{238,3 + \theta_0}\right) \tag{3.24}$$

The humidity driven load expression denoted in (3.17) can then be complemented with th ELD variable, giving the final expression for the humidity related load in this model:

$$y_H = b_{H1}Hum_{Avg} + b_{H2}Hum_{Min} + b_{H3}Hum_{Max} + b_{H4}ELD \tag{3.25}$$

$$y_H = \left[\begin{array}{cccc} b_{H1} & b_{H2} & \cdots & b_{H4} \end{array}\right] \left[\begin{array}{cccc} Hum_{Avg} & Hum_{Min} & Hum_{Max} & ELD \end{array}\right]^T \tag{3.26}$$

$$y_H = B_H U_H \tag{3.27}$$

Similarly to the HDD and CDD parametrization of temperature, multiple ELD variables can be employed, specially if the forecasting area has several humidity controlled spaces with custom moisture levels, such as clean rooms for electronics manufacturing or biopharmaceutical research.

3.4.1.3 Wind speed and convection heat transfer

The energy performance of HVAC equipment depends on the heat transfer coefficients existent between the interior controlled enviroment and the exterior ambient. These coefficients dictate how much the HVAC must absorb or emit heat to keep indoor conditions at the programmed setpoints. The temperature, heating and cooling degree day inputs are mostly related to the conduction heat flux. However, when the incident wind speed is considerable, the effects of the forced convection can become the dominant mode of heat flux. As a consequence, the wind can have observable effects in electricity consumption.

By definition, the heat transfer is defined by expression (3.28):

$$\dot{q} = h_c S(T - T_0) \tag{3.28}$$

where the heat transferred per unit time $\dot{q}$ is a function of the convective heat transfer coefficient h_c, the contact area S and the diference between T and T_0, respectively the temperatures of the object and the fluid.

For the forced convective flow regime, the convection coefficient is usually correlated to the wind speed at a reference location. Usually, the mean wind speed in the undisturbed flow at a height of 10 m above the ground is employed, which is the standard arrangement for weather station anemometers. The wind speed correlation is mostly reported as linear or power-law correlations in papers [25], whose studies are both based on measurements, wind tunnel simulations and Computational Fluid Dynamics (CFD).

The convection coefficient is also dependent of the wind direction [70]. Not only the contact surface shape and area change accordingly to the wind direction, as also the landscape is hardly symetrical and near obstacles can further change the effective shape and area either by shadowing or concentrating the air flux in the contact surface. In the impossibility of mapping the coefficient at every recorded direction, it is possible to decompose the mean wind speed and wind direction information contained in the METAR in four directional wind inputs, arbitrarily chosen to be aligned with the cardinal points.

To model such dependencies, the directional wind inputs are transformed by means of power laws to provide the convection coefficients h_c for the heat transfer modelling, which is shown in Eqs (3.28) and (3.29).

$$h_c = \rho \left(v\right)^{\alpha} \tag{3.29}$$

where this power law approximation relates the coefficient h_c to the air speed v_w raised to power of α, and ρ is a proportionality constat. Literature reports a narrow range for exponents, usually between 0.8 and 0.9. In close agreement with the results found in [25, 30, 17], the exponent $\alpha = 0.82$ is chosen to create four additional variables to model heat convection on facades oriented to each compass point. The resulting model for the wind dependent load becomes:

$$y_v = b_{v1}\left(v_N\right)^{0.82} + b_{v2}\left(v_S\right)^{0.82} + b_{v3}\left(v_E\right)^{0.82} + b_{v4}\left(v_W\right)^{0.82} \tag{3.30}$$

where v_N, v_S, v_E and v_W denotes the wind speed component at North, South, East and West directions, respectively.

The effect of natural ventilation can also be important. Exponent $\alpha = 2$ is empirically employed in maximum wind and average wind nondirectional inputs to model human confort psycometric functions. This exponent is also related to distributed wind turbines, which if existent do act as a negative load. The final equation is given by (3.31):

$$y_v = b_{v1}\left(v_N\right)^{0.82} + b_{v2}\left(v_S\right)^{0.82} + b_{v3}\left(v_E\right)^{0.82} + b_{v4}\left(v_W\right)^{0.82} + b_{v5}\left(v_{Avg}\right)^{2} + b_{v6}\left(v_{Max}\right)^{2} \tag{3.31}$$

$$y_v = \begin{bmatrix} b_{v1} \\ b_{v2} \\ b_{v3} \\ b_{v4} \\ b_{v5} \\ b_{v6} \end{bmatrix}^T \begin{bmatrix} \left(v_N\right)^{0.82} \\ \left(v_S\right)^{0.82} \\ \left(v_E\right)^{0.82} \\ \left(v_W\right)^{0.82} \\ \left(v_{avg}\right)^{2} \\ \left(v_{Max}\right)^{2} \end{bmatrix} = B_v U_v \tag{3.32}$$

where v_{Avg} and v_{Max} represent the nondirectional average and maximum wind speeds obtained from METAR.

3.4.1.4 Natural illumination

The lighting load is a consequence of the human need for illumination, in order to enhance task safety and performance, improve the appearance of an area or have positive psychological effects on its occupants. Individual lighting systems can be controlled by several means, from simple switches to complex automated systems. There are also lighting systems that are kept permanently lit, such as signaling and some emergency lights in escape routes.

The main driver of lighting load is the amount of sunlight reaching a given area. Below a certain threshould, electric lights are activated in order to provide artificial illumination. Generally, lights are kept off during the day and activated in the night. Cloud cover can reduce sunlight during daytime, leading increasing demand for lighting. Lack of daylight, however, is not the sole factor influencing the lighting load. Due to lack of natural illumination, rooms in large buildings require lights to be activated during the entire work hours. Lights are also employed as a mean to emphasize business signs, outdoor ads and to aesthetically improve the appearance of monuments and buildings.

In order to model the natural illumination, one must account for the amount of global clear sky irradiation arriving at a surface, after being filtered by the cloud cover and by the human eye sensivity, the latter strongly dependent on wavelength. The global clear sky irradiation arriving at a surface can be calculated with the SPCTRL2 model. The SEDES2 cloudy sky model is then applied to account for the effect of clouds over natural lights. More details about the SPCTRL2 and SEDES2 models are presented in Appendix B.

The determination of the illumination surfaces, however, is a complex task. Daylight has a ample access to outdoor areas, while inside buildings the specific size and layout of windows and translucid ceiling apparattus is determinant to the amount of natural light received. Furthermore, the internal room layout, room height, wall and floor colors and furnitures do affect the internal reflection of natural light. Thus, the exact modeling of natural illumination inside a single building is a complex task. A simpler approach is needed in order to obtain a feasible load model.

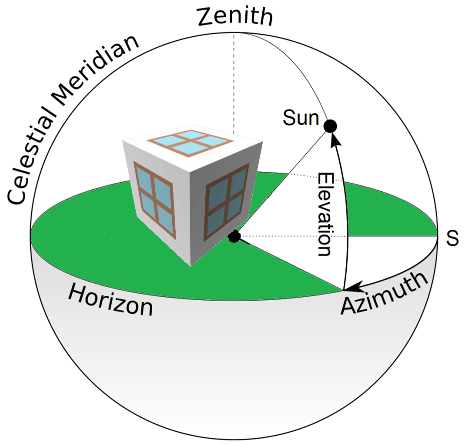


Figure 3.12: Natural illumination model. The five “window” surfaces are used to approximate the sunlight incident into buildings and open spaces. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/commons/f/f7/Azimuth-Altitude_schematic.svg

The objective is not to model the natural lighting in a given building, but in a very large number of buildings and outdoor areas. Certainly, every single building has unique shape, windows, walls, ceilings and interior details. However, for a very large set of structures, these variations tend to cancel out and an average value for the daylight can be calculated, yielding a much simpler approximation for the natural lighting modeling. Only the effect of the sun direction must then be accounted. This is accomplished by approximating the total incident daylight by the sum of five components: one perpendicular to the surface, pointed to the zenith, and four paralel to the surface, each pointed to a compass point. The sunlight dependent component of the load Y_ϕ can then be approximated by the linear relationship shown in equation (3.33):

$$y_\phi = b_{\phi 1}\phi_N + b_{\phi 2}\phi_S + b_{\phi 3}\phi_E + b_{\phi 4}\phi_W + b_{\phi 5}\phi_Z \tag{3.33}$$

$$y_\phi = \begin{bmatrix} b_{\phi 1} & b_{\phi 2} & b_{\phi 3} & b_{\phi 4} & b_{\phi 5} \end{bmatrix} \begin{bmatrix} \phi_N & \phi_S & \phi_E & \phi_W & \phi_Z \end{bmatrix}^T \tag{3.34}$$

$$y_\phi = B_\phi U_\phi \tag{3.35}$$

where ϕ_N, ϕ_S, ϕ_E, ϕ_W and ϕ_z denotes the illuminance component at North, South, East, West and Zenith (Up) directions, respectively.

However, as the load component y_ϕ is a function of illuminance (ϕ), the SPCTRL2 and SEDES2 have outputs measured in irradiance. Illuminance represents the light power of a source incident over a surface as perceived by the human eye. The sky irradiance models only deal with the absolute incident power density, measured in Watts per square meter. In order to model the human perception, the human eye standard sensivity is applied to transform the radiometric (energy related) units obtained from SPCTRL2 and SEDES2 to photometric units (perception related).

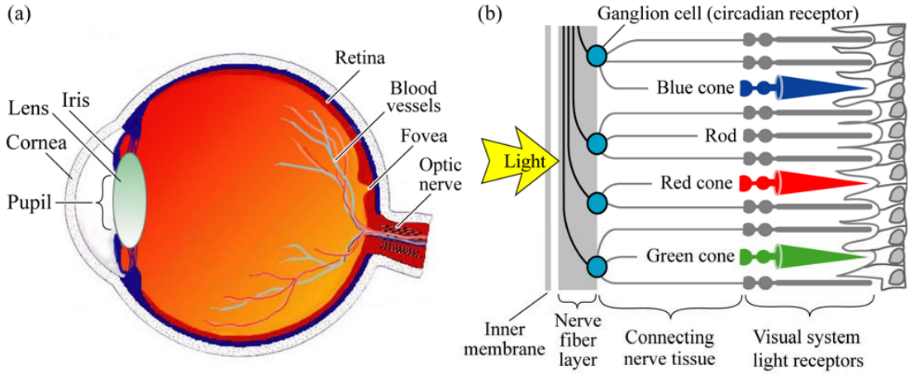


Figure 3.13: (a) Cross section through a human eye. (b) Schematic view of the retina and its photoreceptors (adapted from Encyclopedia Brittannica, 1994 edition)

The reference recipient of natural light is the human eye, which is illustrated in Figure 3.13. The inside of the eyeball is clad by the retina, which is the light-sensitive part of the eye. The illustration also shows the fovea, a cone-rich central region of the retina which affords the high acuteness of central vision. The schematic shows the cell structure of the retina including the light-sensitive rod cells and cone cells. Also shown are the ganglion cells and nerve fibers that transmit the visual information to the brain. Rod cells are more abundant and more light sensitive than cone cells. Rods are sensitive over the entire visible spectrum. There are three types of cone cells, namely cone cells sensitive in the red, green, and blue spectral range. The cone cells are therefore denoted as the red-sensitive, green-sensitive, and blue-sensitive cones, or simply as the red, green, and blue cones.

The eye operates at three different vision regimes, related to the type of receptors which are activated. Photopic vision relates to human vision at high ambient light levels (e.g. during daylight conditions) when vision is mediated by the cones. The photopic vision regime applies to luminance levels greater than 3 cd/m^2 . Scotopic vision relates to human vision at low ambient light levels (e.g. at night) when vision is mediated by rods. Rods have a much higher sensitivity than the cones. However, the sense of color is essentially lost in the scotopic vision regime. At low light levels such as in a moonless night, objects lose their colors and only appear to have different gray levels. The scotopic vision regime applies to luminance levels smaller 0.003 cd/m^2 . Mesopic vision relates to intermediary light levels between the photopic and scotopic vision regime.

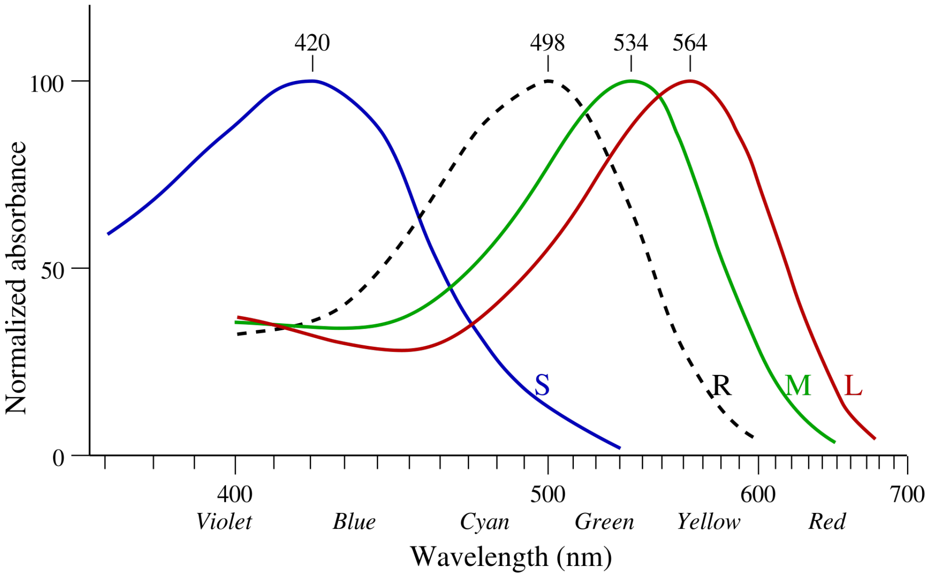


Figure 3.14: Normalized spectral sensivity of rod and cone cells. By Maxim Razin [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://commons.wikimedia.org/wiki/File:Cone-response.svg

The approximate spectral sensitivity functions of the rods and three types or cones are shown in Fig. 3.14. Night-time vision (scotopic vision) is weaker in the red spectral range and thus stronger in the blue spectral range as compared to photopic vision. As the scotopic vision is usually activated only in extremely dark environments, it has little chance to be used in plain daylight or even in twilight. Hanceforth, for the derivations needed to obtain the sensivity curve for lighting load forecast, only the photopic regime is considered.

The physical properties of electromagnetic radiation are characterized by radiometric units. Using radiometric units, light is characterized in terms of physical quantities: the number of photons, photon energy, and optical power. However, the radiometric units are irrelevant when it comes to light perception by a human being. For example, infrared radiation causes no luminous sensation in the eye. To characterize the light and color sensation by the human eye, different types of units are needed. These units are called photometric units. The luminous intensity, which is a photometric quantity, represents the light intensity of an optical source, as perceived by the human eye. The luminous intensity is measured in units of candela (cd), which is a base unit of the International System of Units. A monochromatic light source emitting an optical power of 1/683 watt at 555 nm into the solid angle of 1 steradian (sr) has a luminous intensity of 1 candela (cd). All other photometric units are shown in Table 3.6, which also compares then with the respective radiometric units.

Table 3.6: Photometric and corresponding radiometric units

Photometric unit	Unit	Radiometric unit	Dimension/Symbol
Luminous Flux	lm (Lumen)	Radiant Flux	Watt (W)
Luminous Intensity	$cd = lm/sr$ (Candela)	Radiant intensity	W/sr
Illuminance	$lux = lm/m^2$(Lux)	Irradiance	W/m^2
Luminance	cd/m^2	Radiance	$W/\left(sr \cdot m^2\right)$

The conversion between radiometric and photometric units is provided by the luminous efficiency function or eye sensitivity function, known as $V(\lambda)$. This function was first evaluated in 1924, giving rise to the CIE 1931 photometric standard. A modified $V(\lambda)$ function was introduced by [103] and this modified formulation is known as the CIE 1978 $V(\lambda)$. This function, which is largely employed in visual perception studies [107] and can be considered the most accurate description of the eye sensitivity in the photopic vision regime, is shown in Figure 3.15.

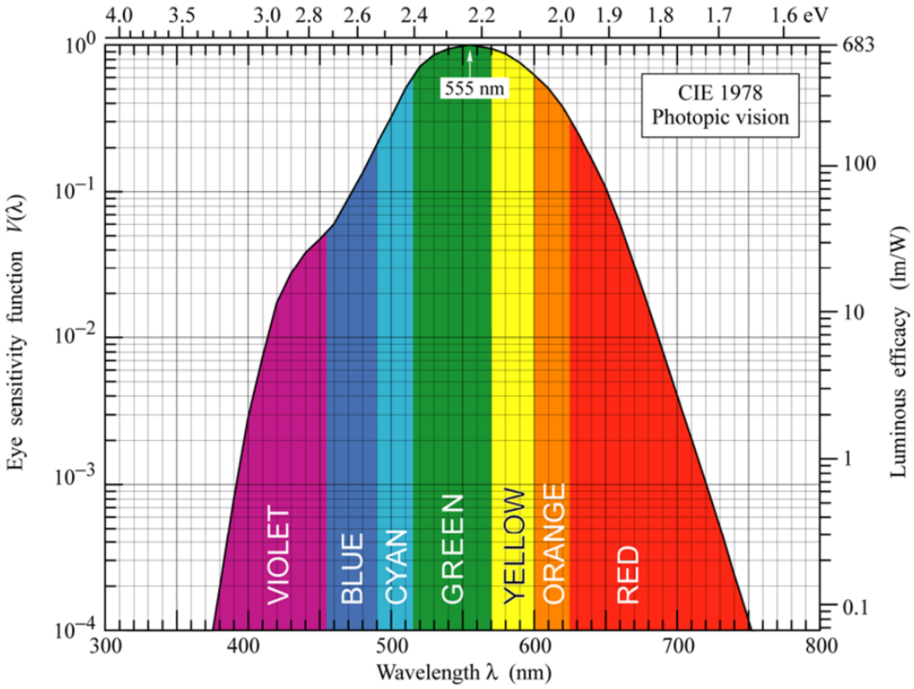


Figure 3.15: Eye sensivity function. The values of CIE 1978 $V(\lambda)$ are shown in the left-hand ordenate, while the correspondent luminous efficacy (conversion factor for Watts to lumens) are shown in the right-hand ordenate. Both are maximum in 555 nm wavelenght. By Jordanwesthoff (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://commons.wikimedia.org/wiki/File%3AHuman_photopic_response.jpg

In order to convert the irrandiance to illuminance, the following integration must be made:

$$\phi_d = 683 \int\limits_{\Lambda} V(\lambda) I_{C\lambda}(\lambda) d\lambda \tag{3.36}$$

where ϕ_d represents the Illuminance at direction d, Λ represents the interval of visible light wavelenghts and $I_{C\lambda}(\lambda)$ the irradiance at wavelenght λ as calculated by SPCTRL2 and SEDES2. Applying this expression to the North, South, East, West and Zenith directions gives the values of ϕ_N, ϕ_S, ϕ_E, ϕ_W and ϕ_z denoted in equation (3.33).

3.4.1.5 Solar irradiation

The Sun's irradiation has two principal effects over the electricity consumption. As well as being the main factor for artificial illumination, the global solar irradiance incident on buildings also increases cooling loads or decreases heating loads.

The problem formulation and modelling difficulties that arise when estimating the solar irradiation induced thermal load is similar to the ones faced when analysing natural illumination. To simulate the induced thermal load irradiated by the Sun over a single building, several factors should be taken into consideration, such as shape and orientation of roof and external walls, their thermal insulation and albedo (absortivity). However, again the goal of this estimation is to model the induced load over a very large number of buildings, not the exact model for a single structure.

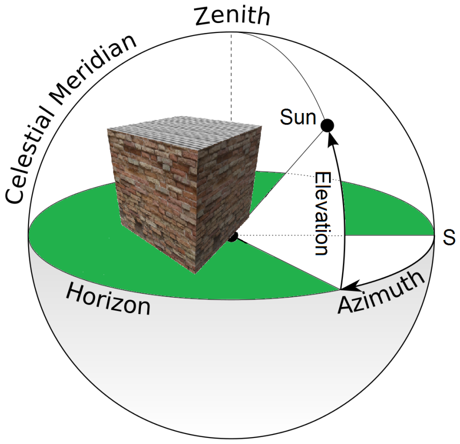


Figure 3.16: Natural solar irradiation. The roof and four wall surfaces are used to approximate the heat absorption by irradiation into buildings. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/commons/f/f7/Azimuth-Altitude_schematic.svg

Similar to the illumination case, only the effect of the sun direction must then be computed. This is accomplished by approximating the total incident irradiation by the sum of five components: one "roof" perpendicular to the surface, pointed to the zenith, and four "wall" paralel to the surface, each pointed to a compass point. The irradiation dependent component of the load y_I can then be approximated by the linear relationship shown in equation (3.37):

$$y_I = b_{I1}I_N + b_{I2}I_S + b_{I3}I_E + b_{I4}I_W + b_{I5}I_Z \tag{3.37}$$

$$y_I = \begin{bmatrix} b_{I1} & b_{I2} & b_{I3} & b_{I4} & b_{I5} \end{bmatrix} \begin{bmatrix} I_N & I_S & I_E & I_W & I_Z \end{bmatrix}^T \tag{3.38}$$

$$y_I = B_I U_I \tag{3.39}$$

where I_N, I_S, I_E, I_W and I_z denotes the irradiance component at North, South, East, West and Zenith (Up) directions, respectively. These are all calculated by means of the SPCTRL2 and SEDES2 models.

3.4.2 Socioeconomic variables

Other important aspects of electricity demand are associated with social and economic facts. Economic growth, industrial production and running stock of electric appliances. The causal relationship between economic growth, characterized in diverse indicators, and the electricity consumption is investigated in numerous of papers. In [11], Granger tests indicate short-run causality from energy consumption to income for India and Indonesia, while the test points to bidiretional relationship for Thailand and the Philippines. The study presented in [22] conclude that causality is stronger in developed OECD countries than in developing countries. Several variables are used to assert the dependencies between energy consumption and economic activities: Gross Domestic Product (GDP), population and price indexes [7].

The use of socioeconomic variables to load forecast is not novel. In reference [36], aggregate energy supply, macroeconomic data such as gross domestic product (GDP), population growth, buildings construction and demolition rate are employed to forecast hourly demand in the city of Abu Dhabi. In developing countries whose economies are continuously growing, the trend in GDP is highly correlated to long term trends in electric demand [89]. According to reference [46], residential load density is dependent on socioeconomic factors such as population, income level, living space per person, household appliances capacity. The commercial load density is influenced by the urban GDP, international economic situation, speed of urban economic development, commercial cyclical fluctuation. The industrial load density is influenced by technical progress, energy saving policy and production output. Some studies also link industrial consumption to GDP [15]

For the population input, time series can usually be obtained in the regional or national statistical database. Since these time series usually have monthly or annual values, daily values must be obtained through curve fitting or forecast. Generally, daily values can be determined by means of exponential smoothing, Grey models or cubic splines. The agricultural load density is influenced by farmland area, technical progress.

A similar approach is executed for the GDP input. Dividing GDP by the population, one can obtain the per capita GDP. A modified rolling grey algorithm as described in [98] is applied to simulate forecasts of these candidate input variables, as most of these indicators cannot be collected in real time. Additional variables are taken into account, such as inflation indicators, dollar exchange rate, the fraction of low income households, the relative sales volume index and the energy intensity indicator for industries with low, medium and high specific energy consumption, relative imports and export indexes, life expectancy at birth, basic sanitation at residences and birth rate.

Table 3.7: List of Socioeconomic variables obtained per Case study

Symbol	Variable	Brasília	Leipzig
SE_1	Population	Yes	Yes
SE_2	GDP	Yes	Yes
SE_3	GDP per capita	Yes	Yes
SE_4	Price Index	Yes	No
SE_5	Currency Exchange rate	Yes	Yes
SE_6	Light Industry Energy Intensity	Yes	No
SE_7	Medium Industry Energy Intensity	Yes	No
SE_8	Heavy Industry Energy Intensity	Yes	No
SE_9	Sales volume index	Yes	No
SE_{10}	Low income households	Yes	No
SE_{11}	Relative import index	Yes	No
SE_{12}	Relative export index	Yes	No
SE_{13}	Life Expectancy	Yes	No
SE_{14}	Basic Sanitation	Yes	No
SE_{15}	Birth rate	Yes	No

A regression similar to the models presented in [76, 14] is proposed. The socioeconomic dependent load y_{SE} is then estimated by the linear combination of the variables listed in Table 3.7:

$$y_{SE} = b_{SE1}SE_1 + b_{SE2}SE_2 + ... + b_{SE15}SE_{15} \tag{3.40}$$

$$y_{SE} = \begin{bmatrix} b_{SE1} & b_{SE2} & \cdots & b_{SE15} \end{bmatrix} \begin{bmatrix} SE_1 & SE_2 & \cdots & SE_{15} \end{bmatrix}^T \tag{3.41}$$

$$y_{SE} = B_{SE}U_{SE} \tag{3.42}$$

3.4.3 Electricity tariffs

Increasing prices of goods and services often leads to decreased demand. However, as electricity is an essential service, its price elasticity may be very low. This translates in reduced price sensivity on the consumer part, specially from residential customers. On the other hand, industries and large consumers tend to be attentive and prudent with their energy consumption, as increased costs could shrink profit margins.

This capacity on part of the consumer to curtail part of its consumption due to higher prices is a object of study, as this fact can be a new tool to keep generation and load in balance. Known as Demand Response (DR), this practice has been suggested as a potentially valuable resource in future electricity systems, as it could constitute an alternative to potentially more costly means of system operation, such as backup generation, network expansion and physical electricity storage [41].

The benefits of DR programs are market-wide. An overall electricity price reduction is expected because of a more efficient utilization of the available infrastructure, reducing demand from expensive electricity generating units and avoiding losses on busy distribution feeders during peak load. Moreover, DR programs can increase short-term capacity using market-based programs, which in turn, results in an avoided or deferred capacity costs [6]. However, all demand response programs require hourly meter readings, which in practice necessitate automatic meter reading systems and a deregulated spot market for energy, at last for part of the consumer classes. Not every country has both technical and regulatory prerequisites to operate DR programs. Without a real time, Smart Grid like environment, it is not known how quickly does the consumer react to changes in electricity prices.

In Brazil, currently there is no hourly metering for low voltage consumers, neither a real time energy market. In 2015 the Brazilian Electricity Regulatory Agency (ANEEL) approved a monthly variable economic signal in the energy price called Tariff Flags. The Flags are Green, Yellow and Red (two levels), in analogy with traffic lights. They represent tariff differences to the small consumer and impart extra fees to the energy price, giving the consumer an economic indicator to conserve energy when the generation is costlier. It aims to minimize eventual differences between costs and revenues of the utilities and contribute to the optimization of the system's electricity and energy resources.

In a analysis of the Tariff Flag program, it was noticed that the intensity of the demand response of each economic sector depends not only on price elasticity, but also on the energy tariff that is applied. The industrial sector is expected to be the most affected, with reductions in the order of 3.5% and 7.0% (including fees) according to the Yellow and Red flags, respectively [69]. Thus, a monthly change in tariff can cause visible load reductions in a few months, a relatively short term. This gives motivation to investigate the impact of regular changes in the electricity tariff as a variable for short-term load forecast.

Brazilian electricity tariff system is complex. Low voltage customers only have access to the conventional monomial tariff, in which there is a fixed tariff for energy ($/kWh). High voltage clients must adhere to a binomial tariff contract, in which there two rates: one for energy ($/kWh) and another for demand ($/kW). There is a surtax if the demanded power is higher than the contract limit. The tariff type can be conventional, hourly seasonal type green or hourly seasonal type blue, moving from fixed rates for demand, energy and surtaxes to different rates due to seasons and peak hours. As there is a 60 day delay between measurements and the payment of energy bill, 60 day moving averages are employed to smooth the transitions. In Fig. 3.17 a moving average representation of tariffs (conventional type) by consumer classes is presented.

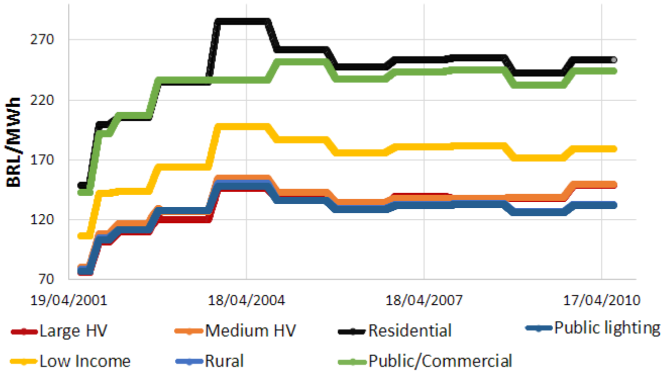


Figure 3.17: 60 day moving average of electricity tariffs in Brasília, in Brazilian Reais (BRL) per MWh, by consumer class

In the period from 2001 to 2010, due to the multitude of classes, types and seasonal periods, the historic tariffs time series is composed of 75 candidate variables, being 11 low voltage conventional, 10 high voltage conventional, 18 hourly seasonal type green and 36 hourly seasonal type blue. Variables represent the full set of unitary cost of energy, demand and overdemand fee, at each tariff type and voltage level. Attributing each variable a coefficient, the tariff dependent load component y_τ is given by (3.43):

$$y_\tau = b_{\tau 1}\tau_1 + b_{\tau 2}\tau_2 + ... + b_{\tau 75}\tau_{75} \tag{3.43}$$

$$y_\tau = \begin{bmatrix} b_{\tau 1} & b_{\tau 2} & \cdots & b_{\tau 75} \end{bmatrix} \begin{bmatrix} \tau_1 & \tau_2 & \cdots & \tau_{75} \end{bmatrix}^T \tag{3.44}$$

$$y_\tau = B_\tau U_\tau \tag{3.45}$$

In the german forecasting scenario, the tariff history was not found in an online database. Due to lack of data, the tariff variables are only are used in the Brazilian forecasting scenario.

3.4.4 Calendar and Weather events

The load profiles have markedly distinct behavior in working days, holidays and weekends. There are also atypical days [24] with different load curves, such regular day preceding or following a holiday. Large media and sports events can also lead to uncommon behavior in the electricity demand.

While introducing additional variability to the forecasting problem, calendar events have the advantage of being previsible, which can be represented as a binary variable. These can be described as boolean time series that have a true value when the event is expected, being it false otherwise. The use of this type of input in load forecasting methodologies is not new [82, 90, 81]. Some scholars advocate that the workday-holiday parametrization can be improved if all seven days of the week are separately described in the variables[75].

Table 3.8: List of event variables

Symbol	Variable
e_1	Monday
e_2	Tuesday
e_3	Wednesday
e_4	Thursday
e_5	Friday
e_6	Saturday
e_7	Sunday
e_8	Holiday
e_9	Attypical day
e_{10}	Summer Saving Time
e_{11}	Fog
e_{12}	Rain
e_{13}	Thunderstorm
e_{14}	Snow

Expanding on other papers characterizations, for this load forecasting system there are binary variables for each day of the week, for summer saving time, for holidays and attypical days. The latter are classified as such due to proximity to other holidays or the occurrence of major media or sports events, such as soccer matches. Also present are boolean variables for the four weather events informed by the METAR: Fog, Rain, Snow and Thunderstorm. These weather phenomenons do alter energy consumption due to outages, increased heating and lighting demand. The event dependent load component y_e is then estimated by:

$$y_e = b_{e1}e_1 + b_{e2}e_2 + ... + b_{e14}e_{14} \tag{3.46}$$

$$y_e = \left[\begin{array}{cccc} b_{e1} & b_{e2} & \cdots & b_{e14} \end{array} \right] \left[\begin{array}{cccc} e_1 & e_2 & \cdots & e_{14} \end{array} \right]^T \tag{3.47}$$

$$y_e = B_e U_e \tag{3.48}$$

3.4.5 The Load Model

Time-series approaches are among the oldest methods applied in load forecasting [47]. These were developed in order to directly incorporate a specific time-dependent structure in the analysed data, i.e. the dependence of a variable on its previous values [18]. This approach presents advantages, as it can model trends and periodical variations in a time series without requiring detailed knowledge about the inner dynamics of the system. A very simple class are the so-called autoregressive moving average or ARMA models, depicted in expression (3.49):

$$y[k+1] = \sum_{i=1}^{n} a_k y[k-i+1] + z[k] + \sum_{i=1}^{q} c_k z[k-i+1] \tag{3.49}$$

where $Y[k]$ is the time series to be modeled, $Y[k-i]$ are its previous values and $z[k]$ is a white noise component with zero mean and σ^2 variance. The model order n (autoregressive part) and q (Moving Average) must be determined, and then the coefficients a_i and c_i are calculated by Maximum Likelihood or Least Squares variants. This model can be expanded to include m exogenous variables, giving the autoregressive moving average exogenous (ARMAX) model:

$$y[k+1] = \sum_{i=1}^{n} a_i y[k-i+1] + \sum_{j=1}^{m} \sum_{i=1}^{p} b_{ij} u_j[k-i+1] + z[k+1] + \sum_{i=1}^{q} c_i z[k-i+1] \tag{3.50}$$

Some considerations about the nature of load measurements can be made to simplify this expression. Supposing that $z[k]$ is a measurement error and that its previous values do not affect the current measurement $Y[k]$ gives:

$$y[k+1] = \sum_{i=1}^{n} a_i y[k-i+1] + \sum_{j=1}^{m}\sum_{i=1}^{p} b_{ij} u_j[k-i+1] + z[k+1] \tag{3.51}$$

In terms of electric load modeling, equation (3.51) can be understood as a linear dependence of the future output $y[k+1]$ to an autoregressive term $y_{AR}[k]$, to an exogenous input term $y_U[k]$ with p delays and the noise $z[k+1]$:

$$y[k+1] = y_{AR}[k] + \sum_{i=1}^{p} \left(y_U[k-i+1]\right) + z[k+1] \tag{3.52}$$

Rewriting (3.51) in vectorial form yields:

$$y[k+1] = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T \begin{bmatrix} y[k] \\ y[k-1] \\ \vdots \\ y[k-n+1] \end{bmatrix} + \sum_{i=1}^{p} \left(\begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{im} \end{bmatrix}^T \begin{bmatrix} u_1[k-i+1] \\ u_2[k-i+1] \\ \vdots \\ u_m[k-i+1] \end{bmatrix} \right) + z[k+1] \tag{3.53}$$

The summation term can be equivalently rewritten as a product of two vectors p times stacked:

$$y[k] = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T \begin{bmatrix} y[k] \\ y[k] \\ \vdots \\ y[k-n+1] \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{12} \\ \vdots \\ b_{1m} \\ b_{21} \\ b_{22} \\ \vdots \\ b_{2m} \\ \vdots \\ b_{p1} \\ b_{p2} \\ \vdots \\ b_{pm} \end{bmatrix}^T \begin{bmatrix} u_1[k] \\ u_2[k] \\ \vdots \\ u_m[k] \\ u_1[k-1] \\ u_2[k-1] \\ \vdots \\ u_m[k-1] \\ \vdots \\ u_1[k-p+1] \\ u_2[k-p+1] \\ \vdots \\ u_m[k-p+1] \end{bmatrix} + z[k+1] \quad (3.54)$$

In the previous subsections, the main exogenous variables affecting electric load have been defined. The input dependent load $y_U[k]$ is obtained by the following expressions:

$$y_U[k] = y_T[k] + y_H[k] + y_v[k] + y_\phi[k] + y_I[k] + y_{SE}[k] + y_\tau[k] + y_e[k] \quad (3.55)$$

where y_T is obtained in equation (3.15), y_H in (3.27), y_v in (3.32), y_ϕ in (3.35), y_I in (3.39), y_{SE} in (3.42), y_τ in (3.45) and y_e in (3.48). Substituting yields:

$$\sum_{i=1}^{p} (y_U[k-i+1]) = \sum_{i=1}^{p} \left(\begin{bmatrix} B_{iT} \\ B_{iH} \\ B_{iv} \\ B_{i\phi} \\ B_{iI} \\ B_{iSE} \\ B_{i\tau} \\ B_{ie} \end{bmatrix}^T \begin{bmatrix} U_T[k-i+1] \\ U_H[k-i+1] \\ U_v[k-i+1] \\ U_\phi[k-i+1] \\ U_I[k-i+1] \\ U_{SE}[k-i+1] \\ U_\tau[k-i+1] \\ U_e[k-i+1] \end{bmatrix} \right) \quad (3.56)$$

$$\sum_{i=1}^{p} (y_U[k-i+1]) = \sum_{i=1}^{p} \left(\widetilde{B}_i \widetilde{U}[k-i+1]\right) \tag{3.57}$$

$$\sum_{i=1}^{p} (y_U[k-i+1]) = \left[\begin{array}{cccc} \widetilde{B}_1 & \widetilde{B}_2 & \cdots & \widetilde{B}_p \end{array}\right] \left[\begin{array}{c} \widetilde{U}[k] \\ \widetilde{U}[k-1] \\ \vdots \\ \widetilde{U}[k-p+1] \end{array}\right] \tag{3.58}$$

By inspection of (3.54) and (3.58), it is possible to define the coupling vector $\widetilde{B}$ and input vector $U[k]$ and input coupling matrix B:

$$\widetilde{B}U[k] = \left[\begin{array}{c} b_{11} \\ b_{12} \\ \vdots \\ b_{1m} \\ b_{21} \\ b_{22} \\ \vdots \\ b_{2m} \\ \vdots \\ b_{p1} \\ b_{p2} \\ \vdots \\ b_{pm} \end{array}\right]^T \left[\begin{array}{c} u_1[k] \\ u_2[k] \\ \vdots \\ u_m[k] \\ u_1[k-1] \\ u_2[k-1] \\ \vdots \\ u_m[k-1] \\ \vdots \\ u_1[k-p+1] \\ u_2[k-p+1] \\ \vdots \\ u_m[k-p+1] \end{array}\right] = \left[\begin{array}{cccc} \widetilde{B}_1 & \widetilde{B}_2 & \cdots & \widetilde{B}_p \end{array}\right] \left[\begin{array}{c} \widetilde{U}[k] \\ \widetilde{U}[k-1] \\ \vdots \\ \widetilde{U}[k-p+1] \end{array}\right] \tag{3.59}$$

Defining the state vector $X[k]$, and decomposing the noise term $z[k]$ into a measurement component $v[k]$ and a process noise $W[k]$, both gaussian i.i.d.:

$$X[k+1] = \begin{bmatrix} y[k+1] \\ y[k] \\ \vdots \\ y[k-n+1] \end{bmatrix} ;\ X[k] = \begin{bmatrix} y[k] \\ y[k-1] \\ \vdots \\ y[k-n] \end{bmatrix} ;$$

$$z[k+1] = v[k+1] + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T W[k] \tag{3.60}$$

It is now possible to transform the difference equation (3.54) in a State Space model in the companion form, expressed in matricial form as:

$$X[k+1] = \begin{bmatrix} a_1 & a_2 & \cdots & a_{n+1} & a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} X[k] + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{pm} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} U[k] + W[k] \tag{3.61}$$

The output equation is given by expression (3.62):

$$y[k+1] = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} X[k+1] + v[k+1] \tag{3.62}$$

Rewriting the equations in matricial form gives the final load State space model:

$$X[k+1] = \mathbf{A}X[k] + \mathbf{B}U[k] + W[k] \tag{3.63}$$

$$Y[k+1] = \mathbf{C}X[k+1] + v[k+1] \tag{3.64}$$

The State space formulation is preferred because it allows a more concise presentation and manipulation. Unlike Box-Jenkins models, there is no need for stationarity for the load and inputs time series, permiting he use of a large set of exogenous variables of different types, such as boolean, integer and real valued. State space is also the base of online and adaptive Kalman filter predictors, which adds robustness through the application of an double state/measurement noise model. There are, however, additional difficulties to estimate two noise covariance matrices [74].

3.5 Photovoltaic Generation Model

Photovoltaic (PV) energy is one of the most promising renewable generation technologies. The cost of PV modules decreased fivefold between 2008 and 2013, while the cost of full PV systems decreased by 66% in the same period. The levelised cost of electricity of decentralised solar PV systems is approaching or falling below the utilities energy tariffs in some markets, across residential and commercial segments. As such, cumulative PV capacity grew at a rate of 49% per year from 2003 to 2013 [51], as illustrated in figure 3.18.

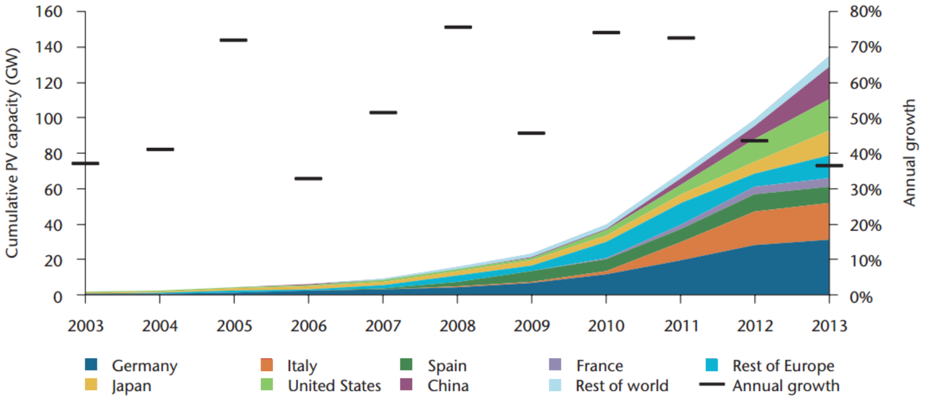


Figure 3.18: Global cumulative growth of PV capacity. Source: reproduced from IEA Solar photovoltaic roadmap 2014 [51]

PV cells are the most basic unit in a photovoltaic power producing device, typically available in 12,5 cm and 15 cm square sizes. In general, these can be classified as either silicon-based crystalline (monocrystaline and polycrystalline silicon), Thin-film, or organic. Currently, crystalline silicon technologies account for more than 94% of the overall cell production in the IEA countries participating into the Photovoltaic Power System Programme [86].

Monocrystaline silicon (mono-Si) cells are produced from a single crystal growth method, having commercial efficiencies between 16% and 25%. Polycrystalline silicon (poly-Si) cells are usually manufactured from a cheaper cast solidification process, produced by cooling and solidifying molten silicon, then cutting it into thin plates. The solidification of the material results into cells that contain many crystals, making the surface of the poly-Si cell less perfect than a mono-Si device. Due to these defects, polycrystalline are also less efficient than mono-Si. However, they have remained popular because they are less expensive but cost-effective, with average conversion efficiency around 14-18%.

Thinfilm cells are formed by the deposition of extremely thin layers of photovoltaic semiconductor materials onto a inert substrate material such as glass, stainless steel or plastic. They are potentially less expensive to manufacture than crystalline cells, but have conversion efficiencies slightly below poly-Si, in average. Some expensive high end Thin-film have efficiencies comparable to mono-Si. Thin-film semiconductor materials commercially used are cadmium telluride (CdTe), and Copper-Indium-(Gallium)-Selenide (CIGS and CIS). In the past, amorphous silicon (a-Si) had a significant market share but lately failed behind in both cost reductions and efficiency gains. In terms of efficiencies, in 2016, CdTe cells reached 22% in labs.

Organic thin-film PV cells, using dye or organic semiconductors, have created interest and research, development and demonstration activities are underway. In recent years, perovskites solar cells have reached efficiencies higher than 20% in labs but have not yet resulted in stable market products.

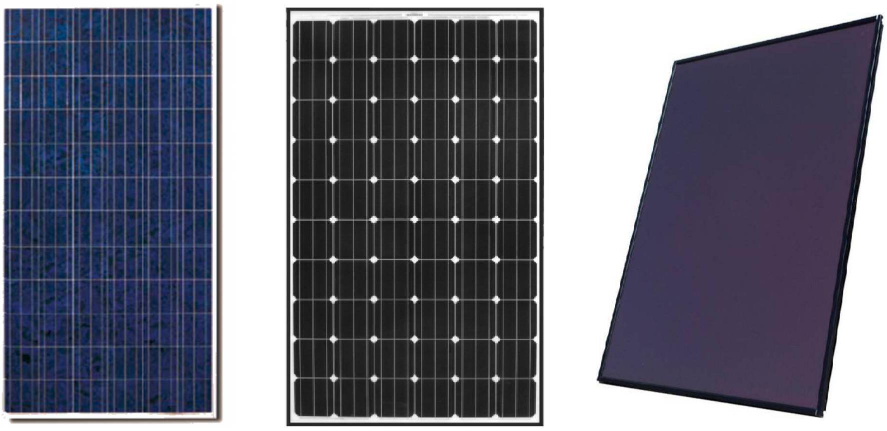


Figure 3.19: Apparent difference between module types. From left to right. Polycrystaline Silicon module, Monocrystaline Silicon and Thin Film module.

Excepting extreme latitudes and locations under prolonged shadow due to geographical features, solar power is available worldwide and, unlike thermal power sources, photovoltaic systems' efficiencies do not substantially decrease from a utility sized plant to a rooftop residential system. Due to these characteristics favoring decentralized deployment, PV power is the main power source for distributed generation directly connected to the distribution network. In grid-connected PV systems, an inverter is used to convert electricity from direct current (DC) to the alternating current (AC) supplied to the electricity network. Conversion efficiency is in the range of 95% to 99%, varying with inverter size and temperature. Most inverters incorporate a Maximum Power Point Tracker (MPPT), which continuously adjusts the load impedance to provide the maximum power from the PV array. At the end of 2015, 227 GWp of photovoltaic panels have been installed worldwide. Germany, Greece and Italy had more of 7% of their electricity demand supplied by photovoltaic arrays [85]. Variability of solar resource poses difficulties in grid management as solar penetration rates rise continuously. This level of PV penetration can substantially alter the electric load behavior, adding a new variable to the load forecasts conducted by the power system operators to ensure stability and economical dispatch.

In order to forecast solar photovoltaic power, a realistic yet concise model of this electricity source is required. Forecasting methodologies can be largely characterized as physical or statistical. The physical approach combines solar irradiation and PV system models to predict generation, whereas the statistical approach primarily confides on past data to generate forecast, with little or no reliance on irradiance and PV models. Predictions could be for short-term, done up to one week ahead, medium term covering one week to one month, and long-term for forecasting months or years ahead. Literature favors statistical approaches for short-term forecast, while physical methods are well suited to perform long-term prediction [108]. There are also hybrid-physical methods that combine statistical with a simplified physical input set to improve performance.

Very similar to the tools employed for load forecasting, examples of statistical approaches for predicting PV generation include the persistence (naive) model, linear regressions, ARMA time series, exponential smoothing, Artificial Neural Networks, support vector approaches and fuzzy inference set. Artificial Neural Networks in several variants is the model of choice for almost 25% of the recent papers, while regression and ARMA derived approaches amount to 18%. Hybrid-physical modeling has comparatively few publications, comprising 6% of the studies reviewed in [10].

A hybrid-physical adaptive PV generation modeling framework consisting of an State space linear representation feeded with weather variables and simulated physical inputs is proposed in this research. The first step simulates the solar radiation in clear sky conditions, the second step simulates the radiation in cloudy conditions, the third calculates the PV panel generation per area accordingly to a defined model, and the fourth and final step simulates the growth rate of the installed area in order to simulate the total mean and maximum PV generation. The framework box diagram is shown in Fig. 3.20.

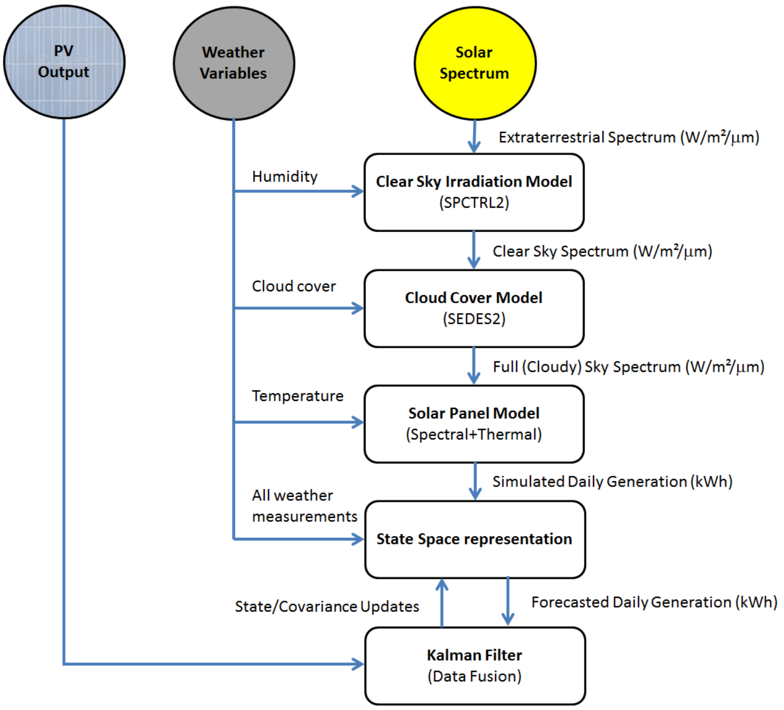

PV
Output
Weather
Variables
Solar
Spectrum
Extraterrestrial Spectrum (W/m²/μm)
Humidity
Clear Sky Irradiation Model
(SPCTRL2)
Clear Sky Spectrum (W/m²/μm)
Cloud cover
Cloud Cover Model
(SEDES2)
Full (Cloudy) Sky Spectrum (W/m²/μm)
Temperature
Solar Panel Model
(Spectral+Thermal)
Simulated Daily Generation (kWh)
All weather
measurements
State Space representation
State/Covariance Updates
Forecasted Daily Generation (kWh)
Kalman Filter
(Data Fusion)

Figure 3.20: Box diagram of the solar photovoltaic simulational framework

Each step generates its outputs using its specific inputs, which in turn are either static parameters related to PV panel and growth model, daily weather measurements or outputs generated by the previous step. This simulational framework determines the mean and maximum PV generation for a given day, so it must be executed once for every time step of the predicting algorithms. The relationship between these inputs, outputs and simulation steps are shown in Table 3.9, as well as the data flow.

Table 3.9: Inputs and outputs per PV forecasting step

Step	Inputs	Outputs
Clear Sky	Geometry/Site parameters Humidity Parameters	Diffuse Irradiation (Clear) Direct Normal Irradiation (Clear)
Cloudy Sky	Cloud Cover Parameter Diffuse Irradiation (Clear) Direct Normal Irradiation (Clear)	Diffuse Irradiation (Cloudy) Direct Normal Irradiation (Cloudy)
Solar Panel	Diffuse Irradiation (Cloudy) Direct Normal Irradiation (Cloudy) Cell and Surface material Temperature Cell temperature coefficient Panel effective area Fill factor	Simulated PV generation Conversion efficiency
State Space Model	Simulated PV generation Weather variables State/Covariance Updates	Forecasted PV generation

3.5.1 The Solar irradiation model

A slightly modified Bird Simple Spectral Model (SPCTRL2) is applied in this forecasting/simulation algorithm to compute the solar irradiation source [16]. Its results are then modified by the SEDES2 Cloud Cover model, and the resulting spectra is then used as input to the Photovoltaic Panel model, in order to compute the electricity generation. These models are chosen because of its public nature, relative simplicity and open license [1, 77]. Mostly, they do not require access to detailed or special meteorological data and its results offer acceptable agreement with more strict/detailed models and field measurements, after some calibration in its input parameters.

SPCTRL2, in its original implementation, it computes clear sky spectral direct beam, hemispherical diffuse, and hemispherical total irradiances on a tilted or horizontal receiver plane at a single point in time. For tilted planes, the user specifies the tilt and azimuth of the plane, geometric properties displayed in figure 3.21.

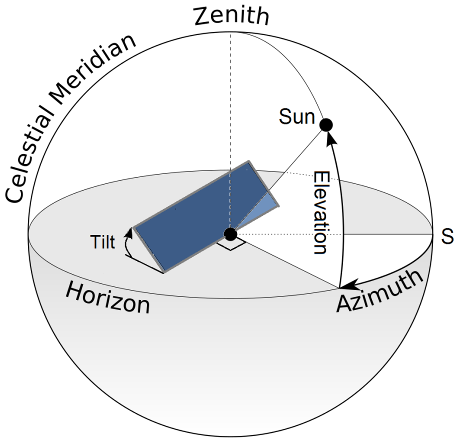


Figure 3.21: Sun azimuth angle and elevation angle. Azimuth reference is the geographical north pole. Modified from original provided by By TWCarlson [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons. Accessible at https://upload.wikimedia.org/wikipedia/commons/f/f7/Azimuth-Altitude_schematic.svg

The wavelength spacing is irregular, covering 122 wavelengths from 305 nm to 4000 nm. Aerosol optical depth, total precipitable water vapor (cm), and equivalent ozone depth (cm) must be specified by the user. The model does not take into account variations in atmospheric structure or constituents, and it also lacks a separate computation of circumsolar radiation, as the direct irradiance is assumed to contain this radiation component within a 5 degree solid angle.

The Clear Sky model was implemented in MATLAB environment and made to recursively calculate the irradiances over several points in time, in order to be able to compute daily, weekly or yearly irradiation spectra. The model inputs are: Latitude, Longitude, Panel tilt and Azimuth angles, Atmospheric Pressure, time step, Precipitable Water Vapor, Albedo, Ozone Column thickness and Aerosol Optical Depth. Outputs are Direct Normal (shown in figure 3.22), Diffuse and Global solar irradiation with clear skies.

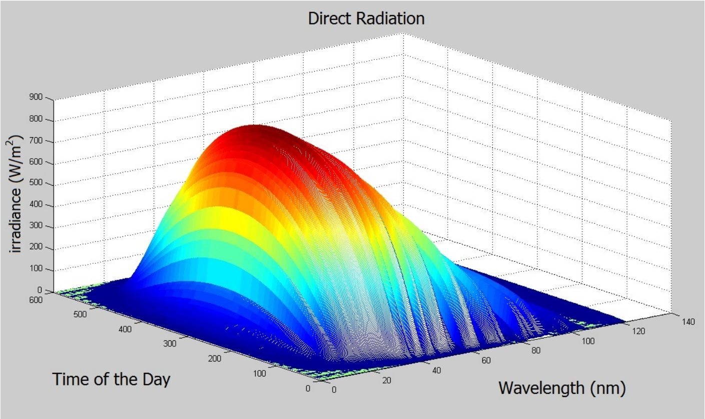


Figure 3.22: Direct Normal Irradiation as a function of time and wavelength

Special care was taken with the Albedo, Precipitable Water Vapor Column, Ozone Column Thickness and Aerosol Optical Depth, as those are not readily available neither straightforwardly measured in weather stations.

The Albedo input is related to the diffuse reflectivity of the nearby environment. It typically varies from 5% from asphalt pavement up to 55% from fresh concrete [34]. Typical albedo values for western cities range from 10 to 20% [97]. Proximity to deserts could increase albedo significantly, on the other hand proximity to deep water bodies tends to decrease it. Given complete weather station data, occurrence of snow could be accounted in the algorithm by increasing the base albedo value for a few percent, accordingly to [19] research on satellite imaging over Hartford, USA.

As shown in [34], the seasonal albedo variation in a large city without snow events is relatively small, although diurnal variation can be as high as 50%. However, there is little need to correct base Albedo values in cities without snow as the variation mostly occurs near sunrise and sunset, and there is a approximate linear slope around 12PM. As there is an approximate symmetry in the daily irradiation pattern between late morning and early afternoon, this kind of albedo variation cancels itself out for PV generation purposes.

The Ozone Column was estimated through Latitude and Longitude inputs by means of the Heuklon empirical model [48]. The Heuklon model parameters, however, were updated accordingly with data shown in [57].

As the original model reaches numerical singularities when computing sunrise and sunset irradiation, a logical switch was implemented in order to make a linear interpolation when those singularities are calculated at high zenith angles (Sun near or below horizon). This procedure introduces computation errors, but those are negligible when considering the much lower magnitude of irradiation at those extreme angles of incidence and a sufficiently small type step, typically smaller than 15 minutes.

A direct and diffuse solar spectrum is achievable in function of the location, date and hour of the day [1]. Climatic relevant information for PV power generation are obtained [2] from METARs, as well as historic of events related to raining, snow, haze, mean visibility, air pressure and relative air humidity.

Applying the SPCTRL2 model as described in Appendix B, daily solar incidence is simulated, obtaining the values of direct, diffuse and global solar radiation spectrum. This irrandiance represents the solar power arriving at a tilted plane in a cloudless sky. The spectral global irradiance on an tilted surface is represented by the expression shown in (B.34):

$$I_\lambda = I_{d\lambda}\cos(\theta) + I_{s\lambda}\left\{\frac{I_{d\lambda}\cos(\theta)}{H_{0\lambda}D\cos(Z)} + \left[\left(\frac{1+\cos(T)}{2}\right)\left(1-\frac{I_{d\lambda}}{H_{0\lambda}D}\right)\right]\right\} + \\ + \frac{(I_{d\lambda}+I_{s\lambda})\,r_{g\lambda}\,(1-\cos(T))}{2} \tag{3.65}$$

The angle of incidence θdepends on the solar zenith angle Z, tilt angle T, Sun azimuth A and surface azimuth A_φ, as shown in equation (B.35):

$$\theta = \cos^{-1}\left(\cos(Z)\cos(T) + \sin(Z)\cos(A-A_\varphi)\sin(T)\right) \tag{3.66}$$

The SEDES2 model, also described in Appendix B, modifies this irrandiance spectrum according to the reported cloud cover index in order to account for the additional scattering and reflections. These modifiers use a quadratic equation with the clearness index K_t and six empirically derived constants, as shown in (B.36):

[1] http://www.nrel.gov/solar radiation/data.html

[2] http://www.wunderground.com/

$$I_{C\lambda} = \left[A1_\lambda + \frac{A2_\lambda}{\cos\left(Z\right)} + \left(B1_\lambda + \frac{B2_\lambda}{\cos\left(Z\right)}\right) K_t + \left(C1_\lambda + \frac{C2_\lambda}{\cos\left(Z\right)}\right) K_t^2\right] I_\lambda \tag{3.67}$$

where the clearness index is defined as the ratio between the Global Horizontal Irradiance (H) and the extraterrestrial irrandiance H_0 projected over the surface area:

$$K_t = \frac{H}{H_0 \cos\left(Z\right)} \tag{3.68}$$

In practice, the clearness index K_t is approximated from the cloud cover C_w by means of the approximated model studied in [95]:

$$H = H_0 \left[z_1 \left(\sqrt{T_{Max} - T_{Min}}\right) + z_2 \left(\sqrt{1 - \frac{C_w}{8}}\right)\right] \tag{3.69}$$

Equation (3.69) was developed as a method to provide estimates of daily global radiation as input for the Crop Growth Monitoring System of the European Union. T_{max} and T_{Min} denote the maximum and minimum daily temperature as informed in the METAR. The cloud cover index C_w in this formulation must be given in octas, a measurement of how many eighths of the sky are obscured by clouds. The coefficients z_1 and z_2 have to be fitted to the observations, as they are location and season dependent. Denoting the temperature variation as ΔT and the complement of C_w as $\overline{C_w}$, substituting (3.69) in (B.36) yields:

$$\begin{split} I_{C\lambda} = {} & \left[A1_\lambda + \frac{A2_\lambda}{\cos\left(Z\right)} + \left(B1_\lambda + \frac{B2_\lambda}{\cos\left(Z\right)}\right)\left(\left[z_1\left(\sqrt{\Delta T}\right) + z_2\left(\sqrt{\overline{C_w}}\right)\right]\right) \sec\left(Z\right)\right] I_\lambda + \\ & + \left[\left(C1_\lambda + \frac{C2_\lambda}{\cos\left(Z\right)}\right)\left(\left[z_1\left(\sqrt{\Delta T}\right) + z_2\left(\sqrt{\overline{C_w}}\right)\right]\right)^2 \sec^2\left(Z\right)\right] I_\lambda \end{split} \tag{3.70}$$

$$I_{C\lambda} = I_\lambda \begin{bmatrix} A1_\lambda + A2_\lambda \sec(Z) \\ (B1_\lambda + B2_\lambda \sec(Z))\sqrt{\Delta T} \\ (B1_\lambda + B2_\lambda \sec(Z))\sqrt{\overline{C_w}} \\ 2\,(C1_\lambda + C2_\lambda \sec(Z))\sqrt{\Delta T\overline{C_w}} \\ (C1_\lambda + C2_\lambda \sec(Z))\,\Delta T \\ (C1_\lambda + C2_\lambda \sec(Z))\,\overline{C_w} \end{bmatrix}^T \begin{bmatrix} 1 \\ z_1 \\ z_2 \\ z_1 z_2 \\ (z_1)^2 \\ (z_2)^2 \end{bmatrix} \tag{3.71}$$

where the coefficients z_1 and z_2 are seasonally updated by means of the model optimization routine. In practice, the model optimizes the six irradiation components denoted in (3.71) according to the recent weather and generation history. The spectral profile determined by the wavelenght dependent coefficients $A1_\lambda$, $A2_\lambda$, $B1_\lambda$, $B2_\lambda$, $C1_\lambda$ and $C2_\lambda$ is shown in the Appendix B. If there uncertainties about the solar panels' azimuth and tilt angle, multiple cloudy skies spectral irradiation profiles $I_{C\lambda}$can be used as inputs, calculated with slightly different geometric parameters.

3.5.2 PV panel model

In order to estimate the daily power density (W/m^2), the cloudy sky irradiance $I_{C\lambda}$ calculated by SPCTRL2 and SEDES2 must be further modified by solar panels' spectral response, electrical and thermal parameters. The solar panel is an array of solar cells electrically interconnected. The cells are protected from weather and intemperism by an inert encapsulment material, while a protective film coating is applied on the backside to provide chemical stability. The glass coating in the Sun facing side provides chemical protection and additional mechanical support. There is a hard frame that provides structural integrity to the panel and support to its electric output terminals. The schematic of a crystaline solar panel is shown in figure 3.23.

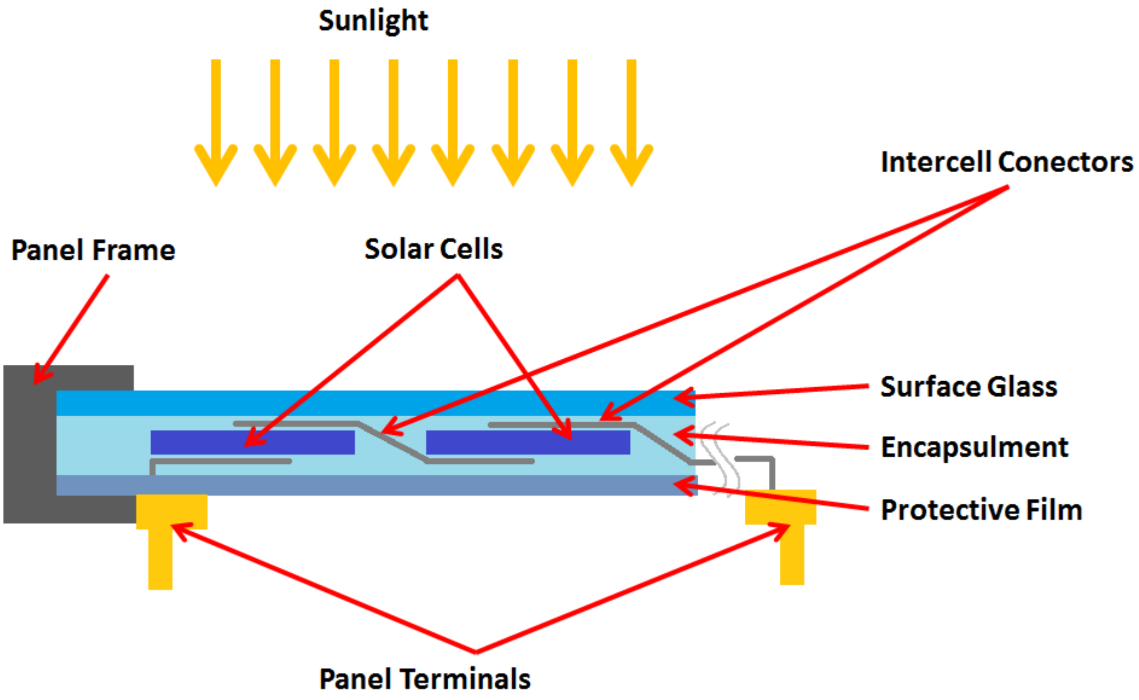


Figure 3.23: Schematic of a crystalline silicon solar panel

The reflectivity of the glass coating is the first component that modifies the incident irradiation, directly reflecting a fraction of it back to the sky. It is modelled by the reflectivity profile of the glass, which is a function of wavelength. Silicon and glass reflective coefficients are provided by [39].

The spectral response is an important cell characteristic that informs how much energy is absorbed from a photon in a given wavelength. It peaks at the optimum wavelength, the point which the photon has the exact energy to move an electron over the band-gap to the conduction band. Shorter wavelengths have excess energy that dissipates through emission of new photons with lower energy and bigger wavelengths. Too large wavelengths simply do not have enough energy to move the electron to the conduction band. The typical spectral response of a silicon cell is shown in Fig. 3.24.

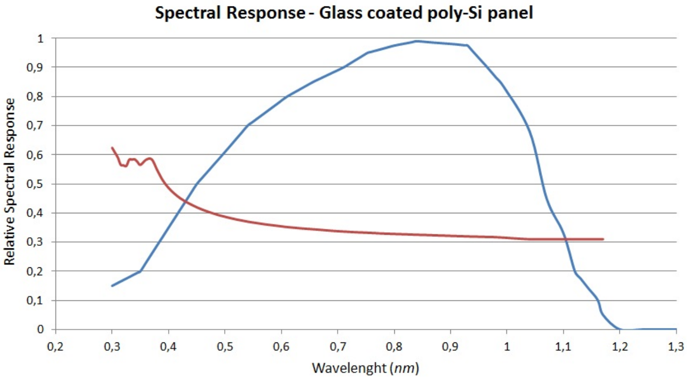


Figure 3.24: Typical spectral response of the polycrystalline silicon cell (blue) and glass reflectivity per wavelenght (red)

In order to predict PV power generation, the first step is to encounter the typical spectral response of a polycrystalline Silicon panel. Points[3] were captured and, by means of simple linear interpolation, the response coefficients for each wavelength can be obtained, giving an approximation of the spectral response function $g(\lambda)$. The imperfections in the ability of solar cells to convert photons into electricity is modeled by the External Quantum Efficiency η_{EQE} (EQE). The maximum EQE of a photosensitive device refers to the percentage of photons hitting the device's photo-reactive surface that produce charge carriers at the peak of the spectral response. It is typically circa 80% for most monocrystalline Silicon, 2 to 3% lower to poly-Si. Maximum quantum efficiency for thin film solar cells depends on the photovoltaic material, and is usually lower than both mono-Si and poly-Si. Due to different semiconductor bandgaps, the optimum wavelengths also depends on the material.

Combining the sky's irradiance $I_{C\lambda}$, the response function $g(\lambda)$, the cell maximum External Quantum Efficiency (EQE) η_{EQE} and the glass spectral reflectivity function $r(\lambda)$ gives the average power absorbed by the cell during a given period of time:

$$u_{CELL} = \frac{1}{\tau} \int_0^{\tau} \int_{\Lambda} I_{C\lambda}(\lambda) \eta_{EQE} g(\lambda) r(\lambda) d\lambda d\tau \tag{3.72}$$

The set of results is then integrated in time (24 hours) in order to bring forth the average daily PV power density, using a 15 minute window, giving the PV absorbed spectrum, illustrated in figure 3.25. This model employs information contained in the

[3] http://sst-solar.com/images/downloads/solarsysdatenblattqcellsQ6LTT.pdf

solar panel's data-sheet, and is applicable to mono-Si, Poly-Si cells. Thin film cells can also be simulated, as long as spectral and quantum efficiency data is available.

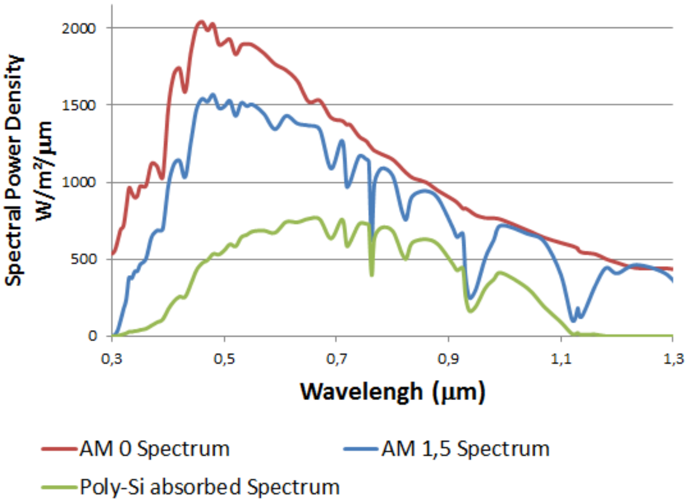


Figure 3.25: Example of Extraterrestrial Solar irradiation (AM0), ASTM Standard Spectrum (AM1,5) in cloudless sky and Absorbed Spectrum by a typical poly-Si cell (AM1,5)

Equation (3.72) models the photovoltaic absorption an losses. Further electric losses occur in the solar panel and inverter due to The model parameters are the effective area, the solar cell fill factor and the panel and inverter thermal coefficients. The resulting solar panel power can then be written in (3.73) as a function of these parameters:

$$u_{PV} = A_{eff}\eta_{FF}\eta_{\Delta T}y_{CELL} \tag{3.73}$$

The effective area A_{eff} is the solar panels' light absorbing area, which is the total area minus the area occupied by electric contacts and structural elements. Most data-sheets for mono-Si and poly-Si cells presents planform drawings of the solar panels which can be employed to obtain the effective area. Due to the different manufacturing process employed for the thin film cells, electrical contacts are not always visible and are sometimes not shown in these data-sheet drawings. An estimate must be made at this case employing information from handbooks or from another data-sheet of a cell made with the same material and with similar efficiency. For the crystalline silicon solar panels simulated in this article, the value of 93% was found by averaging the effective area calculated from three different solar cell manufacturers. The total area can be either taken from the solar plant specifications or estimated from its parameter and PV generation history.

The cell fill factor measures the squareness of the solar cell current-voltage curve. It is a reasonable indirect measurement of the internal quantum and electrical losses, as the physical phenomenons that limit voltage and current at maximum power point are related to charge carrier recombination and internal thermodynamic losses, internal and electrical contact joule effect losses (shunt and series resistance). It can be calculated with the cell's nominal Short Circuit Current I_{SC}, Open Circuit Voltage V_{OC} and Nominal Peak Power P_{Nom}, accordingly to the relation shown in equation (3.74):

$$\eta_{FF} = \frac{P_{Nom}}{V_{OC} I_{SC}} \tag{3.74}$$

As most semiconductor materials have properties highly dependent on temperature and the standards usually demands the panel to be tested at 25 degrees Celsius, temperature coefficients k_T are usually given in data-sheets and are employed to model the efficiency variation due to ambient temperature.

$$\eta_{\Delta T} = 1 - k_T \Delta T \tag{3.75}$$

where ΔT is the difference between the ambient temperature and the 25 Celsius reference.

The maximum quantum efficiency measures the percentage of electrons that get into conduction band for each absorbed photon in the optimum wavelength.

Gathering all these information, we have validated our estimation for the overall panel efficiency by comparing the simulations to the real efficiency of the module under the pattern spectrum AM 1.5. The real efficiency and the simulations results stood in a narrow range around 15%.

3.5.3 State space representation

State space models are relatively rare in papers, but recently scholars have been employing this tool to produce hybrid-physical PV forecasting algorithms [10]. Usually, state space approaches employ the Kalman filter to provide the entire forecast [101], tuning parameters for ANN or machine learning approaches [13], or to simplify and evaluate solar irradiation models [44].

In this dissertation, the Kalman filter is used to combine the PV generation history, weather measurements and estimates of the PV production based on solar irradiation and solar panel models featured in subsections 3.5.1 and 3.5.2. Mathematically, equation (3.51) represents the next day PV generation $y_{PV}[k+1]$ as linearly dependent of an autoregressive term $y_{AR}[k]$, an exogenous input term $u[k]$ with p delays and the noise $z[k+1]$:

$$y_{PV}[k+1] = y_{AR}[k] + \sum_{i=1}^{p} (u[k-i+1]) + z[k+1] \tag{3.76}$$

The input term $u[k]$ is a linear combination of the several weather variables provided by the METARs (Table 3.2), the solar irradiance calculated with SPCTRL2 and SEDES2 shown in equation (3.71), and the estimated PV generation as calculated in equation (3.73). It is advisable to produce more than one estimation of PV generation, using multiple combinations azimuth and tilt angles in order to account for unknown geometric parameters or positioning errors. Unknown parameters require wider separation of the azimuth and tilt pairs, while when compensating positioning uncertainties, the parameters just need to be in the vicinity of the measured instalation angles.

As shown in subsection 3.4.5, it is possible to transform the difference equation (3.76) in a state space model in the companion form, expressed in matricial form as:

$$X_{PV}[k+1] = \begin{bmatrix} a_1 & a_2 & \cdots & a_{n+1} & a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} X_{PV}[k] + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{pm} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} U[k] + W[k] \tag{3.77}$$

The output equation is given by expression (3.78):

$$y_{PV}[k+1] = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} X_{PV}[k+1] + v[k+1] \tag{3.78}$$

Rewriting the equations in matricial form gives the final load State space model:

$$X_{PV}[k+1] = \mathbf{A}X_{PV}[k] + \mathbf{B}U[k] + W[k] \tag{3.79}$$

$$y_{PV}[k+1] = \mathbf{C}X_{PV}[k+1] + v[k+1] \tag{3.80}$$

For the PV generation forecast, the state space formulation allows a wider degree of freedom when manipulating inputs. There is no need for stationarity for the generation and inputs time series, as is required by ARMA and other Box-Jenkins derived methods. Due to its linear formulation, it is more resilient to the curse of dimensionality than ANN and machine learning approaches. State space also is used in conjunction with Kalman filter predictors, which adds robustness through the application of an double state/measurement noise model and is a dependable data fusion technique, a useful trait for adaptive hybrid-physical forecasting with several input variables.

4 LOAD AND GENERATION FORECASTING

Economic development, throughout the world, depends directly on the availability of electric energy, especially because most industries depend almost entirely on its use. The availability of a source of continuous, cheap, and reliable energy is of foremost economic importance. Load forecasting is vitally important for the electric industry in the deregulated economy. It has many applications including energy purchasing and generation, load switching, contract evaluation, and infrastructure development. A large variety of mathematical methods have been developed for load forecasting. In this chapter, various approaches to load forecasting are discussed.

High renewable energy penetration grids are challenging to balance due to inherently variable generation weather-dependent energy resources. Forecasting photovoltaic generation is a tool for mitigating resource uncertainty and reducing the need for scheduling of ancillary generation. Several forecasting methodologies have been developed to target different forecast time horizons.

The objective of this chapter is to study the dynamic state estimation problem and its applications to electric power system analysis. Furthermore, the different approaches used to solve this dynamic estimation problem are also discussed in this chapter. Section 4.1 proposes the Kalman based forecasting algorithm, while Section 4.2 deals with photovoltaic generation forecasting.

4.1 Load Forecasting

Load forecasting is way of estimating what future electric load will be for a given forecast horizon based on the available information about the system. The forecast horizon refers to the prediction time horizon, which can be long-term, medium-term or short-term. While there are not normative boundaries between these three horizons, authors usually define long term as forecasts aiming at load prediction for more than a year ahead, medium term from one week and up to one year ahead, and short term forecasts as predictions targeting the next hours and up to one week in the future [47, 90]. In this work, only short-term load forecasting is analysed.

In short-term load forecasting (STLF), the future load on a power system is predicted by extrapolating a predetermined relationship between the load and its influential variables, namely time and/or weather. Determining this relationship is a two stage process that requires identifying the relationship between the load and the related variables and quantifying this relationship through the use of a suitable parameter estimation technique. A prerequisite to the development of an accurate load-forecasting model is an in-depth understanding of the characteristics of the load to be modeled. This knowledge of the load behavior is gained from experience with the load and through statistical analysis of past load data. Utilities with similar climatic and economic environments usually experience similar load behavior, and load models developed for one utility can usually be modified to suit another.

However, as shown in Chapter 3, the number of variables which are are related to the load and/or the photovoltaic generation can amount to a very large number of inputs. This poses a risk of overfitting the model due to the so-called curse of dimensionality, and relates to the fact that the convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow [60], requiring an unpractical volume of data to optimize the model parameters. In that sense, it is advisable to perform a feature selection procedure to reduce the dimensionality of the prediction problem.

In this work, a methodology that combines feature selection by means of Principal Component Analysis (PCA), prediction by a modified Kalman filter with series Gray regression and full variance tracking is proposed. The forecasting system is illustrated in the schematics shown in Figure 4.1:

The general data model assumes three main sets of variables: input, output and measurement variables, as shown in Figure 4.1. The raw input vector $\widehat{U}[k]$ drive the real system, while the forecasting scheme is driven by the $U[k]$ input, a transformation of $\widehat{U}[k]$ that creates new variables through the nonlinear transformations discussed in Chapter 3, then applies a normalization procedure and perform principal component analysis to reduce dimensionality. The input $U[k]$ can be corrupted by the noise term $W[k]$, which represents uncertainties about the filter state. The measurement variable $y[k]$ represent the output of the real system corrupted by a measurement noise $v[k]$. Both are also inputs for both the modified Kalman filter and the Grey model predictor. The Grey model produces a forecast $\widehat{y}_G[k]$ that is used to enhance predictions for the modified Kalman filter state. If the filter contains a reasonable state space model of

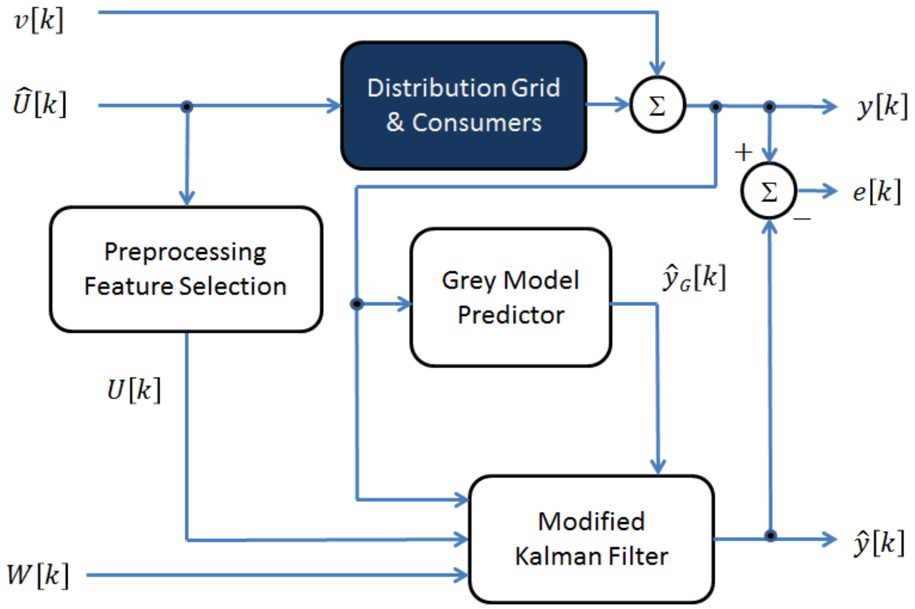


Figure 4.1: Proposed data model for Load forecasting

the real system, its output $\hat{y}[k]$ is a forecast of the real system output. The prediction error $e[k]$ can be obtained by subtracting $\hat{y}[k]$ of $y[k]$. In this paper, k denotes the process time step, which is equal to a day.

The load prediction is performed for two different electric distribution systems, located in the cities of Brasilia and Leipzig, in Brazil and Germany, respectively. Brasilia's time series contains information about peak, average and base load for the period between 2001 and 2010, aggregated as a single substation. Leipzig demand history contains similar data from 2001 to 2003, however measured in eight substations.

For both cities, the forecasting system employs an extensive set of candidate input variables. Eight different input sets are used to evaluate the impact of the additional variables. These sets are labeled from "A" to "Z" in order to shorten the notation in the Chapter 5. The nine input sets, their designation, where they are listed, the number of variables and their labels are shown in Table 4.1:

Input set A is only concerned with temperatures, and represent the classical short-term load forecast variables. Input set B uses all variables present in the METAR reports, taking full advantage of the several measurements taken in the aerodrome weather stations. Input set C contains the temperature in both logarithm and degree-days parametrizations. Input set D contains all weather variables discussed in subsection 3.4.1, set E the variables discussed in subsection 3.4.2, set F concerns tariffs and is

Table 4.1: Listing of the candidate input variables

Set description	Variables listed in	Size	Input Sets
Temperatures	Eq. (3.12)	3	A
METARs	Table 3.2	22	B
Degree-days	Eq. (3.14)	10	C
Full weather	Eqs. (3.14), (3.25), (3.31), (3.33) and (3.37)	39	D
Socioeconomic	Table 3.7:	15/3	E
Tariffs	Eq. (3.43)	75/0	F
Events	Table 3.8	14	G
Sunlight	Eq. (3.33) and Eq.(3.37)	10	H
All	All variables from input sets D, E, F and G.	153/66	Z

discussed in subsection 3.4.3 and input set G the event variables described in subsection 3.4.4. Set H deals with solar irradiation and illuminance variables, described in subsections3.4.1.4 and 3.4.1.5. The last input set, "Z", is the union of sets D, E, F, G and H. For load forecasting in Brasilia, all inputs are available, amounting to 153 variables in input set Z. For Leipzig, forecasts are done with a maximum of 66 variables, because tariff history is not available and only 3 socioeconomic variables are employed: population, GDP and GDP per capita.

The state space model can accomodate input delays in the form of extra inputs generated by cascade lag operators. In the proposed forecasting algorithm, however, the state space model parameters are unknown. Every additional input means an additional coefficient that requires periodical optimization. Adding too much inputs is thus detrimental to the quality of the parameter fitting, either increasing the estimation errors in the model coefficients or requiring a larger set of data to obtain a given precision in the parameter optimization. It is thus advisable to reduce the model dimensionality to enhance its computability and forecast accuracy, which is accomplished by means of preprocessing and feature selection.

The underlying nonlinearities of a power system and some of its physical parameters are usually known a priori. However, some of the, mostly minor, nonlinearities cannot be modelled accurately due to the system complexity and constraints on physical ability to measure. This is thus seen as a partially known system and may be modelled as a grey box [59]. The proposed forecasting algorithm employs an autoregressive modified rolling grey model to account part of these nonlinearities and unknown dynamics. This grey box model adds its prediction as an additional input to the Kalman filters.

The remainder of this section is divided in three parts. In subsection 4.1.1, the preprocessing and feature selection steps are explained. In subsection 4.1.2, the Grey model predictor is presented. Subsection 4.1.3 details the Kalman filter application used to provide the load forecasts.

4.1.1 Preprocessing and Feature selection

The feature selection routine begins with a preprocessing step that prepares the candidate variables to be combined and selected in the principal component analysis, normalizing mean and variance of the candidate inputs.

The mean and variance normalization is a simple procedure designed to enforce uniformity in the amplitude scale of the candidate variables, except for those boolean. This is important to minimize numerical errors. Taking a sample of a given length n of the ith candidate variable $\hat{U}_{0i}$, which has mean $\overline{U_{0i}}$ and variance σ_{0i}^2, it can be normalized to zero mean and unitary variance by the linear operation as follows:

$$\hat{U}_{1i} = \frac{(\hat{U}_{0i} - \overline{U_{0i}})}{\sigma_{0i}} \tag{4.1}$$

This forecasting system employs PCA to search and select the input variable set that better explains the variance in electric demand, by means of linear combination of the candidate variables that generate a set of orthogonal inputs, called principal components. A method to reduce dimensionality is to select the j components with higher variance that explain a given percentage of the candidate set total variance, discarding the other components altogether.

Composed of more than two hundred variables, the original set displays high crosscorrelation between the input themselves, as presented in Fig. 4.2.

PCA is applied at a training sample of the d_0 candidate variables, assembled in this forecasting system from their previous training period values. The size of d_0 can be as high as 280, when all candidate inputs presented in Chapter 3 are used. The objective is to reduce the dimensionality of the input set from d_0 to d.

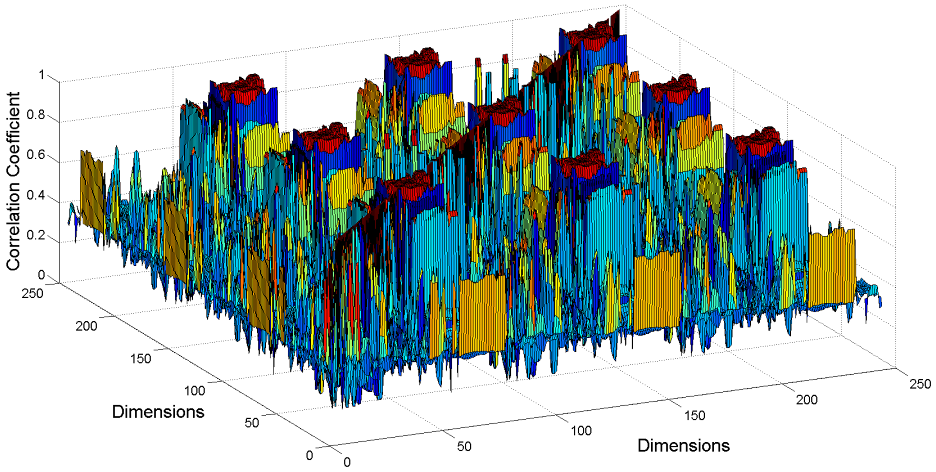


Figure 4.2: Correlation between several candidate variables and peak demand, which corresponds to the projection in the planes $x = 0$ or $y = 0$

By means of a SVD decomposition, the left-singular vectors, the singular values and the right-singular vectors are obtained. The d singular values that represent a given percentage of the total variance are selected, their quantity determining the dimension in the selected input set. The left and right-singular vectors are then employed to produce the transformation matrix T. Size of d is chosen by exhaustive search, as a compromise between the mean and maximum error metrics, mitigating overestimation.

For prediction, as the next day d_0 values of the candidate variables become available, they are transformed by T in a optimized input of d variables, which are used for the prediction of next day electric load.

In order to adapt to seasonal variations, this process is repeated at every model update iteration. Illustrating the reduction in dimensionality, the crosscorrelation of an optimized input set with 126 variables is shown in Fig. 4.3, obtained from 280 candidate inputs at the first iteration.

4.1.2 Grey model forecasting

During the last two decades, the grey systems theory has been showcased in several papers [59]. Its main advantage is the ability to deal with partially parametrized nonlinear systems without requiring vast amounts of high quality information. It has been widely and successfully applied to various systems in the most diverse fields, such as science, technology, economics, finance, sociology and forecasting. In [38], a Grey-

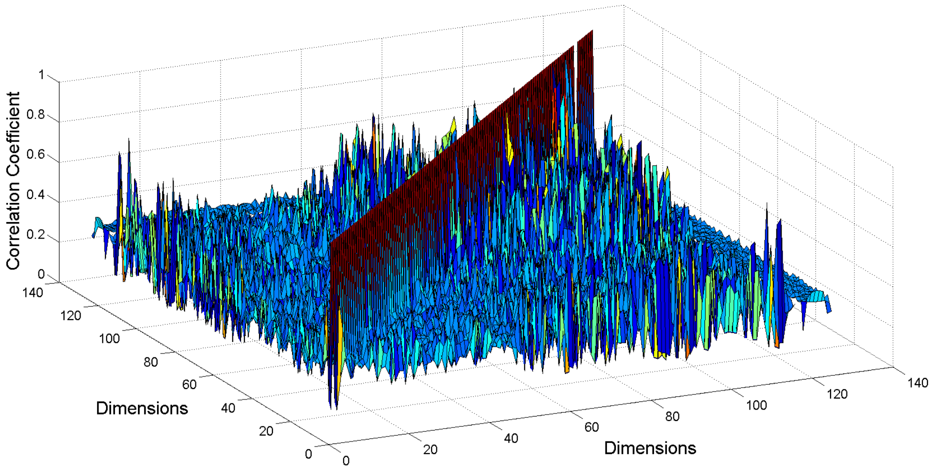


Figure 4.3: Correlation between the 126 selected variables and peak demand, which corresponds to the projection in the planes $x = 0$ or $y = 0$

regression variable weight combination model achieves good precision for MTLF and LTLF without needing additional explicative variables. [106] proposes a grey model with a time varying weighted generating operator to extract information concealed in recent data. The method is validated in five case studies, the first regarding hourly prediction of a grid connected photovoltaic system and the third applying the method to forecast Russia's yearly energy consumption.

In this work, the Grey model is employed to enhance the Kalman based predictions, adding robustness and support to nonlinearities and unknown dynamics. As this algorithm is executed in series with the Kalman filter and also requires parameter optimizations, the simplest autoregressive case is chosen, leaving the input processing to the State space model.

Using the same state vector $X[k]$ (size N) defined in equation (3.60) as the Grey input, the second step recursively employs the FGM(1,1) model presented in [98] to forecast the next day load. In order to extract more information from the Grey input, a constant is concatenated as the first entry in X to form X_F, as shown in eq. (4.2).

$$X_F[i] = \begin{cases} 0 & i = 1 \\ X[i-1] & i = 2, 3, ..., N+1 \end{cases} \tag{4.2}$$

The accumulated generating operation (AGO) is then applied to the grey input by means of eq. (4.3) to generate the intermediate variable X_G:

$$X_G(i) = \sum_{j=1}^{i} X_F(j), \ i = 1, 2, ..., N \tag{4.3}$$

The Grey exponential model, based on X_G is generated by eq. (4.4):

$$\frac{dX_G[k]}{dk} + \alpha X_G[k] = \beta \tag{4.4}$$

Eq. (4.4) is called the first order Grey differential equation, where the Grey developmental coefficient α and Grey control parameter β constants have to be estimated. The solution with initial condition $X_F[1] = 0$ is given by eq. (4.5):

$$X_G[k+1] = \frac{\beta}{\alpha}\left(1 - e^{-\alpha k}\right) \tag{4.5}$$

The developmental coefficient and Grey control parameter are determined by least-squares method in eq. (4.6):

$$[\alpha, \beta]^T = (F^T F)^{-1} F^T X_G \tag{4.6}$$

where F is defined by (4.7):

$$F = \begin{bmatrix} -0.5(X_G(1) + X_G(2)) & 1 \\ -0.5(X_G(2) + X_G(3)) & 1 \\ \vdots & \vdots \\ -0.5(X_G(N) + X_G(N+1)) & 1 \end{bmatrix} \tag{4.7}$$

With α and β obtained, eq. (4.5) can be used to forecast a future value of the intermediate variable X_G. As the first element of the state vector $X[k]$ is also the output variable, performing an inverse AGO (eq. (4.8)) over the predicted $X_G[k+1]$ yields a forecast $\hat{Y}_G[k+1]$ for the future electric load $Y[k+1]$:

$$\hat{Y}_G[k+1] = X_G[k+1] - X_G[k] \tag{4.8}$$

The grey prediction $\hat{Y}_G[k+1]$ is then used as an additional input in the Kalman based predicting algorithm.

4.1.3 Proposed Kalman based adaptive prediction scheme

Unlike the batch filters, which consider the complete set of data in order to assemble a model and then forecasting future data, Kalman filters are adaptative. Therefore, after each prediction step, new values for state space parameters are updated in Kalman filters. Such feature reduces considerably the computational complexity of the Kalman filter in comparison with Box-Jenkins time series approaches, such as AR and ARMA models.

In addition, the input variables of AR and ARMA filters should be stationary and unbiased. The bias is frequently removed by means of differentiation, which cannot be applied to data with exponential behavior, as population. Kalman filters, instead, are able to work with every sort of data without previous mathematical treatment. However, the large amount of input variables would end up bringing distortion to the comparison between batch and Kalman filters. Therefore, a first step was adopting the same set of batch filters inputs for Kalman ones. This prevents that different data structure leds to misleading conclusions.

The Kalman filter, in its most basic form, is a linear recursive data processing algorithm that makes optimum estimates of a variable of interest by combining the knowledge of system dynamics (embedded in its state-space model), the statistical description of system noises and measurement errors and the information about the initial conditions of the system [56]. The state-space representation is a discrete time domain model that relates inputs, output and state variables through two sets of difference equations, shown in (4.9) and (4.10).

$$\mathbf{X}[k+1] = A_K\mathbf{X}[k] + B_K\mathbf{U}[k] + \mathbf{W}[k] \tag{4.9}$$

$$\mathbf{Y}[k+1] = C_K\mathbf{X}[k+1] + D_K\mathbf{U}[k+1] + \mathbf{V}[k+1] \tag{4.10}$$

This work presents an application of the discrete-time Kalman Filter as a load forecasting tool that does not need information about the distribution grid topology. This filter combines the state space model, initial conditions, the previous electricity demand time series and its related exogenous input variables as presented in Appendix A. The output is a daily recursive prediction for electricity demand in the next day. The processing is done in 4 phases and 8 steps.

- Phase Zero - Model Optimization
  - Step 1: Calculate State Space Model Coefficients
- Phase I - Prediction (occurs before the observation)
  - Step 2: State Vector Estimation;
  - Step 3: Error Covariance Matrix Estimation;
- Phase II - Filter Update (occurs after the observation)
  - Step 4: Kalman Gain determination;
  - Step 5: State Vector update with output observation;
  - Step 6: Error Covariance Matrix update with output observation;
- Phase III - Variance Estimation
  - Step 7: R vector Estimation;
  - Step 8: Q Matrix Estimation.

In the predicting scheme, each time step represents a day. At the very first time step, an initialization procedure is executed in order to set the filter model order and initial conditions. Phase Zero is not performed at every time step, as it is the most computationally expensive phase. It is perfomed for the first filter iteration, and then at every T-nth time step it is executed in order to update the state space model coefficients. Phases I to III are recursively performed at each time step. The process flow is shown in Fig. 4.4.:

Each phase produces its outputs using its specific inputs, which in turn are measurement values or the outputs previously generated by the other phases. The relationship between these inputs, outputs and predicting phases are shown in Table 4.2, as well as the data flow.

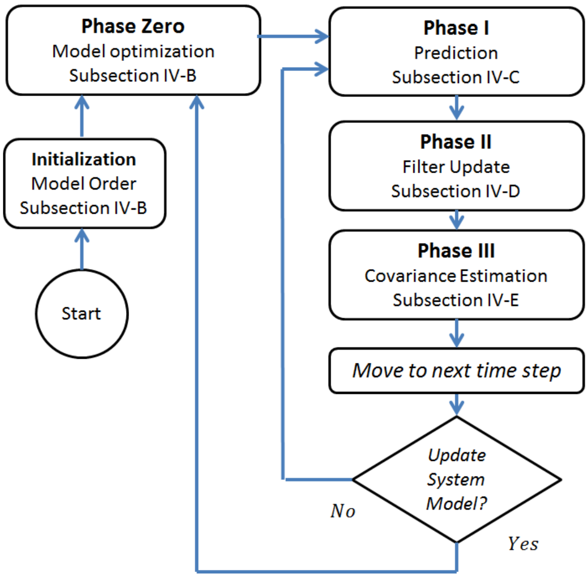


Figure 4.4: Box Diagram of the proposed Kalman based predicting scheme

Table 4.2: Inputs and outputs per predicting phase

Phase	Inputs	Outputs
Initialization	Demand history Exogenous inputs history	Model Order Initial conditions
Phase Zero	Model Order Initial conditions Previous Demand Previous exogenous inputs	Model parameters
Phase I	Model parameters Initial conditions Updated state vector Updated error covariance Q matrix estimate	Demand Prediction State vector estimate Error covariance estimate
Phase II	State vector estimate Error covariance estimate Demand Measurement R covariance estimate	Updated state vector Updated error covariance Kalman Gain
Phase III	State vector estimate Updated state vector Demand Prediction Demand Measurement	Q matrix estimate R covariance estimate

The initialization procedure requires a sizeable sample of both demand and exogenous inputs historical time series in order to select the model order and provide reasonable initial conditions for Phase Zero and Phase I. These initial conditions are only needed for the first time step, while the selected model order is permanently used by the predicting scheme. For the second time step onwards, Phase Zero employs previous values of demand and exogenous inputs, which were sampled during the predicting scheme operation. Phase I relies on the corrected state vector, error covariance and estimated Q matrix, which provide all the information needed to perform the prediction steps.

4.1.3.1 State space Model

The Kalman filter [50, 56] requires a state-space representation of the system. It is a discrete time domain model that relates inputs, output and state variables through two sets of difference equations, shown in (4.9) and (4.10). In this Kalman based predicting scheme, the following choices and premises were chosen to simplify the optimization of the model parameters:

1. The state variables are the last N values of the Electricity Demand, where N is the model order;
2. The first state variable is a linear combination of all state variables and the inputs;
3. The inputs only affect the first state variable;
4. The first state variable is also the Prediction Output
5. The inputs do not affect the Prediction Output (no instant transmission term).

Analisys of the electricity demand time series shows that there is a correlation between the demand in a particular day and the demand of the previous days. Also there is correlation between it and some exogenous variables related to climate, population and economy. The core of this predicting scheme is to represent the demand as a linear combination of its previous values and the exogenous inputs. The particular choice of state variables is employed in order to permit direct calculation of this linear combination. That also justifies why the first state variable is modelled to be the predicted output. As the exogenous inputs can not affect the demand of the previous days, those inputs could only act over the current prediction. Lastly, as the output and the first state variable, there is no need for direct coupling between inputs and output.

A state-space model that perfectly fits this situation is one in the canonical controllable form, except for the lack of normalization towards Matrix B_K. The first state variable is updated at every time step as a linear combination of all state variables and inputs. The other state variables are the last $N-1$ values of the output. The output at every step is arbitrarily set as equal to the first state variable, and no instant transmission term (Matrix D_K) is employed. Mathematically, this model has the representation shown in (4.11) and (4.12).

$$\mathbf{X}[k+1] = A_K\mathbf{X}[k] + B_K\mathbf{U}[k] + \mathbf{W}[k] \tag{4.11}$$

$$\mathbf{Y}[k+1] = C_K\mathbf{X}[k+1] + \mathbf{V}[k+1] \tag{4.12}$$

Disregarding the noise inputs $\mathbf{W}$[k] and $\mathbf{V}$[k], the matricial representation for this model is shown in equations (4.13) and (4.14):

$$\begin{bmatrix} x_1[k+1] \\ x_2[k+1] \\ \vdots \\ x_n[k+1] \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{n-1} & a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1[k] \\ x_2[k] \\ \vdots \\ x_{n-1}[k] \\ x_n[k] \end{bmatrix}$$

$$+ \begin{bmatrix} b_1 & b_2 & \cdots & b_m \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} u_1[k] \\ u_2[k] \\ \vdots \\ u_m[k] \end{bmatrix} \tag{4.13}$$

$$y[k+1] = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1[k+1] \\ x_2[k+1] \\ \vdots \\ x_n[k+1] \end{bmatrix} \tag{4.14}$$

In equations (4.9) and (4.13), Matrix A_K models the dependence of the next state (and consequently the next output) with the current system states. In a comparison with the batch Schemes, it models the autorregressive behaviour of the system. Matrix B_K models the dependence of the next state with the current system inputs. The remaining $\mathbf{W}$[k] terms models the imperfections in the state-space model, as uncertainties in the inputs, non-linearities and intrisic stochastic process that occur in the real system. Ideally, $\mathbf{W}$[k] approaches a normally distributed random vector with zero mean and covariance matrix Q. In equation (4.10) and (4.14), Matrix C_K models the coupling between the output and the system state. The variable $\mathbf{V}$[k] models noise in the energy demand measurements, as well as imperfections in the state to output coupling. Ideally, $\mathbf{V}$[k] approaches a normally distributed random variable with zero mean and variance R.

4.1.3.2 Initialization and Phase Zero

The initialization procedure sets the initial parameters that the Kalman based predicting scheme needs in order to operate reliably. The first parameter to be set is the model order, which sets how many state variables are to be employed. As the training dataset, the scheme needs from 180 to 365 time steps of past data, which are the previous electricity demand and exogenous input time series. A range of candidate model orders is then simulated over the training dataset, and total squared error (TSE) of predictions is calculated. The candidate model order N that achieves the lowest sum of TSE's is selected for the predicting scheme. The system state, error covariance matrix, Q and R parameter values are stored in order to provide initialization values for the forthcoming processing phases.

Phase Zero is performed after the scheme's initialization and at every t time steps. Its objective is to optimize the coefficients of the state space model. Having defined the model order N, it is needed to determine the matrices' elements. Due to the specific state space representation that is employed for the predicting scheme, shown in (4.13) and (4.14), only the elements in the first row of A_K and B_K matrices must be determined by linear least squares. Isolating the first row terms in equation (4.13) and ignoring the noise terms, one obtains equation (4.15):

$$x_1[k+1] = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}^T \begin{bmatrix} x_1[k] \\ \vdots \\ x_n[k] \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}^T \begin{bmatrix} u_1[k] \\ \vdots \\ u_m[k] \end{bmatrix} \tag{4.15}$$

From (4.15) and due to the structure of C_K, it is straightforward that:

$$y[k+1] = \begin{bmatrix} y[k] \\ \vdots \\ y[k+1-n] \\ u_1[k] \\ \vdots \\ u_m[k] \end{bmatrix}^T \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{4.16}$$

For the sake of stability and precision, the elements are computed via an iterative Generalized Minimum Residual (GMRES) method.

4.1.3.3 Phase I - Prediction

Phase I is performed in two steps, resulting in a estimate of the system state $\hat{X}[k+1]$, a prediction $\hat{y}[k+1]$ for the electricity demand in next day and in a estimate of the error covariance matrix $\hat{\mathbf{P}}[k+1]$. electricity demand prediction for the next day. It is important to notice that Phase I occurs before observation, which means that the estimates and prediction are calculated before the measurement of the electricity demand $y[k+1]$ is available.

The first step employs equations (4.17) and (4.18) to estimate the system state $\hat{X}[k+1]$ and calculate the predicted output $\hat{y}[k+1]$ for next time-step. This estimation employs the stored data $X[k]$ of current time-step state and the exogenous inputs $U[k]$. Climate variables are calculated with the forecasted values of temperature for the next day, population variable is extrapolated from the census trends and the calendar variables have exact values.

$$\hat{\mathbf{X}}[k+1] = A_K\mathbf{X}[k] + B_K\mathbf{U}[k] \tag{4.17}$$

$$\hat{\mathbf{y}}[k+1] = C_K\hat{\mathbf{X}}[k+1] \tag{4.18}$$

The second step evaluates the impact of the system noise over the predictions. This objective is reached with the estimation of a Error Covariance Matrix $\hat{\mathbf{P}}$[k + 1] for the estimated values of the state variables. The estimation employs the A_K matrix, the corrected Error Covariance Matrix calculated in the previous time-step $\mathbf{P}$[k] and $Q[k]$, the current estimation for covariance of $\mathbf{W}$[k]. The step is shown in equation (4.19).

$$\hat{\mathbf{P}}[k+1] = A_K\mathbf{P}[k]{A_K}^T + Q[k] \tag{4.19}$$

4.1.3.4 Phase II - Filter Update

The Filter update phase occurs after the measurement of electricity demand, performed in three steps. It compares the prediction $\hat{y}[k+1]$ with the measured value $y[k+1]$ and with this information the estimated state $\hat{X}[k+1]$ and error covariance matrix $\hat{\mathbf{P}}$[k + 1]are respectively updated to$X[k+1]$ and $\mathbf{P}$[k + 1].

The first step evaluates the probable impact of the observation's variance to the correction of the state estimation. The Kalman Gain is the wheighting factor by which it is determined how much the observation will be taken into account when updating the State Vector and the Error Covariance Matrix. The higher the observation error variance $R[k]$, less confidence will be placed over the observation values to update the filter state. The Kalman Gain can be obtained by equation (4.20):

$$\mathbf{K}[k+1] = \hat{\mathbf{P}}[k+1]{C_K}^T(C_K\hat{\mathbf{P}}[k+1]{C_K}^T + R[k])^{-1} \tag{4.20}$$

The second step corrects the estimated system state with the observation information wheighed in by the Kalman Gain, which is given in (4.21).

$$\mathbf{X}[k+1] = \hat{\mathbf{X}}[k+1] + \mathbf{K}[k+1](\mathbf{Y}[k+1] - C_K\hat{\mathbf{X}}[k+1]) \tag{4.21}$$

The third step of Phase II updates the state estimate error covariance$\mathbf{P}$ with the observation information, also weighed in by the Kalman Gain, as shown in Eq. (4.22).

$$\mathbf{P}[k+1] = (I - \mathbf{K}[k+1]C_K)\hat{\mathbf{P}}[k+1] \tag{4.22}$$

In (4.22), **I** represents the Identity Matrix.

4.1.3.5 Phase III - Variance Estimation

One of the biggest challenges to Kalman filtering schemes is the determination of suitable values for the Q and R covariance terms. Previous knowledge of these parameters is seldom available, specially when the model does not represent a definite physical system. Phase III adresses this shortcoming in this proposed Kalman based predicting scheme. There is a recursive procedure that estimates the most probable value for R and Q at every time step.

In the first step, a R variance tracking routine was employed based on the estimation of V[k]. Isolating it in (4.10) gives the equation (4.23):

$$\mathbf{V}[k] = \mathbf{y}[k] - C_K\mathbf{X}[k] \tag{4.23}$$

It is then possible to estimate $\mathbf{V}$[k] by subtracting the predicted output $C_K\mathbf{X}$[k] of the measured output $\mathbf{Y}$[k]. By definition, R is the variance of the$\mathbf{V}$[k] from the first to the k-th time step. As the demand measurements are consequence of a very high number of stochastic process (multiple loads, multiple measurement systems, faults, grid losses and reading errors), one can suppose that abrupt changes in statistical parameters of a isolated process does not necessarily translates into a abrupt change of the statistical parameters of the measurement process. As it is very unlikely that several of those stochastic processes will change in coordination, one can conclude that abrupt variations in the R parameter are also improbable. This approximate continuity is modelled in the tracking routine by weighing in the value of R estimated for the previous step, as shown in (4.24):

$$R[k+1] = k^{-1}R[k] + (k-1)k^{-1}Var(\mathbf{V}[k]) \tag{4.24}$$

In the second step, the estimation of the Q Covariance Matrix starts by isolating the $\mathbf{W}$[k] vector from its definition:

$$\mathbf{W}[k] = \mathbf{X}[\mathrm{k}] - \hat{\mathbf{X}}[k] \tag{4.25}$$

Also by definition, Q is the covariance matrix of the vector $\mathbf{W}$[k]. Considering also that Q does not change abruptly, a similar weighing routine is employed to determine it. However, X[k] is a function of the Kalman Gain (4.21), which in its turn is a function of R. As by definition Q and R measure diferent model imperfections, they are thus modelled as independent variables and it is necessary to subtract the$R$ variance from Q in the innovation $\triangle Q$:

$$\triangle Q = \sqrt{(Var(\mathbf{W}[k]) - I_n \cdot Var(\mathbf{V}[k]))^2}$$

$$Q[k+1] = k^{-1}Q[k] + (k-1)k^{-1}\triangle Q \tag{4.26}$$

Where I_n denotes the identity Matrix of order n. After this last update, the algorithm can move ahead to the next Time Step (which would be$k+2$) and repeat the process, starting from step 1.

4.2 Photovoltaic Generation Forecasting

It is a widely reported fact that photovoltaic (PV) energy has been undergoing a rapid development in recent years [85, 105]. Unlike conventional power sources, PV electricity output is not dispatchable, as it depends entirely on the solar irradiance incident over the solar panels, which is a stochastic variable. Integration of this kind of intermittent energy sources is challenging in terms of power system management in both large and small grids. Indeed, PV energy is a variable resource that is difficult to predict due to meteorological uncertainty. As such, being able to predict the future behavior of a PV plant is very important in order to schedule and manage the alternative supplies and the reserves.

The main challenge of forecasting PV generation is its variability. Apart from occasional technical failures, conventional sources are easily dispatchable in the sense that future production can be precisely planned. This is not the case with PV power, which closely depend on the solar resource, site geography and weather conditions. Extensive reviews of the state of the art in solar power forecasting are available in [10]. Forecasting methodologies can be largely characterized as physical or statistical. The physical approach combines solar irradiation and PV system models to predict generation, whereas the statistical approach primarily confides on past data to generate forecast, with little or no reliance on irradiance and PV models. Hybrid approaches employ both irradiation and PV modeling with time series analisys.

In this work, a Kalman based adaptive method for day ahead short-term PV generation forecasting is presented. Very similar to the methodology developed to electric load forecasting, the predicting algorithm combines feature selection with PCA, autoregressive Grey box modeling and a modified adaptive Kalman filter, producing a robust yet computationally light algorithm to forecast PV production. Expanding on recent applications of Kalman filters and state space modeling for photovoltaic forecast [44, 13, 101], the proposed method employs extended input sets comprised of weather measurements and solar irradiation estimations obtained from SPCTRL2 and SEDES2 models [16, 77]. The input set can be generated from either a single weather or from a group of weather stations. The forecasting system is illustrated in the schematics shown in Figure 4.5:

The general data model assumes three main sets of variables: input, output and measurement variables, as shown in Figure 4.1. The raw input vector $\widehat{U}[k]$ drive the real system, while the forecasting scheme is driven by the $U[k]$ input, a transformation of $\widehat{U}[k]$ that creates new variables through the nonlinear transformations detailed in Apendix B and discussed in Section 3.5, then applies a normalization procedure and perform principal component analysis to reduce dimensionality. $\widehat{U}[k]$ is approximated by the weather measurements taken from one or from several weather stations. The input $U[k]$ can be corrupted by the noise term $W[k]$, which represents uncertainties about the filter state. The measurement variable $y[k]$ represent the output of the real

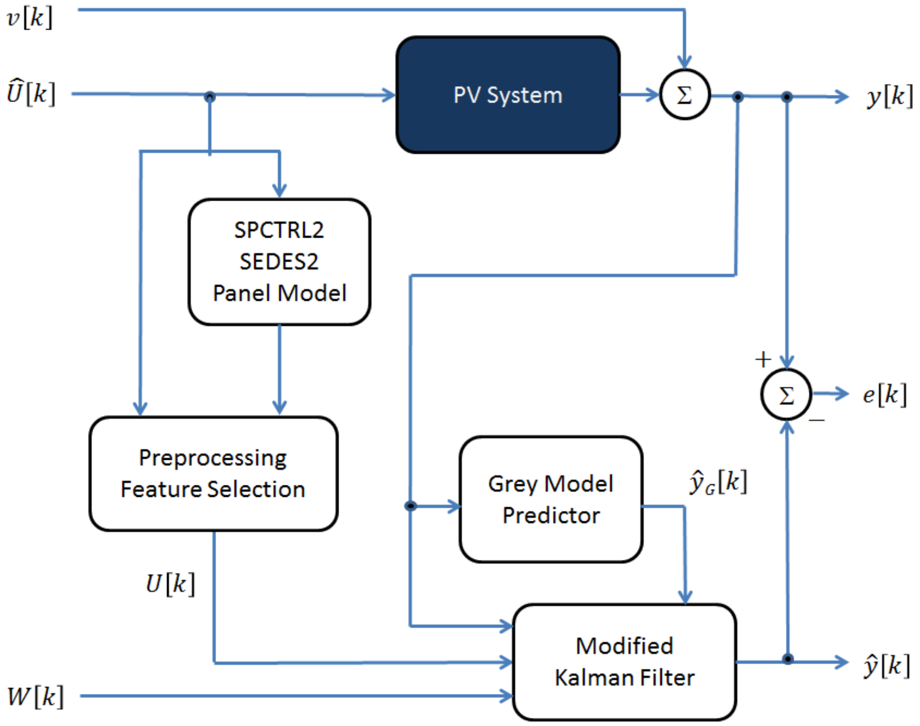


Figure 4.5: Proposed data model for Load forecasting

PV system corrupted by a measurement noise $v[k]$. Both are also inputs forthe modified Kalman filter and the Grey model predictor. The Grey model produces a forecast $\widehat{y}_G[k]$ that is used to enhance predictions for the modified Kalman filter state. If the filter contains a reasonable state space model of the real system, its output $\hat{y}[k]$ is a forecast of the real PV system generation. The prediction error $e[k]$ can be obtained by subtracting $\hat{y}[k]$ of $y[k]$. In this paper, k denotes the process time step, which is equal to a day.

4.2.1 Grey box model for PV

In this work, the Grey model is employed to enhance the Kalman based predictions, adding robustness and support to nonlinearities and unknown dynamics. As this algorithm is executed in series with the Kalman filter and also requires parameter optimizations, the simplest autoregressive case is chosen, leaving the input processing to the State space model.

Equations (4.2) to (4.7) are evaluated at each time step, yielding the Grey box prediction for the PV generation:

$$\hat{Y}_G[k+1] = X_G[k+1] - X_G[k] \tag{4.27}$$

4.2.2 Proposed Kalman based adaptive PV prediction scheme

The Kalman filter [56] is a time domain technique that relates inputs, output and state variables through two sets of difference equations, (4.9) and (4.12). In this PV application, the predicting algorithm consists of the recursive repetition of Eqs. (4.28) to (4.33). Symbols with a hat stand for predictions, while its absence represent a corrected estimation. **K** is the Kalman gain, **P** is the error covariance matrix for the state estimate **X**, and $\mathbf{I_N}$ denotes the identity matrix of order N.

$$\hat{\mathbf{X}}[k+1] = A\mathbf{X}[k] + B\mathbf{U}[k] \tag{4.28}$$

$$\hat{\mathbf{y}}[k+1] = C\hat{\mathbf{X}}[k+1] \tag{4.29}$$

$$\hat{\mathbf{P}}[k+1] = A\mathbf{P}[k]A^T + \mathbf{Q}[k] \tag{4.30}$$

Note that Eqs. (4.28) to (4.30) are calculated before the measurement of the electricity demand, while the remaining filter equations improve the predictions with the information gained by the measurement.

$$\mathbf{K}[k+1] = \hat{\mathbf{P}}[k+1]C^T\left(C\hat{\mathbf{P}}[k+1]C^T + R[k]\right)^{-1} \tag{4.31}$$

$$\mathbf{X}[k+1] = \hat{\mathbf{X}}[k+1] + \mathbf{K}[k+1](\mathbf{Y}[k+1] - C\hat{\mathbf{X}}[k+1]) \tag{4.32}$$

$$\mathbf{P}[k+1] = (\mathbf{I} - \mathbf{K}[k+1]C)\hat{\mathbf{P}}[k+1] \tag{4.33}$$

Adding to the original set of Kalman filter equations, the predicting block also employs variance estimation steps, shown in Eqs. (4.34) to (4.38)

$$\mathbf{V}[k] = \mathbf{y}[k] - C\mathbf{X}[k] \tag{4.34}$$

$$R[k+1] = k^{-1}R[k] + (k-1)k^{-1}Var(\mathbf{V}[k]) \tag{4.35}$$

$$\mathbf{W}[k] = \mathbf{X}[\mathrm{k}] - \hat{\mathbf{X}}[k] \tag{4.36}$$

$$\triangle\mathbf{Q} = \sqrt{(Var(\mathbf{W}[k])^2 - \mathbf{I_N} \cdot Var(\mathbf{V}[k])^2)} \tag{4.37}$$

$$\mathbf{Q}[k+1] = k^{-1}\mathbf{Q}[k] + (k-1)k^{-1}\triangle\mathbf{Q} \tag{4.38}$$

After Eq. (4.38), the algorithm moves ahead to the next time step and repeat the process, starting from Eq. (4.28). The load forecasting system has the input set and state space model refreshed at every 60 to 120 time steps, depending on the number of available weather stations.

5 RESULTS

In order to validate the proposed load forecasting systems performance, the load time series have been forecast by concurrent methods of linear and nonlinear natures. Several state-of-art methods were tested. Results have been divided between electric load forecast and photovoltaic generation forecast, presented in Sections 5.1 and 5.2, respectively.

5.1 Electric load forecasting

In order to validate the proposed load forecasting systems performance, the load time series have been forecast by concurrent methods of linear and nonlinear natures. Several state-of-art methods were tested, such as:

1. Kalman Filter with PCA (PKF),
2. Classical Kalman Filter (KF) without PCA,
3. Classical multilayer perceptron Artificial Neural Network trained by Backpropagation (BP),
4. MLP ANN trained by BP with PCA (PBP),

The above described benchmark models are used to forecast peak, average and base load. Peak forecasting is directly related to the maximum power that will be demanded for the system in a given day, which is important to plan the operation at its power limits, spinning reserves and ancillary systems. Average load is more related to the energy demand in the day, directly related to the electric energy supplied through contracts or hydraulic/fuel reserves. Base load is necessary to plan the operation at light loads, optimizing the shutdown of generation units and grid equipments with high operation and maintenance cost.

For each prediction the Mean Squared Error (MSE), Mean Average Percentual Error (MAPE), Maximum Percentual Error (MPE) and Correlation Coefficient (r^2) error metrics are calculated, using equations (5.1), (5.2), (5.3) and (5.4) respectively.

$$MSE = \frac{1}{n} \sum_{k=1}^{n} (y[k] - \widehat{y}[k])^2 \tag{5.1}$$

$$MAPE = \frac{100}{n} \sum_{k=1}^{n} \left| \frac{y[k] - \widehat{y}[k]}{y[k]} \right| \tag{5.2}$$

$$MPE = \frac{100}{n} \max_{k} \left| \frac{y[k] - \widehat{y}[k]}{y[k]} \right| \tag{5.3}$$

$$r^2 = \frac{\operatorname{cov}(y, \widehat{y})}{\sigma_y \sigma_{\widehat{y}}} = \frac{\sum_{k=1}^{n} (\widehat{y}[k] - \overline{y})^2}{\sum_{k=1}^{n} (y[k] - \overline{y})^2} \tag{5.4}$$

where $y[k]$ and $\widehat{y}[k]$ respectively denote the measured and forecasted electric load for day k, $\overline{y}$ is the time series mean of the loads, σ_y and $\sigma_{\widehat{y}}$ the standard deviation from mean in the measurements and predictions.

Three forecasting scenarios are used, based on real power systems. The first scenario comprises 8 power substations in Leipzig, from years 2001 to 2003, and its results are presented in Subsection 5.1.1. The second scenario features Brasilia, also from years 2001 to 2003 during an electricity crysis period. These results are shown in Subsection 5.1.2. The third showcases the electric load demanded by Brasilia, from years 2004 to 2010, a period of huge populational and economic growth. The forecasting results for the third scenario are listed in Subsection 5.1.3.

5.1.1 First forecasting scenario - Leipzig 2001-2003

Leipzig is the largest city in the german state of Saxony, with a population of more than 570.000 inhabitants. In 1930 the population reached its historical peak of over 700,000. It decreased steadily from 1950 until 1989 to about 530,000. In the 1990s the population decreased rather rapidly to 437,000 in 1998. This reduction was mostly due to outward migration and suburbanization. After almost doubling the city area by incorporation of surrounding towns in 1999, the number stabilized and started to rise again with an increase of 1,000 in 2000, as shown in Figure 5.1.

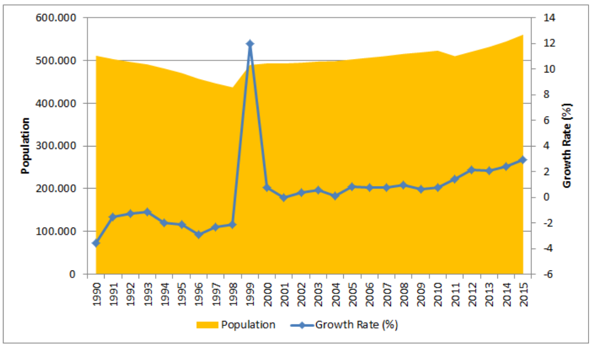


Figure 5.1: Leipzig population and population growth rate from 1990 to 2015. The 1999 growth peak is due to the incorporation of surrounding towns. Credits: EUROSTATs

The city has a temperate climate. Winters are variably mild to cold, with an average around 1 Celsius. Summers are generally warm, albeit not hot, averaging 19 Celsius with daytime maxima of 24 Celsius. Precipitation is higher in the summer, but there is no dry season in the winter. The amount of sunshine differs quite between winter and summer, with an average of 51 hours of sunshine in December and 229 hours of sunshine in July.

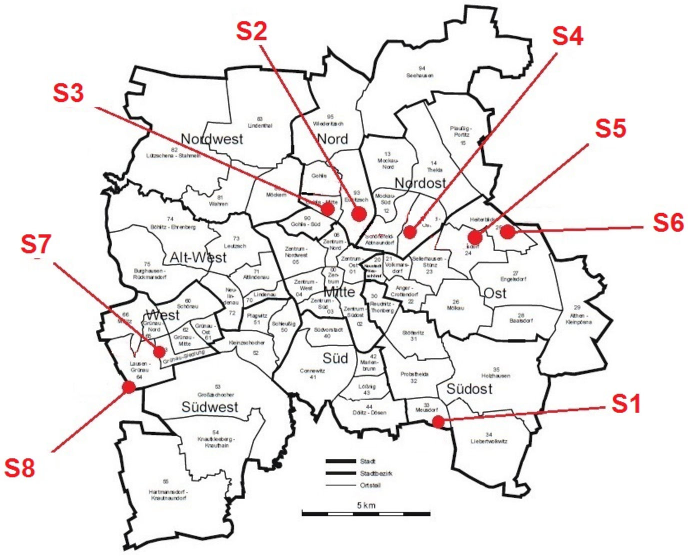


Figure 5.2: Locations of the eight substations in Leipzig. Credits: Jayme Milanezi Jr. [73]

The proposed and benchmark prediction methods are employed to forecast daily electric demand in power substations of Leipzig's distribution system, without any information about the topology or electrical parameters of the grid. The measurement variable is comprised of historical demand data, collected from eight substations located in different neighborhoods, as shown in Fig. 5.2. It contains daily values of minimum, mean and maximum demand from year 2001 to 2004, as illustrated in Figure 5.3.

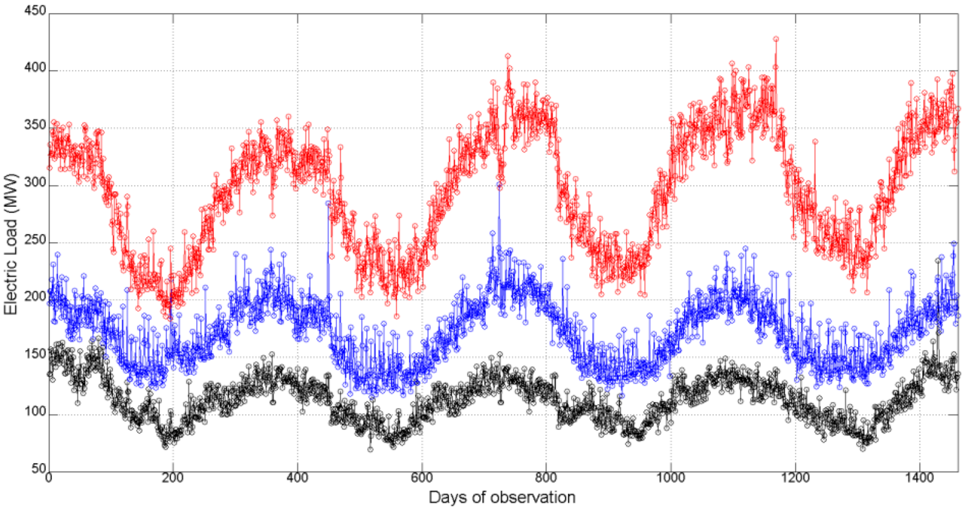


Figure 5.3: Evolution of electric load in Substation S1, from 2001 to 2004. Base load is plotted in black, Average load in blue and Peak load in red.

The corresponding historical weather data has been collected from the Leipzig-Halle (LEJ) weather station. Due to a gap in the METAR time series which occurred in January 2004, load predictions for this year have not been attempted in this work. As such, the training period ranges from January 2001 to December 2001, while the prediction period comprises 730 days between January 2002 and December 2003. Error metrics are calculated exclusively for the prediction period.

Two Kalman based predicting schemes are used, the proposed PCA-Kalman and the classical State space Kalman filter approach. For these methods, in the initialization procedure simulations were made with model orders ranging from one and twenty one, employing the year 2001 data. Considering the squared error metric, the best results are found when using a model order with seven state variables, as shown in Fig. 5.4. Model orders lower than seven fail to predict the weekly variations, while higher model orders are more computationally cumbersome, prone to numerical instabilities and numerical oscillations that seems to degrade forecasting performance.

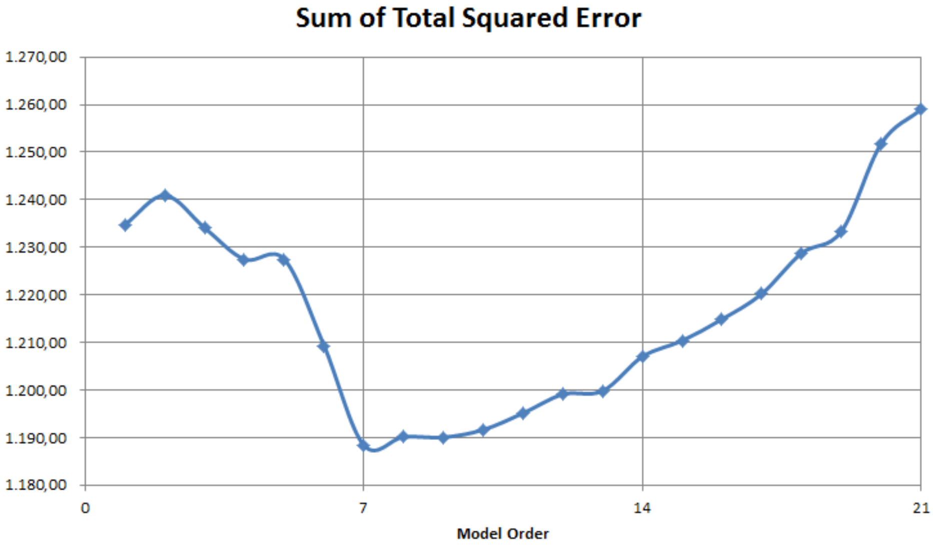


Figure 5.4: Sum of the Total Squared Error for the 8 substations as a function of Model Order. The minimum is achieved when the Order is set to 7.

The T parameter is set to 365, which determines that Phase Zero will be executed once at every year of prediction. The period of 365 days was chosen in order to model the yearly cycles shown in both electricity demand and temperature time series, as well as allowing the required number of data points to optimize the filter parameters with hundreds of inputs. Larger periods could not be reliably evaluated, given that there were only 3 years worth of data, and shorter periods are more prone to numerical oscillations due to data insufficiency. For the least squares optimization, the error tolerance was set to 10^{-11}, and the GMRES iterations are used to make the least squares fitting of the filter coefficients to the previous 365 days of electrical demand and input data.

The same model order adopted for the Kalman filter methods is used for the modified autoregressive Grey Box Model. This simple method is used to demonstrate the capabilities of a time series approach without exogenous inputs, and presents a baseline performance for both Kalman and neural network methodologies.

The two artificial neural network approaches employ MLP with Backpropagation (BP) supervised learning. The weight parameters are calculated by the Levenberg-Marquardt algorithm. The multilayer perceptron architecture employs a single hidden layer, containing 10 neurons. The PCA enhanced ANN employs the same feature selection used by the PCA-Kalman approach, while the standard BP ANN employs the raw inputs also used by the classic Kalman filter method. Results in this scenario are presented per substation, in the following subsubsections. The results are presented in tables showing the forecasting methods' performance over the 9 different input sets described in Table 4.1, as measured by the selected error metrics. The performance analisys covers the best method and input set for each combination, as well as the best method when employing the complete input set Z. Input set F is marked as not available (N/A) because the tariff history of Leipzig has not been obtained. The description of each substation's neighborhood is obtained from [66].

5.1.1.1 Substation S1

Substation S1 lies in the Meusdorf district, southeast of Leipzig. This neighborhood has a very low demographic density, and on average has between 1,9 and 2,0 inhabitants per house. Population growth in this area is estimated to be 9 % to 15 % between 1999 and 2003. In average, 70 % of these residents are economically active. Tables 5.1, 5.2 and 5.4 present the forecasting results for base, average and peak load, respectively.

Table 5.1: Error metrics for Base load, Substation S1

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	5,5	3,6	4,1	3,4	5,6	N/A	5,1	5,0	3,7
	KF	4,8	3,6	4,3	4,2	4,8	N/A	5,1	4,4	4,3
	PBP	7,3	14,7	15,8	13,9	11,0	N/A	11,8	13,0	18,5
	BP	9,9	12,8	18,1	18,8	5,5	N/A	9,8	13,9	16,0
MAPE	PKF	5,01	3,99	4,27	3,81	5,09	N/A	4,74	4,66	3,66
	KF	4,69	3,95	4,39	4,30	4,61	N/A	4,63	4,20	4,08
	PBP	5,70	8,17	8,50	8,10	7,14	N/A	7,36	7,73	9,92
	BP	6,81	7,63	9,18	9,43	4,97	N/A	7,11	7,77	8,60
MPE	PKF	23,2	25,2	28,9	22,9	23,0	N/A	37,7	33,5	33,2
	KF	17,7	32,2	20,7	28,5	20,0	N/A	35,1	31,3	35,5
	PBP	27,2	44,2	59,9	39,2	37,0	N/A	34,6	53,9	45,9
	BP	26,1	38,3	55,0	56,9	22,7	N/A	29,8	49,9	51,1
r^2	PKF	0,872	0,917	0,906	0,922	0,870	N/A	0,881	0,885	0,918
	KF	0,888	0,917	0,901	0,902	0,888	N/A	0,884	0,900	0,903
	PBP	0,836	0,749	0,693	0,740	0,746	N/A	0,717	0,733	0,597
	BP	0,812	0,738	0,668	0,524	0,879	N/A	0,773	0,686	0,675

For base load forecasting, the proposed PCA-Kalman has the best performance with input set D, followed by classic Kalman filter and standard MLP trained by Backpropagation. Overall, set E provides the better performance for the ANN methods. Restricting the input set to Z, the PCA-Kalman slightly outperforms the classic Kalman filter, followed by the PCA BP and the standard Backpropagation ANN. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.5:

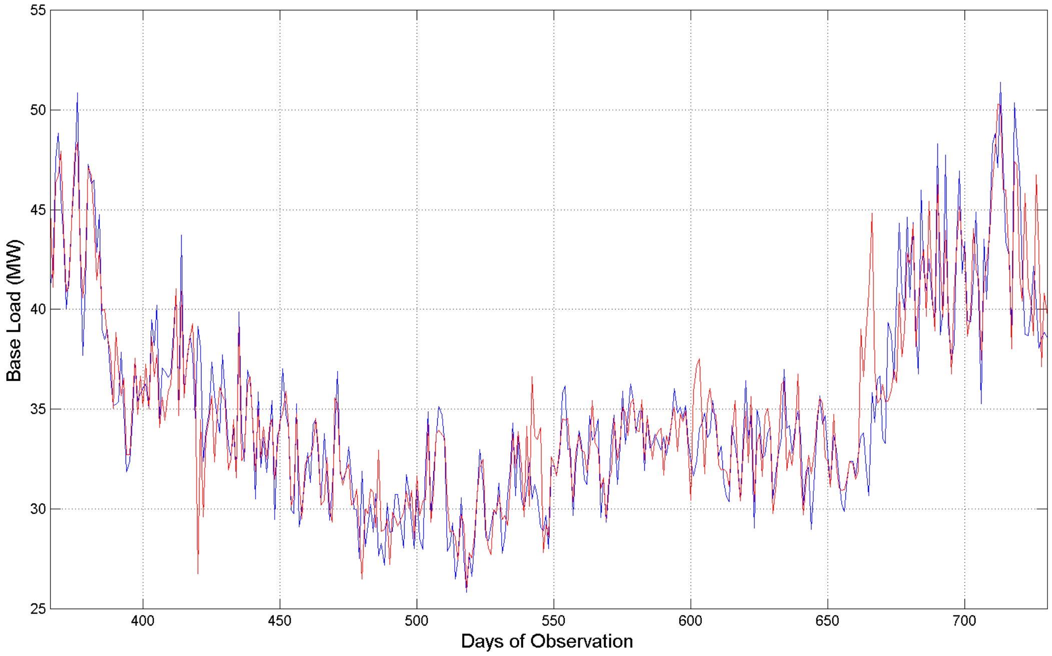


Figure 5.5: Prediction (red line) plotted against the measured base load in Substation S1 (blue line) over 360 days of observation.

Table 5.2: Error metrics for Average load, Substation S1

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	21,8	13,4	17,7	12,2	22,2	N/A	14,9	19,2	8,3
	KF	14,6	10,1	13,9	12,0	15,3	N/A	21,3	11,8	11,5
	PBP	54,3	51,2	76,5	73,2	38,2	N/A	41,7	64,6	69,6
	BP	27,1	81,6	43,0	54,3	27,6	N/A	44,6	92,3	60,1
MAPE	PKF	5,25	3,99	4,68	3,87	5,25	N/A	4,23	4,85	3,04
	KF	4,20	3,42	4,08	3,77	4,28	N/A	4,41	3,84	3,80
	PBP	8,13	7,96	9,47	9,57	6,88	N/A	7,26	8,84	9,11
	BP	5,46	10,31	7,12	8,41	5,60	N/A	7,49	10,81	8,54

Table 5.3: Error metrics for Average load, Substation S1 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	35,1	27,0	27,0	27,1	39,7	N/A	26,8	29,1	19,6
	KF	20,9	19,8	21,6	19,3	20,0	N/A	47,7	18,8	19,7
	PBP	45,7	54,4	50,9	58,4	34,6	N/A	33,7	49,1	46,0
	BP	33,1	65,7	34,2	44,0	31,4	N/A	38,7	61,2	54,8
r^2	PKF	0,882	0,929	0,905	0,935	0,879	N/A	0,921	0,898	0,957
	KF	0,922	0,947	0,926	0,937	0,919	N/A	0,890	0,938	0,939
	PBP	0,712	0,769	0,628	0,680	0,809	N/A	0,783	0,720	0,696
	BP	0,861	0,516	0,808	0,752	0,856	N/A	0,785	0,683	0,690

For average load, the best performance is obtained by the proposed PCA-Kalman filter using the input set Z, followed by the classic Kalman filter using input set B. Using Z inputs, the classic Kalman filter outperforms the PCA Backprogation method and the classic BP, which seems to perform poorly with too many inputs. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.6:

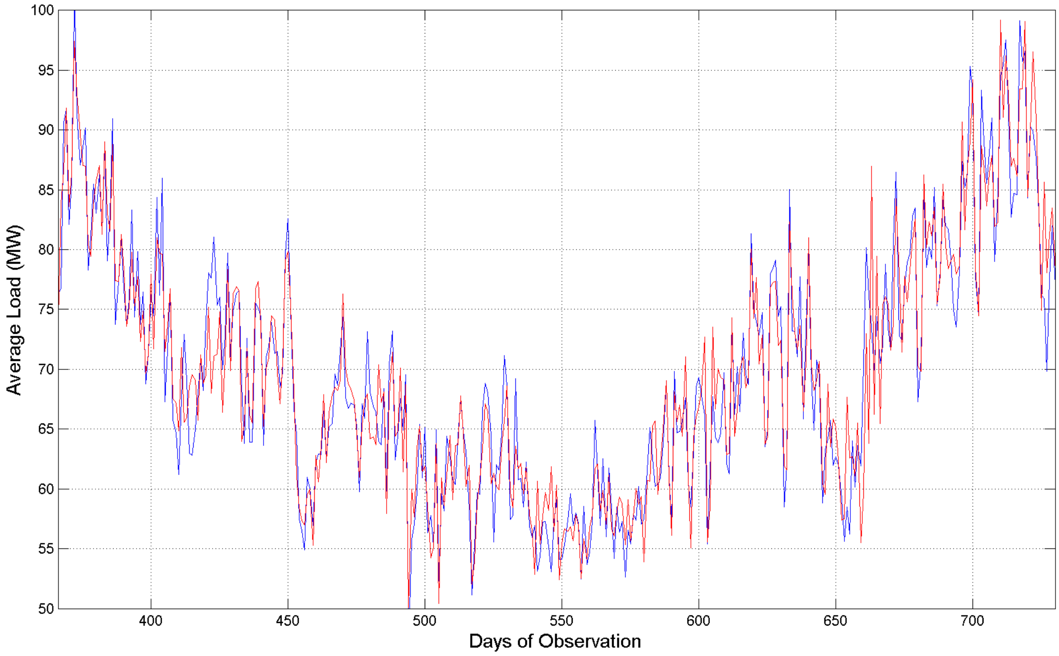


Figure 5.6: Prediction (red line) plotted against the measured average load in Substation S1 (blue line) over 360 days of observation.

Table 5.4: Error metrics for Peak load, Substation S1

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	70,0	47,3	54,6	42,6	67,0	N/A	58,2	63,1	40,2
	KF	58,0	44,2	56,0	51,6	57,1	N/A	81,3	47,2	47,5
	PBP	120,1	173,0	236,0	163,1	121,6	N/A	174,0	196,5	97,5
	BP	88,8	178,0	184,4	202,1	47,3	N/A	128,5	157,7	98,2
MAPE	PKF	6,15	4,80	5,17	4,51	5,97	N/A	5,47	5,59	4,11
	KF	5,55	4,55	5,28	5,08	5,48	N/A	5,66	4,77	4,71
	PBP	8,14	9,68	10,99	9,59	8,21	N/A	10,15	10,54	7,30
	BP	7,02	10,15	9,63	10,46	4,69	N/A	8,25	9,29	7,26
MPE	PKF	28,1	39,9	32,3	34,1	27,5	N/A	28,7	28,6	56,3
	KF	27,6	40,2	31,1	30,5	21,0	N/A	62,2	24,5	26,3
	PBP	38,0	40,2	41,9	51,2	43,2	N/A	55,8	45,1	37,6
	BP	34,7	45,3	48,1	50,1	23,8	N/A	47,0	41,3	37,2
r^2	PKF	0,877	0,919	0,906	0,927	0,882	N/A	0,898	0,892	0,933
	KF	0,899	0,924	0,903	0,911	0,900	N/A	0,862	0,919	0,918
	PBP	0,797	0,746	0,619	0,739	0,801	N/A	0,674	0,677	0,824
	BP	0,848	0,671	0,728	0,709	0,928	N/A	0,776	0,760	0,830

Forecasting peak loads, the better method is the PCA-Kalman filter with input set Z, followed by the classic Kalman filter. In this case, the classic BP ANN outperforms the PCA enhanced ANN when equipped with input set Z. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.7:

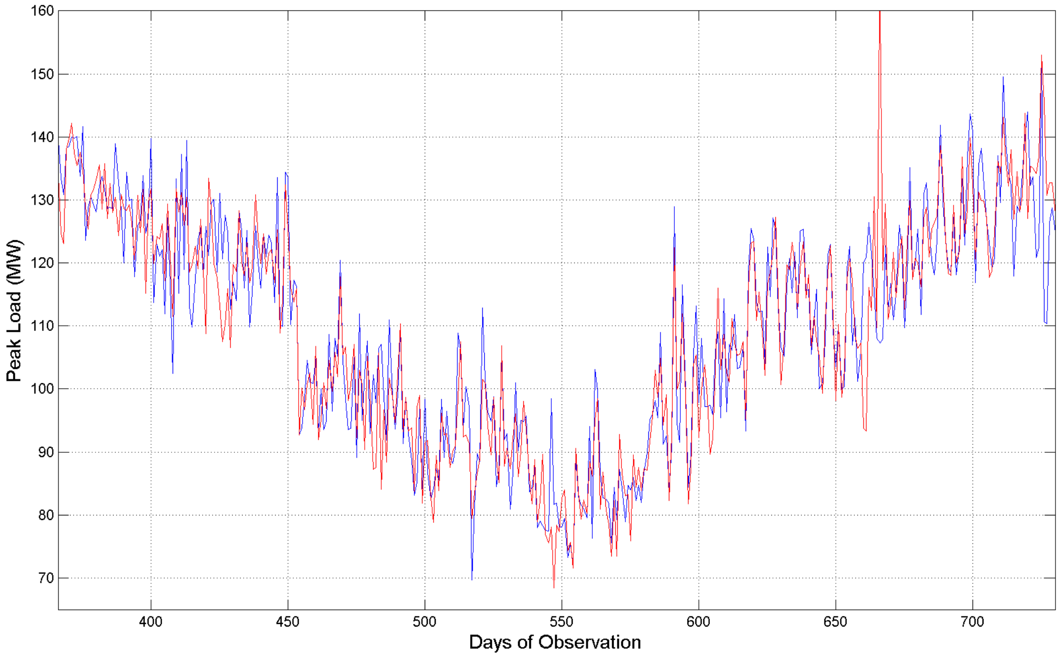


Figure 5.7: Prediction (red line) plotted against the measured peak load in Substation S1 (blue line) over 360 days of observation.

5.1.1.2 Substation S2

Substation S2 is located in the Gohlis-Mitte district, center of Leipzig. This neighborhood has a high demographic density, and on average has between 2,2 or more inhabitants per house. Population growth in this area is estimated to be 9 % to 15 % between 1999 and 2003. In average, 70 % of these residents are economically active. Tables 5.5, 5.6 and 5.7 present the forecasting results for base, average and peak load, respectively.

Table 5.5: Error metrics for Base load, Substation S2

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	7,1	5,5	5,1	4,2	7,5	N/A	6,2	5,6	4,3
	KF	6,1	5,0	5,5	5,2	6,3	N/A	23,9	4,9	4,7
	PBP	16,8	17,2	27,6	15,3	16,5	N/A	13,2	19,3	17,8
	BP	4,0	27,1	28,9	14,9	7,9	N/A	19,6	18,3	24,9
MAPE	PKF	3,59	2,97	2,81	2,60	3,66	N/A	3,27	3,14	2,42
	KF	3,30	2,81	3,18	3,05	3,31	N/A	3,86	2,84	2,73
	PBP	5,62	5,47	7,15	5,46	5,50	N/A	4,91	5,81	5,67
	BP	2,73	6,94	6,81	5,22	4,16	N/A	5,99	5,65	6,71
MPE	PKF	15,8	24,8	34,5	25,3	16,5	N/A	14,9	16,6	24,6
	KF	14,9	19,3	14,3	18,3	21,8	N/A	86,8	19,3	20,7
	PBP	26,7	34,6	28,5	24,7	21,4	N/A	23,5	37,9	35,9
	BP	14,8	30,6	36,4	27,8	20,4	N/A	26,3	31,6	39,0
r^2	PKF	0,921	0,939	0,943	0,953	0,913	N/A	0,929	0,937	0,952
	KF	0,931	0,944	0,938	0,941	0,928	N/A	0,793	0,945	0,947
	PBP	0,808	0,806	0,684	0,836	0,806	N/A	0,848	0,818	0,814
	BP	0,964	0,635	0,693	0,844	0,954	N/A	0,783	0,790	0,735

Forecasting base load in this substation, the lowest MSE overall is obtained by the standard Backpropagation ANN, using input set A, closely followed by the proposed PCA-Kalman method input set D. Comparing methods with input set Z, PCA-Kalman is the best option, offering a very slight performance penalty over the classic BP method at its best input set. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.8:

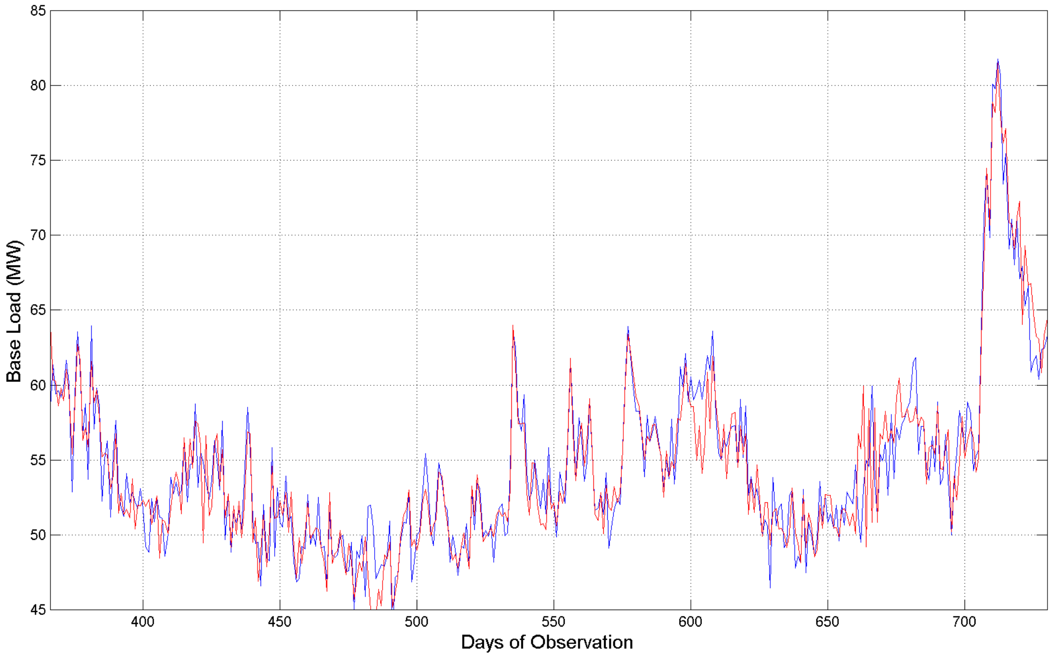


Figure 5.8: Prediction (red line) plotted against the measured base load in Substation S2 (blue line) over 360 days of observation.

Table 5.6: Error metrics for Average load, Substation S2

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	25,0	17,5	22,2	18,2	27,0	N/A	17,6	21,5	10,7
	KF	17,3	11,7	16,8	14,4	18,5	N/A	38,4	14,3	13,6
	PBP	56,7	76,4	97,6	80,9	44,3	N/A	64,9	70,3	110,2
	BP	50,3	84,8	74,7	85,5	39,6	N/A	69,1	61,3	63,9
MAPE	PKF	3,50	2,90	3,01	2,78	3,72	N/A	2,94	3,30	2,13
	KF	2,95	2,40	2,92	2,68	3,09	N/A	3,24	2,62	2,53
	PBP	5,41	6,11	7,23	6,47	4,87	N/A	5,97	6,13	7,43
	BP	4,94	6,36	6,30	6,86	3,91	N/A	6,05	5,94	5,96
MPE	PKF	17,0	24,9	53,1	38,7	18,3	N/A	16,9	19,0	17,2
	KF	16,7	13,8	21,2	15,0	18,9	N/A	52,2	16,4	15,6
	PBP	25,5	47,3	33,6	39,4	19,7	N/A	22,8	32,0	40,5
	BP	26,2	41,6	37,9	31,7	28,2	N/A	23,3	27,5	35,6
r^2	PKF	0,924	0,947	0,932	0,945	0,916	N/A	0,946	0,935	0,968
	KF	0,947	0,965	0,949	0,956	0,943	N/A	0,889	0,956	0,959
	PBP	0,831	0,763	0,738	0,782	0,859	N/A	0,790	0,808	0,717
	BP	0,852	0,745	0,807	0,758	0,897	N/A	0,767	0,813	0,811

For average load, the best performance is obtained by the PCA-Kalman filter using the input set Z, followed by the classic Kalman using input set B. Using Z inputs, the classic Kalman filter outperforms the classic Backprogation method and the PCA enhanced BP. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.9:

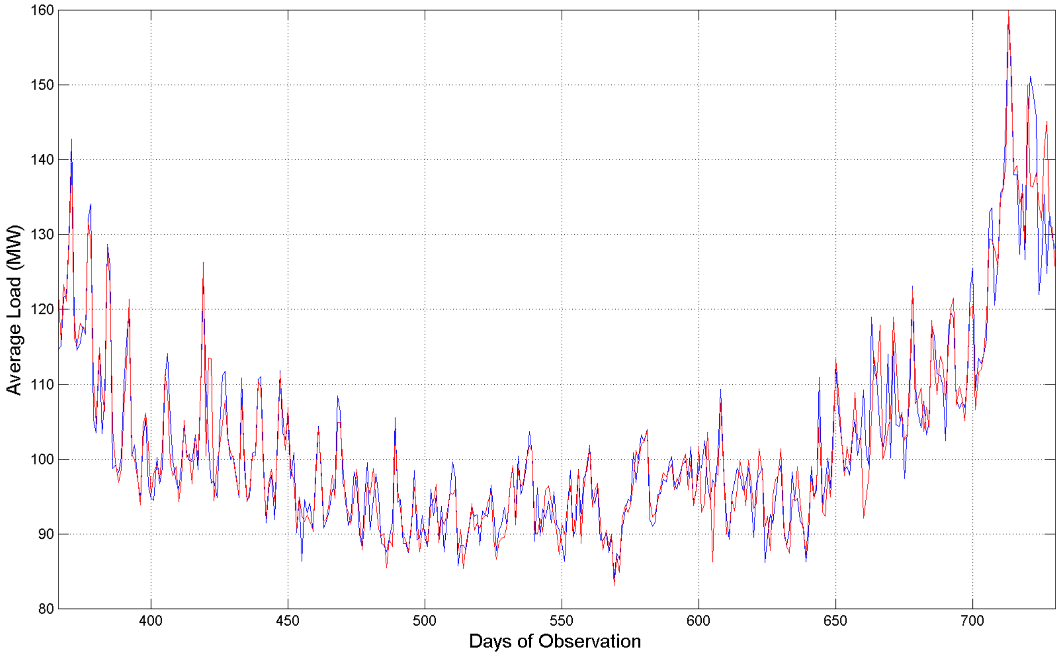


Figure 5.9: Prediction (red line) plotted against the measured average load in Substation S2 (blue line) over 360 days of observation.

Table 5.7: Error metrics for Peak load, Substation S2

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	193,3	160,8	158,6	153,5	187,8	N/A	136,7	166,2	119,3
	KF	142,3	117,5	148,2	138,5	139,3	N/A	468,8	117,1	114,1
	PBP	401,1	549,7	740,4	570,1	571,3	N/A	565,3	492,4	302,0
	BP	290,9	607,9	689,3	489,4	164,0	N/A	440,1	446,2	366,5
MAPE	PKF	6,07	5,09	5,29	4,87	5,92	N/A	5,05	5,44	3,98
	KF	5,15	4,31	5,01	4,80	5,13	N/A	5,67	4,49	4,34
	PBP	8,92	10,69	12,02	10,71	10,66	N/A	11,00	10,37	7,88
	BP	8,05	11,53	11,89	9,96	5,75	N/A	9,52	10,38	8,80

Table 5.8: Error metrics for Peak load, Substation S2 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	33,9	37,8	40,3	38,4	37,1	N/A	29,4	46,7	34,0
	KF	30,8	35,6	36,6	39,5	32,3	N/A	100,0	29,3	29,9
	PBP	39,2	54,4	81,5	43,9	45,6	N/A	69,3	40,2	31,6
	BP	41,0	49,6	77,4	58,2	27,6	N/A	43,5	49,0	42,8
r^2	PKF	0,890	0,910	0,911	0,914	0,893	N/A	0,923	0,907	0,935
	KF	0,920	0,934	0,917	0,922	0,921	N/A	0,771	0,935	0,937
	PBP	0,767	0,701	0,671	0,697	0,713	N/A	0,671	0,716	0,820
	BP	0,839	0,671	0,681	0,740	0,907	N/A	0,739	0,730	0,783

Forecasting peak load, the better method is the classic Kalman filter using input set Z, very closely followed by the PCA-Kalman method with the same input set. BP ANN method performs almost as good as the Kalman filters when using input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.10:

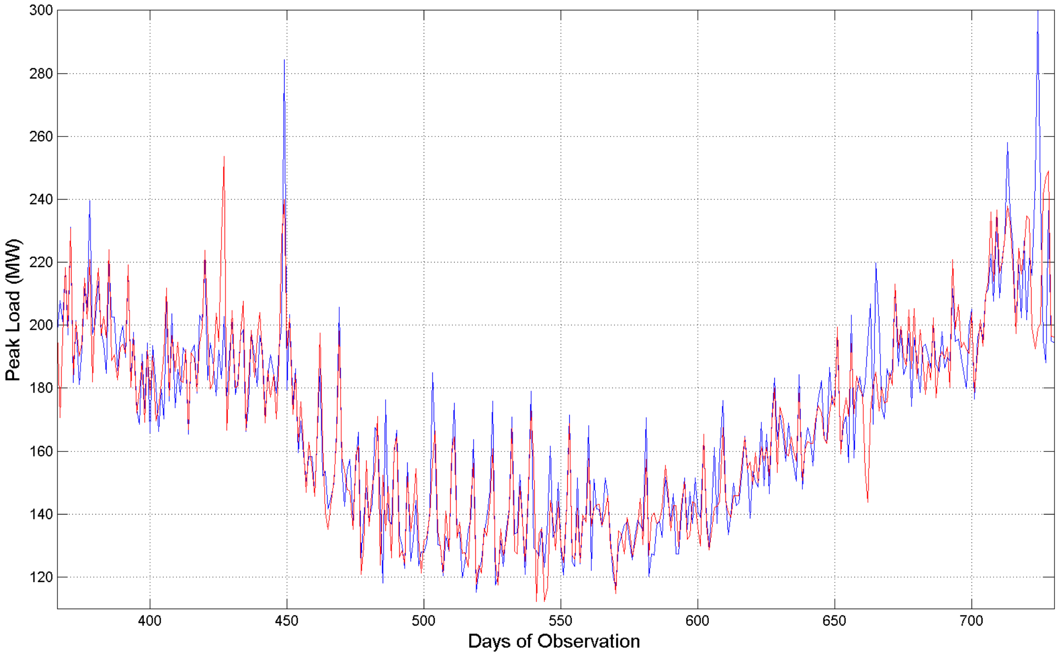


Figure 5.10: Prediction (red line) plotted against the measured peak load in Substation S2 (blue line) over 360 days of observation.

5.1.1.3 Substation S3

Substation S3 is located in the Gohlis-Nord district, northern center of Leipzig. This neighborhood has a very high demographic density, and on average has between 2,2 or more inhabitants per house. Population growth in this area is estimated to be 9 % to 15 % between 1999 and 2003. In average, 80 % of these residents are economically active. Tables 5.9, 5.10 and 5.11 present the forecasting results for base, average and peak load, respectively.

Table 5.9: Error metrics for Base load, Substation S3

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	13,6	10,2	9,7	9,5	14,3	N/A	11,9	12,6	9,6
	KF	11,6	9,2	10,5	9,8	11,2	N/A	45,5	9,7	9,6
	PBP	23,8	39,8	31,3	42,0	22,1	N/A	20,5	33,2	28,1
	BP	20,8	31,3	30,4	36,5	8,5	N/A	19,2	17,5	33,9
MAPE	PKF	2,76	2,29	2,24	2,19	2,82	N/A	2,48	2,55	1,98
	KF	2,52	2,12	2,40	2,29	2,47	N/A	2,96	2,26	2,22
	PBP	3,76	4,68	4,18	4,81	3,66	N/A	3,32	4,30	4,03
	BP	3,49	4,23	4,25	4,34	2,15	N/A	3,29	3,19	4,42
MPE	PKF	13,6	17,7	15,3	14,2	15,5	N/A	14,7	16,6	19,5
	KF	14,8	17,6	13,3	13,3	13,2	N/A	69,7	12,9	14,2
	PBP	13,6	23,7	16,9	28,1	16,0	N/A	21,2	19,2	22,3
	BP	17,2	19,7	19,7	33,6	10,4	N/A	13,1	12,6	24,5
r^2	PKF	0,847	0,886	0,892	0,896	0,833	N/A	0,866	0,860	0,899
	KF	0,868	0,897	0,882	0,890	0,872	N/A	0,687	0,891	0,892
	PBP	0,699	0,640	0,654	0,602	0,733	N/A	0,777	0,696	0,712
	BP	0,799	0,713	0,687	0,642	0,905	N/A	0,775	0,797	0,692

For base load, classic BP with input set E is the method that provides the lowest MSE, followed by the classic Kalman filter with input set B. Using the Z input set, PKF and classic KF perform similarly. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.11:

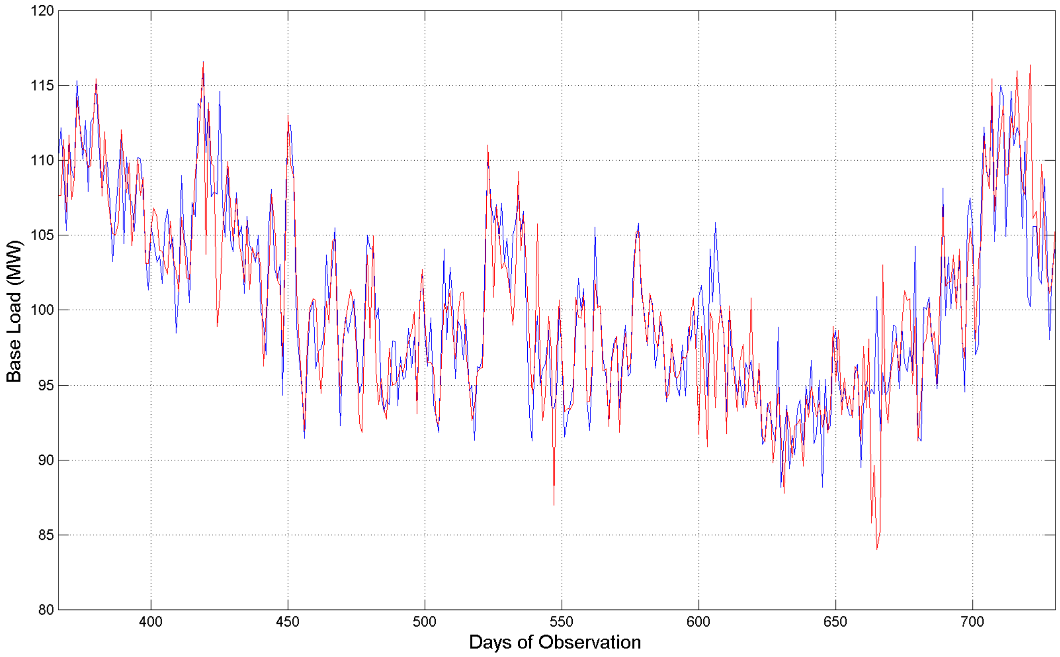


Figure 5.11: Prediction (red line) plotted against the measured base load in Substation S3 (blue line) over 360 days of observation.

Table 5.10: Error metrics for Average load, Substation S3

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	68,1	45,3	57,9	48,4	74,5	N/A	55,3	63,1	35,8
	KF	50,3	37,8	47,7	43,0	54,6	N/A	136,5	39,3	38,1
	PBP	112,7	199,6	183,3	169,0	187,2	N/A	142,8	225,5	174,5
	BP	59,6	190,4	191,8	234,3	61,5	N/A	137,2	335,5	219,4
MAPE	PKF	3,42	2,66	3,02	2,58	3,67	N/A	3,01	3,21	2,32
	KF	2,97	2,46	2,91	2,72	3,12	N/A	3,33	2,59	2,57
	PBP	4,57	6,07	5,74	5,70	5,76	N/A	4,97	6,56	5,68
	BP	3,27	5,71	5,96	6,49	3,27	N/A	4,86	7,18	6,13
MPE	PKF	16,6	25,0	34,2	31,2	23,5	N/A	18,4	25,8	17,8
	KF	15,5	15,0	20,8	16,6	18,7	N/A	59,2	12,4	14,9
	PBP	21,5	29,9	36,6	26,2	27,3	N/A	24,8	31,5	37,1
	BP	15,0	37,2	27,2	35,8	15,9	N/A	22,1	40,3	31,1
r^2	PKF	0,908	0,939	0,922	0,935	0,898	N/A	0,926	0,916	0,953
	KF	0,932	0,951	0,936	0,943	0,926	N/A	0,837	0,948	0,949
	PBP	0,848	0,786	0,769	0,813	0,740	N/A	0,805	0,747	0,777
	BP	0,924	0,781	0,778	0,739	0,921	N/A	0,810	0,563	0,726

For average load, PCA-Kalman with input set Z is the method that provides the lowest MSE, closely followed by the classic Kalman with input set B. Among the ANN methods, the BP approach has the best performance using input set E, but compares poorly with the Kalman filters. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.12:

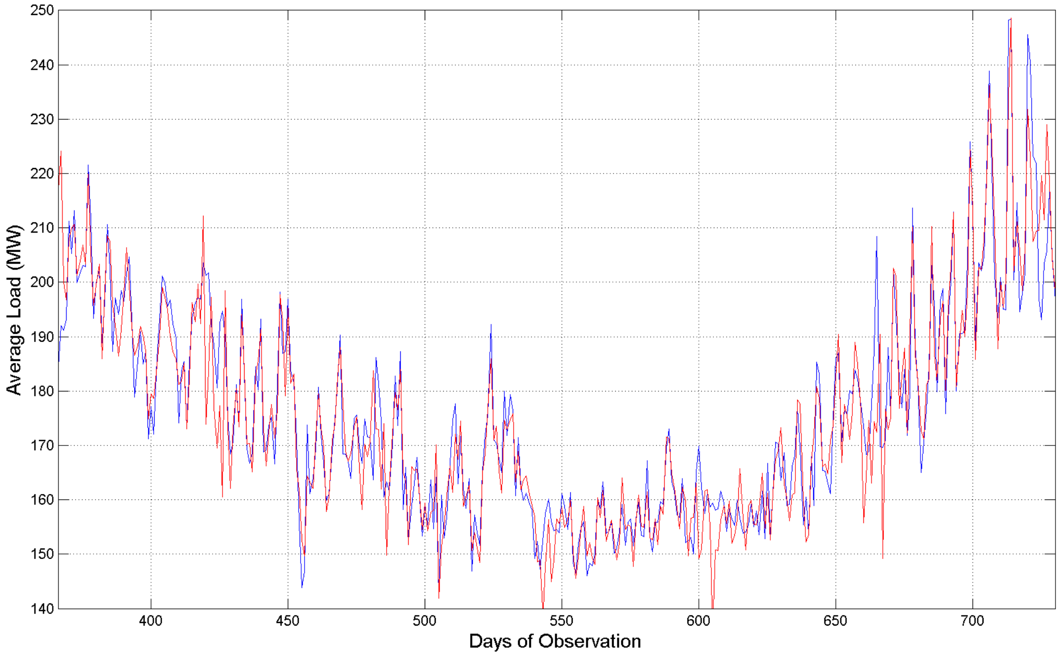


Figure 5.12: Prediction (red line) plotted against the measured average load in Substation S3 (blue line) over 360 days of observation.

Table 5.11: Error metrics for Peak load, Substation S3

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	267,7	186,2	191,8	177,3	254,5	N/A	217,7	238,1	135,3
	KF	211,8	163,4	197,0	184,8	208,5	N/A	565,6	172,6	167,3
	PBP	269,5	723,4	747,9	695,5	607,3	N/A	1140,7	650,0	428,1
	BP	496,3	670,0	699,7	612,7	268,8	N/A	477,4	487,0	750,8
MAPE	PKF	4,47	3,63	3,65	3,48	4,39	N/A	3,90	4,10	2,89
	KF	4,00	3,34	3,74	3,66	4,00	N/A	4,28	3,54	3,43
	PBP	4,49	7,34	7,66	7,12	6,76	N/A	9,52	6,79	5,75
	BP	6,21	7,12	7,07	6,72	4,36	N/A	6,07	6,15	7,47

Table 5.12: Error metrics for Peak load, Substation S3 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	23,3	26,6	25,1	22,1	20,5	N/A	22,6	28,8	24,6
	KF	22,8	23,1	21,5	23,2	19,5	N/A	68,8	22,4	23,9
	PBP	18,9	39,0	40,9	41,0	39,6	N/A	43,5	52,9	25,4
	BP	29,1	35,5	39,3	42,0	20,9	N/A	32,0	36,0	30,0
r^2	PKF	0,947	0,964	0,963	0,966	0,950	N/A	0,958	0,954	0,974
	KF	0,959	0,969	0,962	0,964	0,959	N/A	0,895	0,967	0,968
	PBP	0,949	0,863	0,857	0,869	0,881	N/A	0,751	0,874	0,919
	BP	0,901	0,885	0,875	0,887	0,947	N/A	0,906	0,902	0,877

Forescasting peak load, the PCA-Kalman method offers the better performance when combined with input set Z, followed by classic Kalman with input set B. The better ANN alternative is the classic BP ANN, using input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.13:

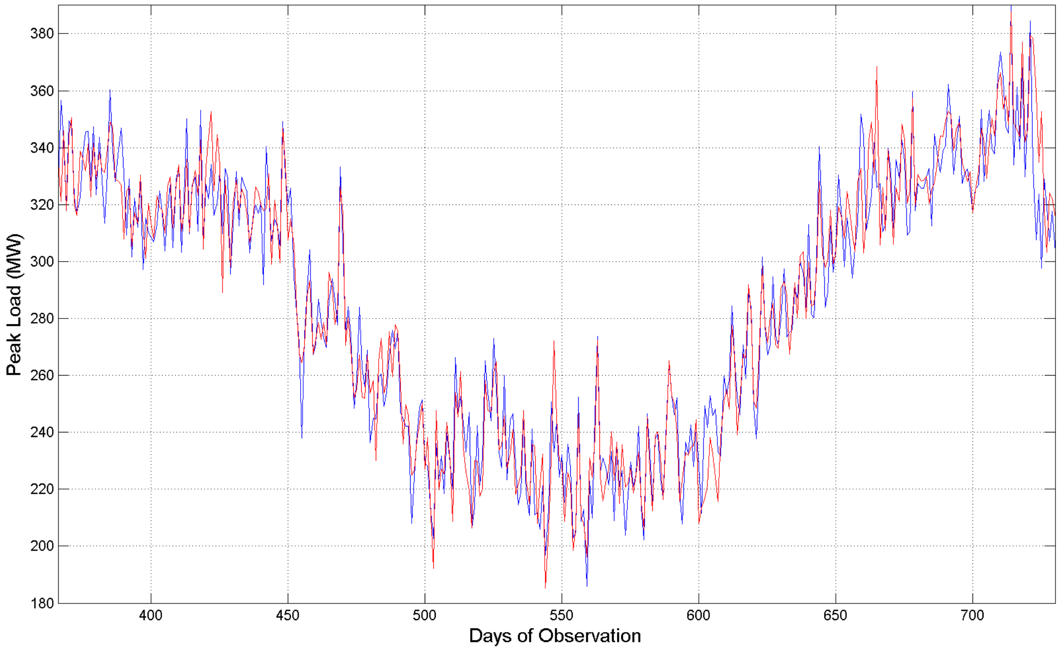


Figure 5.13: Prediction (red line) plotted against the measured peak load in Substation S3 (blue line) over 360 days of observation.

5.1.1.4 Substation S4

Substation S4 is located in the Schonefeld-ost district, center-northeast of Leipzig. This neighborhood has a medium demographic density, and on average has between 2,2 or more inhabitants per house. Population growth in this area is estimated to be 3 % to 9 % between 1999 and 2003. In average, 60% of these residents are economically active. Tables 5.13, 5.14 and 5.15 present the forecasting results for base, average and peak load, respectively.

Table 5.13: Error metrics for Base load, Substation S4

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	2,6	1,7	2,1	1,8	2,5	N/A	2,3	2,3	1,5
	KF	2,2	1,6	1,9	1,8	2,1	N/A	4,3	1,9	1,9
	PBP	3,9	6,7	5,6	7,5	3,4	N/A	5,9	5,4	6,4
	BP	5,3	6,2	6,8	5,4	3,5	N/A	6,0	5,2	7,7
MAPE	PKF	3,48	2,73	2,93	2,69	3,50	N/A	3,25	3,28	2,43
	KF	3,28	2,66	3,05	2,96	3,21	N/A	3,38	2,95	2,91
	PBP	4,34	5,67	5,18	6,16	4,05	N/A	5,58	5,29	5,73
	BP	5,07	5,55	5,84	5,15	4,00	N/A	5,56	5,00	6,27
MPE	PKF	32,9	18,9	28,3	22,5	31,7	N/A	28,6	26,0	23,4
	KF	26,9	17,2	23,5	21,6	30,2	N/A	53,0	24,6	25,5
	PBP	34,8	27,5	24,1	51,5	41,4	N/A	29,2	37,1	32,1
	BP	54,7	53,9	27,4	53,9	47,7	N/A	50,1	42,3	40,7
r^2	PKF	0,833	0,892	0,864	0,886	0,830	N/A	0,850	0,854	0,903
	KF	0,853	0,902	0,877	0,885	0,858	N/A	0,751	0,879	0,880
	PBP	0,726	0,680	0,697	0,603	0,764	N/A	0,580	0,684	0,594
	BP	0,605	0,618	0,617	0,690	0,755	N/A	0,562	0,634	0,514

The PCA-Kalman filter using input set Z very slightly outperforms the classic Kalman method. The better ANN method turns out to be the PCA-BP using input set E. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.14:

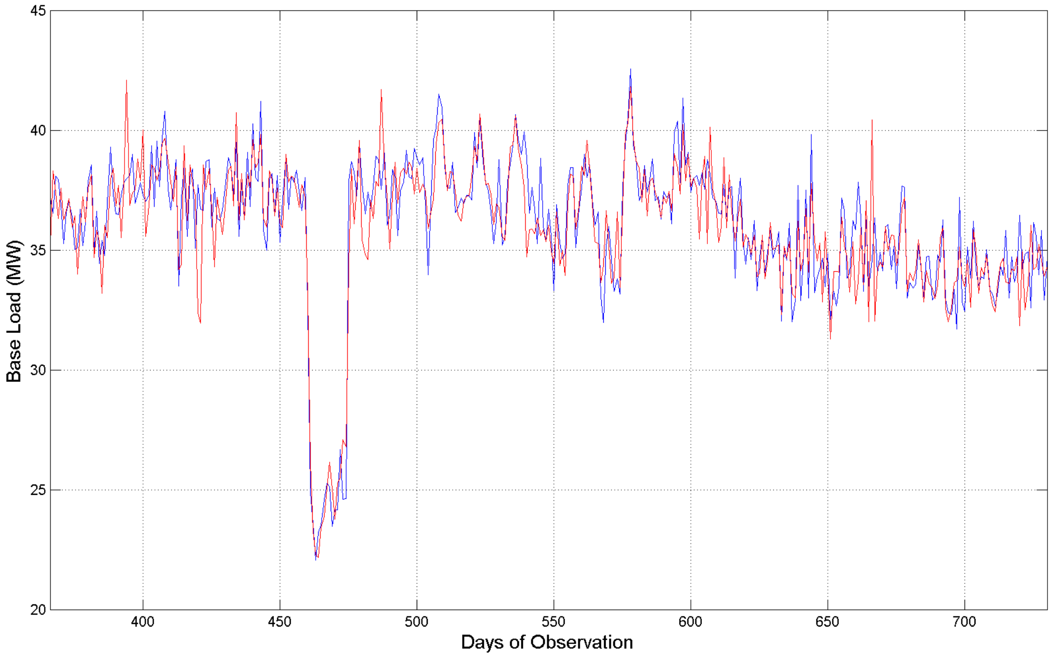


Figure 5.14: Prediction (red line) plotted against the measured base load in Substation S4 (blue line) over 360 days of observation.

Table 5.14: Error metrics for Average load, Substation S4

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	90,8	88,1	102,5	93,5	86,2	N/A	83,8	87,6	56,0
	KF	66,0	58,4	58,4	59,0	63,4	N/A	236,7	69,9	58,5
	PBP	176,8	189,5	278,6	230,6	159,8	N/A	186,2	230,5	151,2
	BP	84,8	305,0	278,2	388,5	103,2	N/A	157,0	299,0	131,7
MAPE	PKF	6,53	5,68	6,12	5,66	6,26	N/A	5,55	6,04	4,55
	KF	5,46	4,91	5,23	5,12	5,48	N/A	6,17	4,92	4,95
	PBP	9,25	9,69	11,88	10,69	9,63	N/A	9,99	10,69	8,76
	BP	7,02	12,72	11,78	13,93	6,99	N/A	9,25	12,74	8,17
MPE	PKF	46,8	59,8	48,4	53,3	47,2	N/A	44,3	40,9	35,1
	KF	38,1	34,7	33,4	30,6	39,2	N/A	100,0	30,3	31,4
	PBP	46,1	53,9	57,7	62,7	52,0	N/A	54,3	57,4	48,0
	BP	30,9	58,0	80,1	73,8	39,3	N/A	50,9	70,9	36,3
r^2	PKF	0,883	0,889	0,870	0,882	0,889	N/A	0,893	0,889	0,930
	KF	0,916	0,927	0,926	0,926	0,920	N/A	0,741	0,915	0,927
	PBP	0,759	0,775	0,606	0,726	0,789	N/A	0,753	0,739	0,798
	BP	0,909	0,656	0,705	0,629	0,867	N/A	0,788	0,714	0,826

Forecasting average loads, the PCA-Kalman method with input set Z slightly outperforms the classic-Kalman method with input set H. Classic BP performs better with input set A. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.15:

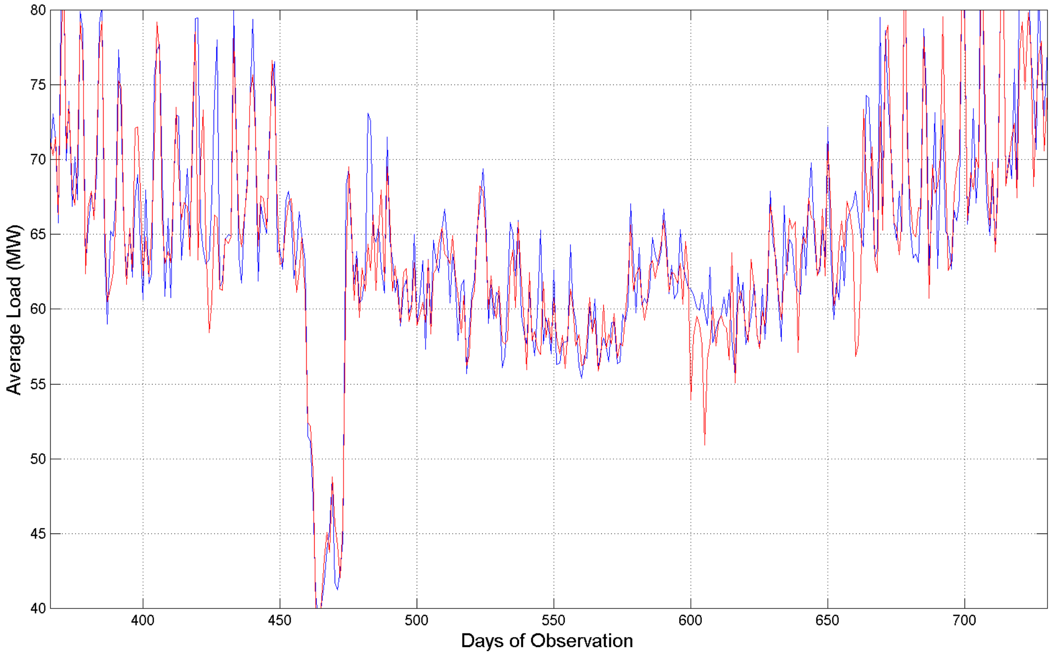


Figure 5.15: Prediction (red line) plotted against the measured average load in Substation S4 (blue line) over 360 days of observation.

Table 5.15: Error metrics for Peak load, Substation S4

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	14,5	10,4	12,5	11,6	15,7	N/A	11,2	14,5	7,1
	KF	10,2	7,8	10,0	8,8	11,2	N/A	16,4	8,9	8,8
	PBP	34,9	42,2	44,4	41,3	32,0	N/A	31,2	55,2	44,4
	BP	43,9	55,6	52,4	54,3	16,4	N/A	23,5	25,4	42,9
MAPE	PKF	4,39	3,50	3,82	3,54	4,51	N/A	3,91	4,30	2,84
	KF	3,77	3,11	3,66	3,44	3,87	N/A	4,06	3,38	3,33
	PBP	7,36	7,57	7,98	7,73	7,04	N/A	7,01	8,98	8,32
	BP	8,23	9,23	8,84	9,07	4,73	N/A	5,85	6,40	7,79

Table 5.16: Error metrics for Peak load, Substation S4 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	34,8	23,3	26,8	29,9	31,7	N/A	25,1	27,1	23,4
	KF	25,1	21,3	19,9	24,4	25,0	N/A	46,1	24,6	24,2
	PBP	43,1	35,5	52,8	55,3	40,6	N/A	45,4	43,3	58,5
	BP	80,8	56,3	62,8	44,1	21,7	N/A	38,9	54,7	47,6
r^2	PKF	0,844	0,891	0,868	0,879	0,830	N/A	0,881	0,845	0,927
	KF	0,892	0,919	0,895	0,908	0,881	N/A	0,828	0,907	0,909
	PBP	0,641	0,587	0,635	0,617	0,662	N/A	0,652	0,564	0,608
	BP	0,536	0,584	0,473	0,526	0,836	N/A	0,755	0,731	0,612

The PCA-Kalman filter with input set Z offers the better performance when forecasting peak load, followed by the classic Kalman using input set B. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.16:

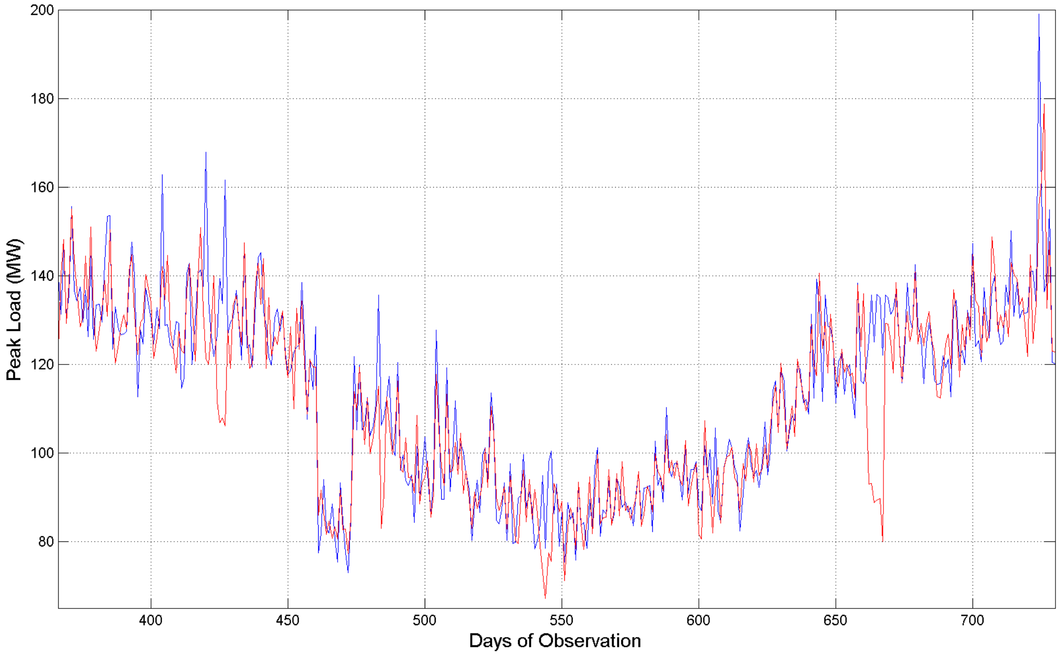


Figure 5.16: Prediction (red line) plotted against the measured peak load in Substation S4 (blue line) over 360 days of observation.

5.1.1.5 Substation S5

Substation S5 is located in the Paunsdorf district, east of Leipzig. This neighborhood has a medium demographic density, and on average has between 2,2 or more inhabitants per house. Population growth in this area is estimated to be -3 % to 3 % between 1999 and 2003. In average, 60 % of these residents are economically active. Tables 5.17, 5.18 and 5.19 present the forecasting results for base, average and peak load, respectively.

Table 5.17: Error metrics for Base load, Substation S5

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	1,6	1,3	1,3	1,4	1,6	N/A	2,6	1,6	1,3
	KF	1,3	1,0	1,3	1,1	1,3	N/A	12,6	1,1	1,0
	PBP	4,7	5,1	4,5	4,8	3,3	N/A	4,5	4,3	4,7
	BP	1,7	6,2	5,2	6,1	1,8	N/A	2,8	5,4	6,5
MAPE	PKF	3,16	2,70	2,69	2,63	3,18	N/A	3,17	2,95	2,39
	KF	2,84	2,42	2,72	2,55	2,84	N/A	3,86	2,48	2,39
	PBP	5,43	5,84	5,30	5,49	4,58	N/A	5,48	5,30	5,45
	BP	3,38	6,38	5,64	6,42	3,39	N/A	4,27	5,82	6,60
MPE	PKF	18,4	19,9	19,5	25,5	17,0	N/A	40,0	24,3	27,8
	KF	14,9	21,2	16,9	14,9	15,4	N/A	100,0	17,2	17,0
	PBP	28,2	32,7	26,3	28,8	21,4	N/A	23,6	25,0	33,7
	BP	15,8	42,6	37,4	35,4	20,8	N/A	23,6	33,7	34,9
r^2	PKF	0,930	0,943	0,942	0,939	0,926	N/A	0,887	0,930	0,945
	KF	0,941	0,954	0,944	0,951	0,941	N/A	0,642	0,953	0,956
	PBP	0,807	0,803	0,811	0,827	0,853	N/A	0,799	0,831	0,816
	BP	0,924	0,766	0,803	0,782	0,918	N/A	0,878	0,765	0,781

For base load forecasting, the classic Kalman filter with input sets B or Z outperforms the PCA-Kalman filter at the input B, C or Z. Classic BP performs better with input set A. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.17:

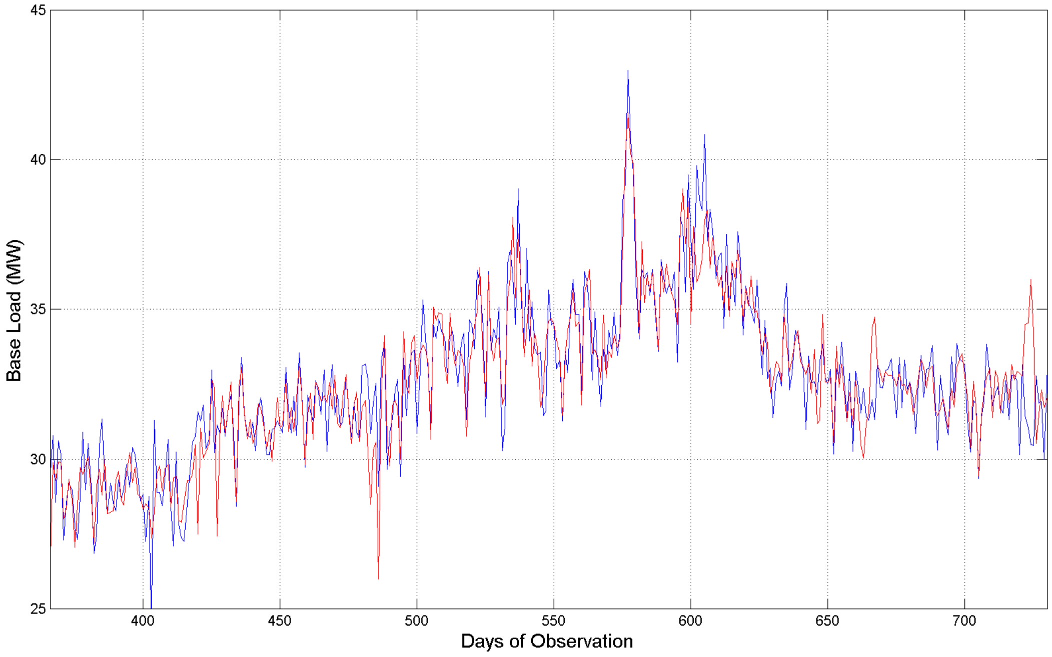


Figure 5.17: Prediction (red line) plotted against the measured base load in Substation S5 (blue line) over 360 days of observation.

Table 5.18: Error metrics for Average load, Substation S5

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	13,0	12,3	13,4	12,3	14,8	N/A	9,1	12,9	7,8
	KF	9,5	8,2	8,8	8,5	8,9	N/A	12,9	8,1	7,8
	PBP	32,5	49,4	53,8	42,8	37,5	N/A	35,9	38,1	40,4
	BP	18,0	65,0	38,6	54,0	25,1	N/A	32,3	53,8	46,2
MAPE	PKF	4,92	4,17	4,37	4,02	5,03	N/A	3,79	4,72	3,16
	KF	4,00	3,43	3,92	3,73	3,85	N/A	4,11	3,53	3,54
	PBP	7,66	9,87	11,15	9,41	8,98	N/A	8,86	9,12	9,18
	BP	5,64	11,52	8,97	10,48	7,02	N/A	8,56	10,81	9,60
MPE	PKF	31,5	34,8	54,4	43,8	36,5	N/A	30,4	34,7	34,5
	KF	31,4	31,4	26,7	27,7	27,2	N/A	45,6	31,2	31,1
	PBP	29,1	43,4	45,6	38,2	36,1	N/A	31,6	33,3	39,6
	BP	30,8	71,0	53,9	80,0	36,0	N/A	34,4	56,5	52,7
r^2	PKF	0,849	0,861	0,848	0,861	0,825	N/A	0,897	0,851	0,913
	KF	0,891	0,907	0,899	0,904	0,898	N/A	0,857	0,911	0,912
	PBP	0,676	0,548	0,535	0,615	0,587	N/A	0,609	0,578	0,605
	BP	0,792	0,273	0,625	0,564	0,701	N/A	0,641	0,524	0,534

For average load, the PCA-Kalman filter with input set Z is the method with lower MSE, followed closely by the classic Kalman filter with input set Z, also. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.18:

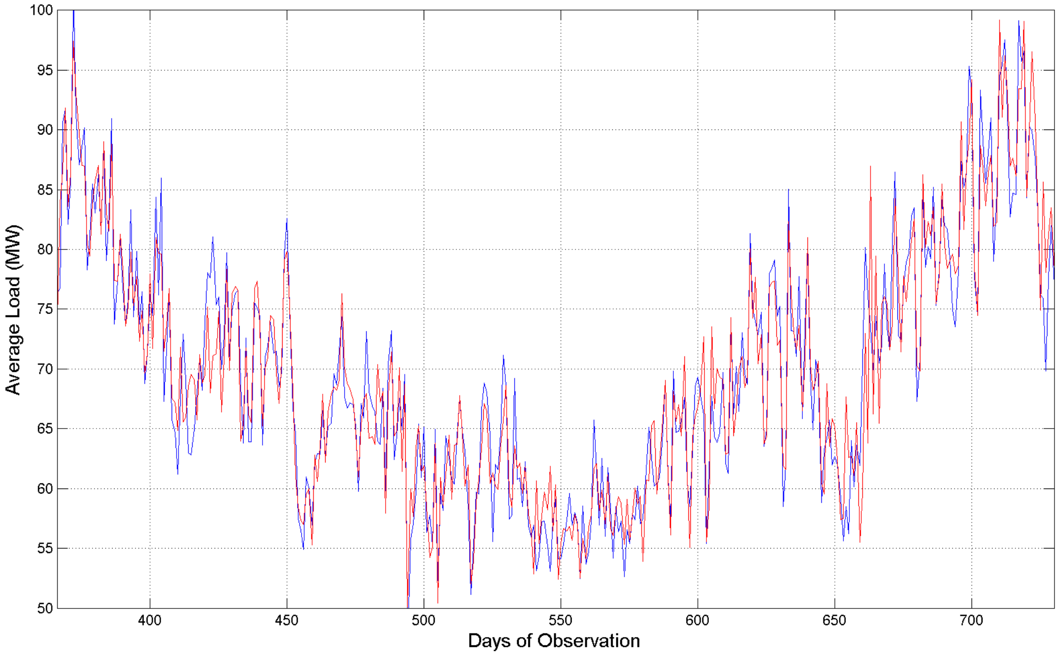


Figure 5.18: Prediction (red line) plotted against the measured average load in Substation S5 (blue line) over 360 days of observation.

Table 5.19: Error metrics for Peak load, Substation S5

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	135,4	109,3	117,3	101,6	129,4	N/A	81,6	129,7	71,5
	KF	76,0	72,8	78,8	82,0	76,3	N/A	192,5	66,0	62,6
	PBP	403,4	322,9	349,6	286,7	287,4	N/A	250,5	377,0	218,4
	BP	158,8	269,7	353,5	384,7	148,2	N/A	254,7	242,5	328,6
MAPE	PKF	8,96	7,33	7,66	6,97	8,80	N/A	6,77	8,22	5,46
	KF	6,85	6,24	6,70	6,68	6,83	N/A	7,37	5,99	5,86
	PBP	16,02	14,10	14,69	13,65	13,11	N/A	12,10	15,45	11,14
	BP	9,59	13,09	15,13	15,16	9,11	N/A	13,23	11,47	13,84

Table 5.20: Error metrics for Peak load, Substation S5 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	40,5	44,9	65,1	59,1	47,7	N/A	42,4	58,5	44,9
	KF	38,2	43,9	50,4	44,2	42,2	N/A	100,0	35,3	30,8
	PBP	105,4	84,0	78,2	52,4	47,5	N/A	57,4	73,6	41,6
	BP	58,4	62,1	54,1	81,5	42,6	N/A	61,1	45,1	57,6
r^2	PKF	0,800	0,844	0,832	0,856	0,807	N/A	0,883	0,814	0,902
	KF	0,891	0,897	0,888	0,884	0,891	N/A	0,757	0,907	0,912
	PBP	0,632	0,640	0,522	0,605	0,594	N/A	0,587	0,418	0,667
	BP	0,768	0,677	0,623	0,550	0,792	N/A	0,560	0,647	0,515

For peak load, the Kalman filter with input set Z is the better method, followed by the PCA-Kalman approach with the same input set. The better ANN method is the BP using input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.19:

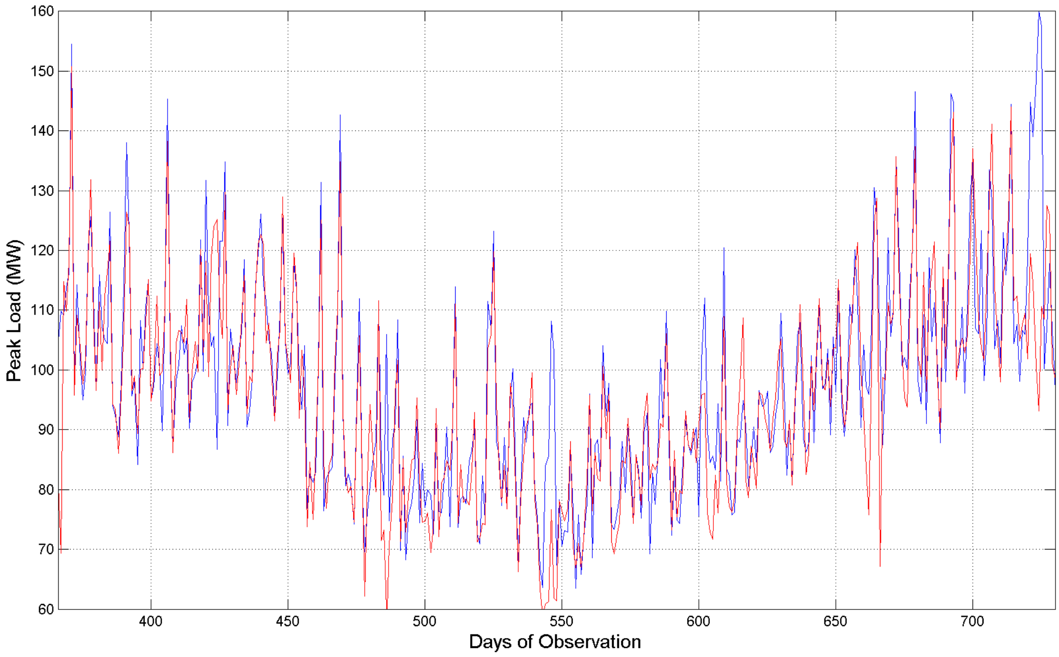


Figure 5.19: Prediction (red line) plotted against the measured peak load in Substation S5 (blue line) over 360 days of observation.

5.1.1.6 Substation S6

Substation S6 is located in the Heiterblick district, east of Leipzig. This neighborhood has a medium demographic density, and on average has 1,9 inhabitants per house. Population growth in this area is negative, estimated to below -3 % between 1999 and 2003. In average, 50 % of these residents are economically active. Tables 5.21, 5.22 and 5.23 present the forecasting results for base, average and peak load, respectively.

Table 5.21: Error metrics for Base load, Substation S6

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	7,1	4,8	5,1	4,2	7,4	N/A	6,1	7,4	3,2
	KF	5,3	3,7	4,9	4,4	5,5	N/A	36,3	4,3	4,4
	PBP	11,4	15,5	11,2	20,2	11,3	N/A	20,8	18,3	14,8
	BP	9,6	19,9	16,0	21,7	7,4	N/A	10,9	11,3	15,9
MAPE	PKF	3,49	2,88	2,90	2,67	3,57	N/A	3,15	3,34	2,28
	KF	3,01	2,53	2,93	2,77	3,10	N/A	3,96	2,67	2,65
	PBP	4,30	5,29	4,47	6,20	4,41	N/A	5,87	5,80	5,25
	BP	4,17	5,86	5,37	6,18	3,55	N/A	4,27	4,47	5,36
MPE	PKF	16,2	16,7	26,5	20,3	18,2	N/A	20,3	25,5	15,4
	KF	16,3	15,4	13,3	15,9	16,3	N/A	100,0	14,7	18,2
	PBP	21,4	24,2	22,9	28,7	16,5	N/A	29,4	39,0	25,1
	BP	20,0	31,9	29,8	22,1	25,9	N/A	18,0	26,3	25,3
r^2	PKF	0,874	0,914	0,909	0,925	0,865	N/A	0,890	0,870	0,944
	KF	0,904	0,935	0,912	0,922	0,900	N/A	0,635	0,924	0,923
	PBP	0,802	0,711	0,808	0,687	0,790	N/A	0,545	0,655	0,780
	BP	0,821	0,691	0,720	0,586	0,871	N/A	0,798	0,789	0,747

The proposed PCA-Kalman method with input set Z offers the lowest MSE when forecasting base load, followed by the Kalman filter with input set B. The better ANN method is the standard BP using input set E. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.20:

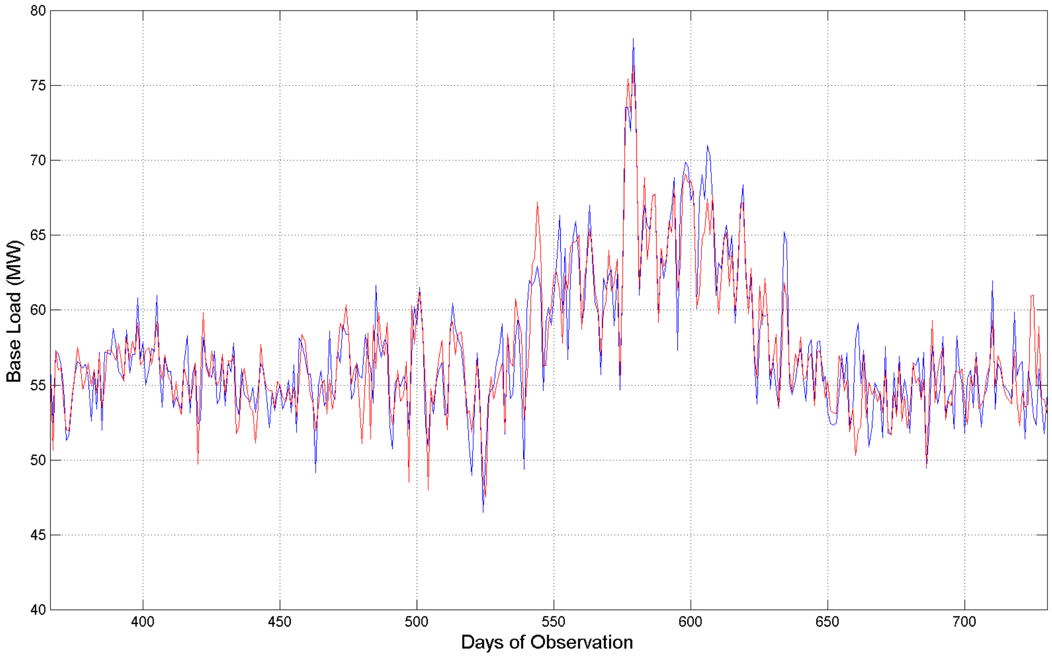


Figure 5.20: Prediction (red line) plotted against the measured base load in Substation S6 (blue line) over 360 days of observation.

Table 5.22: Error metrics for Average load, Substation S6

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	65,3	54,5	56,3	53,8	67,2	N/A	30,9	66,6	26,6
	KF	33,1	25,4	30,1	27,1	33,1	N/A	31,9	28,1	30,2
	PBP	205,6	173,1	193,6	193,5	136,6	N/A	173,8	231,9	176,3
	BP	70,6	220,6	238,9	249,3	91,2	N/A	146,5	237,3	223,5
MAPE	PKF	4,83	4,06	4,38	3,96	4,89	N/A	3,18	4,68	2,73
	KF	3,35	2,84	3,23	3,03	3,32	N/A	3,15	3,03	3,07
	PBP	8,93	8,13	8,70	8,74	7,22	N/A	8,23	9,45	8,29
	BP	5,36	9,44	9,62	9,90	6,03	N/A	7,61	9,23	9,36
MPE	PKF	38,1	38,0	45,6	39,7	32,5	N/A	37,1	30,8	32,2
	KF	35,9	26,4	34,1	30,9	30,1	N/A	36,7	27,9	27,5
	PBP	47,6	37,6	56,4	51,4	37,0	N/A	45,7	65,7	47,9
	BP	27,4	65,5	51,2	48,6	36,3	N/A	41,3	77,2	46,4
r^2	PKF	0,741	0,795	0,780	0,801	0,726	N/A	0,884	0,747	0,904
	KF	0,875	0,906	0,888	0,899	0,875	N/A	0,882	0,895	0,888
	PBP	0,279	0,402	0,280	0,306	0,401	N/A	0,185	0,304	0,413
	BP	0,734	0,148	0,293	0,378	0,595	N/A	0,257	0,243	0,349

The PCA-Kalman with input set Z is also the better method to forecast average load, followed by the Kalman filter with input set B. The better ANN method is again the BP using input set A. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.21:

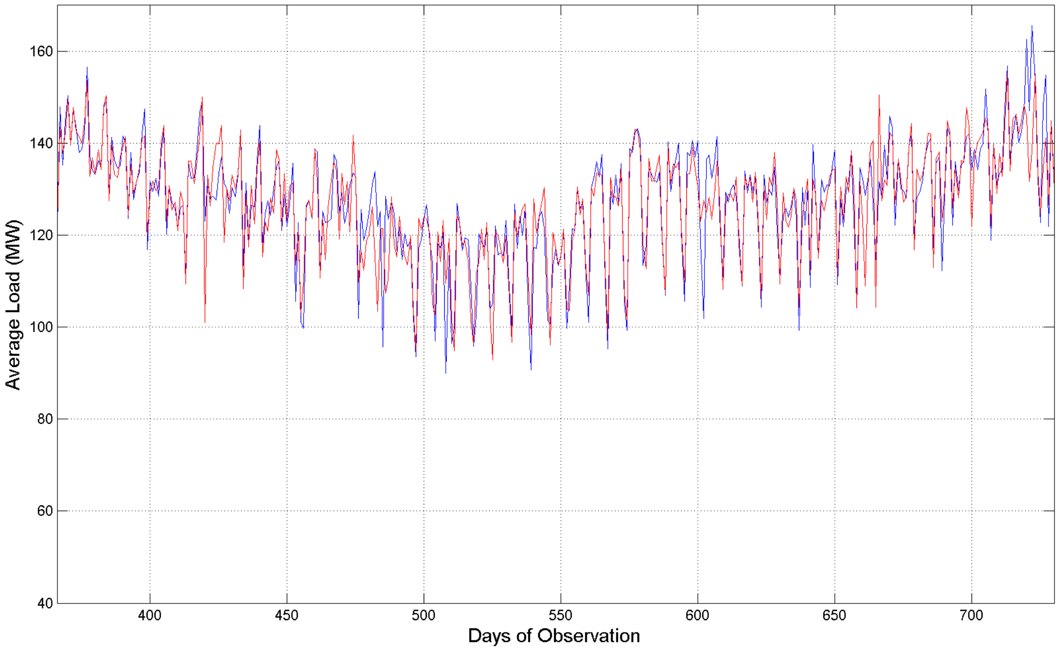


Figure 5.21: Prediction (red line) plotted against the measured average load in Substation S6 (blue line) over 360 days of observation.

Table 5.23: Error metrics for Peak load, Substation S6

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	223,0	154,5	162,7	148,6	208,1	N/A	190,7	233,4	126,2
	KF	170,1	136,1	169,4	164,2	180,3	N/A	671,2	149,6	154,0
	PBP	382,1	644,3	481,9	524,1	463,1	N/A	606,3	431,4	316,8
	BP	282,1	560,8	440,5	491,0	135,3	N/A	562,4	407,0	333,6
MAPE	PKF	5,62	4,55	4,82	4,44	5,45	N/A	5,10	5,54	4,05
	KF	5,02	4,20	4,91	4,77	5,08	N/A	5,79	4,54	4,46
	PBP	7,81	10,05	8,57	8,77	8,90	N/A	10,13	8,09	6,93
	BP	6,52	9,39	8,31	8,63	4,19	N/A	9,43	8,08	7,12

Table 5.24: Error metrics for Peak load, Substation S6 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	35,0	35,4	36,2	35,5	31,0	N/A	32,8	40,7	28,7
	KF	23,3	39,2	39,1	39,0	30,1	N/A	100,0	31,9	33,2
	PBP	30,2	52,6	57,5	50,0	45,6	N/A	46,5	38,2	27,1
	BP	37,4	46,2	40,8	56,9	35,3	N/A	41,3	32,5	32,8
r^2	PKF	0,861	0,906	0,901	0,911	0,870	N/A	0,881	0,860	0,924
	KF	0,895	0,917	0,896	0,900	0,888	N/A	0,673	0,909	0,907
	PBP	0,784	0,642	0,752	0,721	0,677	N/A	0,544	0,771	0,793
	BP	0,841	0,637	0,756	0,739	0,920	N/A	0,594	0,725	0,783

The PCA-Kalman with input set Z outperforms the other methods predicting peak load, closely followed by the BO with input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.22:

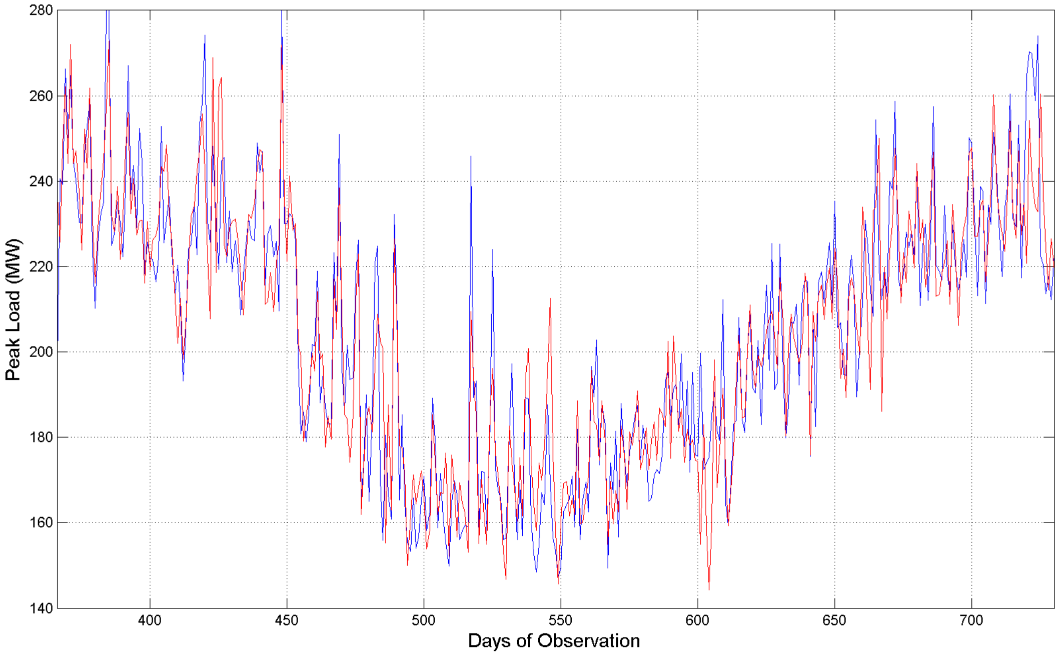


Figure 5.22: Prediction (red line) plotted against the measured peak load in Substation S6 (blue line) over 360 days of observation.

5.1.1.7 Substation S7

Substation S7 is located in the Grunau Siedlung district, west of Leipzig. This neighborhood has a high demographic density, and on average has 2,2 or more inhabitants per house. Population growth in this area is estimated to be above 15 % between 1999 and 2003. In average, 50 % of these residents are economically active. Tables 5.25, 5.26 and 5.27 present the forecasting results for base, average and peak load, respectively.

Table 5.25: Error metrics for Base load, Substation S7

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	27,4	17,3	17,6	15,7	34,0	N/A	27,1	25,9	16,1
	KF	23,2	16,1	22,4	19,3	27,6	N/A	78,2	18,9	17,3
	PBP	58,6	99,3	108,0	114,0	48,1	N/A	92,7	91,8	112,3
	BP	38,6	98,5	75,0	137,4	58,7	N/A	123,4	81,6	89,4
MAPE	PKF	5,13	4,13	4,04	3,86	5,43	N/A	4,96	4,86	3,63
	KF	4,85	3,85	4,66	4,38	5,08	N/A	5,68	4,21	4,11
	PBP	7,75	9,97	10,66	10,93	6,86	N/A	9,87	9,75	10,63
	BP	6,29	9,61	8,78	12,14	5,92	N/A	12,01	9,13	9,58
MPE	PKF	28,2	29,7	25,9	29,4	29,9	N/A	26,0	50,1	39,9
	KF	21,4	33,1	24,5	19,2	25,1	N/A	100,0	19,4	24,3
	PBP	44,2	54,6	93,7	53,6	32,7	N/A	67,8	53,0	86,8
	BP	44,1	58,5	54,6	104,3	62,7	N/A	50,9	54,3	52,2
r^2	PKF	0,967	0,979	0,978	0,981	0,958	N/A	0,967	0,969	0,981
	KF	0,972	0,980	0,973	0,976	0,966	N/A	0,909	0,977	0,979
	PBP	0,931	0,886	0,875	0,864	0,942	N/A	0,883	0,884	0,873
	BP	0,953	0,892	0,908	0,873	0,934	N/A	0,841	0,899	0,894

PCA-Kalman with input set Z and the Kalman filter with input set B perform similarly. Best ANN approach is BP with input set A. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.23:

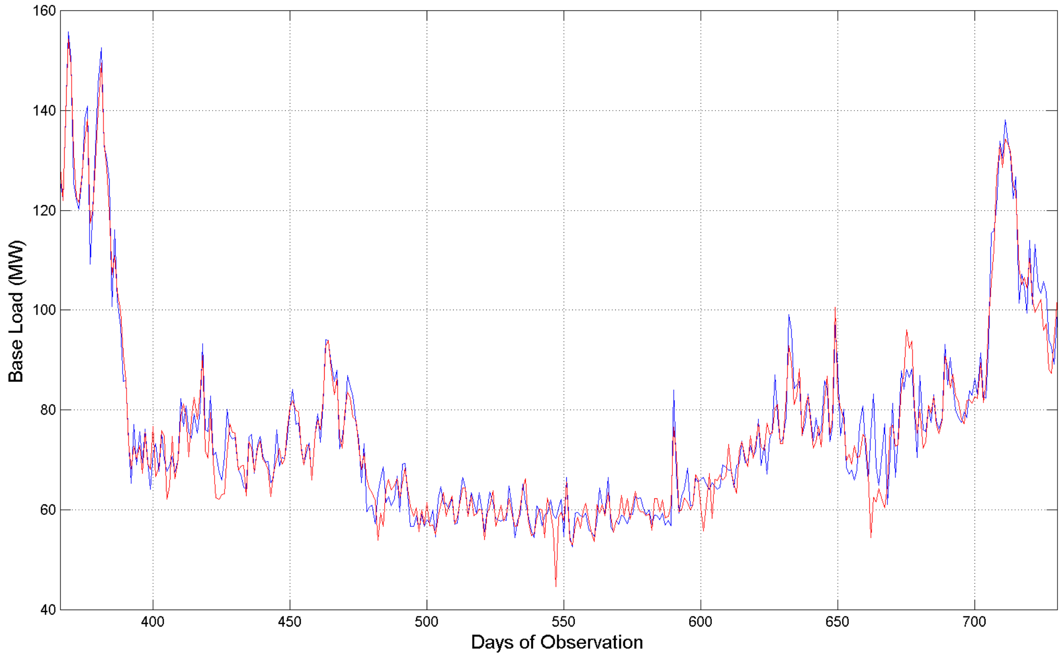


Figure 5.23: Prediction (red line) plotted against the measured base load in Substation S7 (blue line) over 360 days of observation.

Table 5.26: Error metrics for Average load, Substation S7

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	66,5	50,4	51,4	46,3	72,7	N/A	43,5	58,9	26,8
	KF	41,6	32,4	44,0	37,5	47,4	N/A	105,4	35,5	33,6
	PBP	212,1	226,9	224,0	150,4	134,2	N/A	194,0	198,5	220,8
	BP	147,8	216,6	237,0	198,8	63,9	N/A	130,3	181,8	355,2
MAPE	PKF	4,85	4,07	4,09	3,78	5,00	N/A	3,73	4,41	2,86
	KF	3,64	3,24	3,63	3,44	3,80	N/A	4,23	3,27	3,18
	PBP	8,77	9,60	9,61	7,69	7,33	N/A	8,72	9,04	9,03
	BP	6,95	9,49	9,83	9,13	4,56	N/A	7,20	8,54	12,74
MPE	PKF	39,8	32,7	54,0	38,5	41,1	N/A	32,7	40,2	25,1
	KF	38,9	24,1	39,4	31,2	41,4	N/A	80,1	33,8	32,0
	PBP	67,6	46,0	66,8	31,0	44,9	N/A	51,2	53,2	67,6
	BP	68,4	51,5	52,1	48,1	25,3	N/A	29,7	37,0	51,6
r^2	PKF	0,943	0,957	0,956	0,960	0,936	N/A	0,962	0,949	0,977
	KF	0,964	0,972	0,962	0,968	0,959	N/A	0,914	0,970	0,971
	PBP	0,828	0,824	0,818	0,870	0,881	N/A	0,827	0,845	0,836
	BP	0,876	0,840	0,789	0,842	0,946	N/A	0,883	0,833	0,654

The PCA-Kalman method with input set Z predicts the average load with the lowest MSE, followed by the classic Kalman with input set B. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.24:

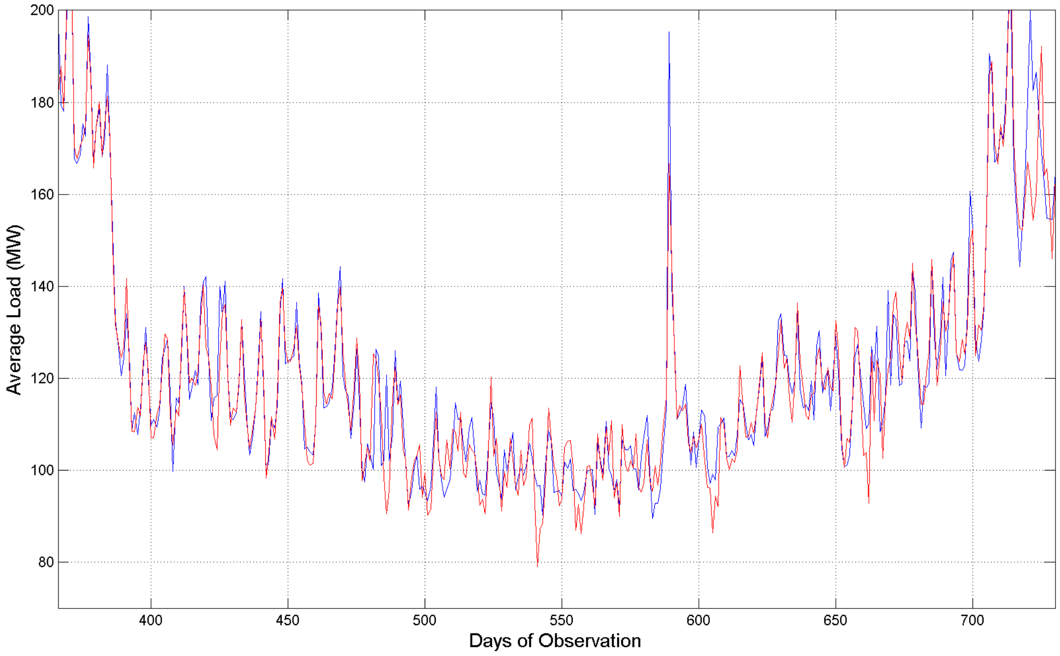


Figure 5.24: Prediction (red line) plotted against the measured average load in Substation S7 (blue line) over 360 days of observation.

Table 5.27: Error metrics for Peak load, Substation S7

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	266,4	193,2	226,4	185,5	272,8	N/A	154,8	249,4	112,1
	KF	153,7	113,9	147,9	130,4	163,6	N/A	313,8	126,4	118,6
	PBP	592,6	859,6	923,2	723,1	504,4	N/A	539,3	926,1	529,1
	BP	463,9	816,1	825,2	1200,1	288,3	N/A	633,0	1019,7	552,4
MAPE	PKF	6,90	5,64	6,15	5,54	6,96	N/A	5,22	6,43	4,08
	KF	5,21	4,39	5,04	4,77	5,38	N/A	5,71	4,67	4,50
	PBP	10,70	12,43	12,97	11,33	9,69	N/A	10,13	13,06	10,19
	BP	9,42	12,34	12,11	16,50	7,42	N/A	11,54	14,11	10,53

Table 5.28: Error metrics for Peak load, Substation S7 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	41,0	46,9	42,9	48,6	38,3	N/A	28,5	53,3	41,7
	KF	26,6	33,3	27,8	36,2	27,4	N/A	76,9	25,2	31,0
	PBP	47,1	56,1	54,2	54,6	36,9	N/A	37,9	68,6	48,2
	BP	87,6	61,9	73,7	84,7	53,9	N/A	49,6	71,6	53,6
r^2	PKF	0,908	0,934	0,922	0,936	0,905	N/A	0,947	0,915	0,963
	KF	0,947	0,961	0,950	0,956	0,944	N/A	0,897	0,957	0,960
	PBP	0,800	0,729	0,740	0,769	0,833	N/A	0,804	0,732	0,815
	BP	0,832	0,755	0,752	0,590	0,906	N/A	0,765	0,704	0,812

The PCA-Kalman approach with input set Z very slightly outperforms the classic Kalman with input set B. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.25:

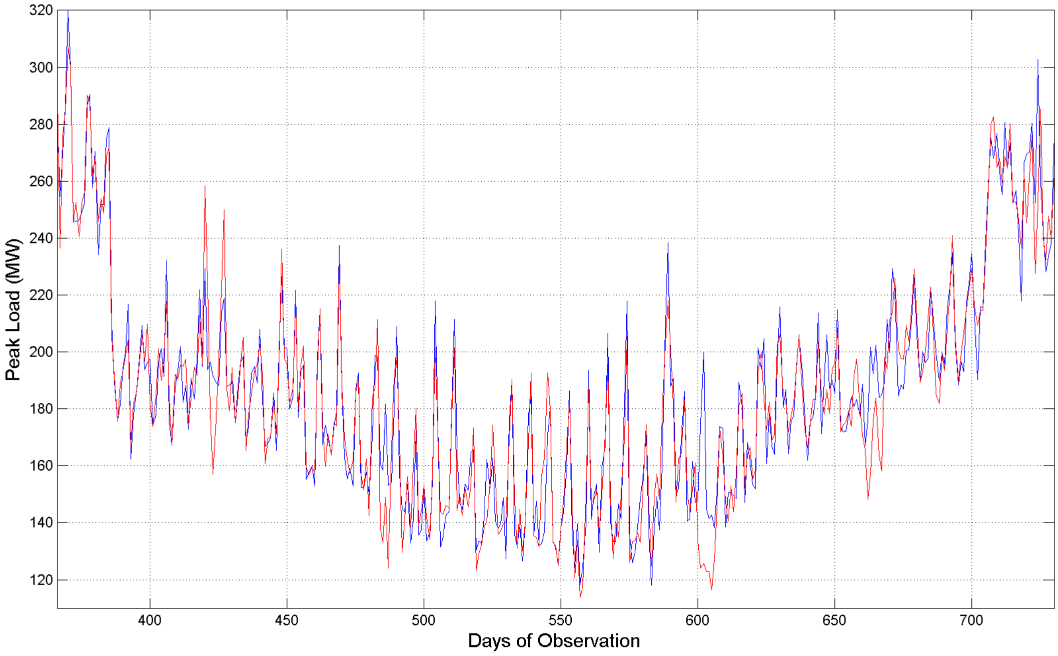


Figure 5.25: Prediction (red line) plotted against the measured peak load in Substation S7 (blue line) over 360 days of observation.

5.1.1.8 Substation S8

Substation S8 is located in the Lausen Grunau district, west of Leipzig. This neighborhood has a very low demographic density, and on average has less than 1,9 inhabitants per house. Population growth in this area is estimated to be between 9 % and 15 % between 1999 and 2003. In average, 60 % of these residents are economically active. Tables 5.29, 5.31 and 5.31 present the forecasting results for base, average and peak load, respectively.

Table 5.29: Error metrics for Base load, Substation S8

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	6,8	5,9	4,7	6,9	7,3	N/A	6,3	6,7	4,2
	KF	5,2	4,3	4,7	4,5	5,6	N/A	45,7	5,6	4,1
	PBP	13,0	12,2	20,2	21,6	14,4	N/A	11,2	19,8	20,7
	BP	9,2	20,3	13,2	21,1	8,9	N/A	10,8	13,4	14,8
MAPE	PKF	2,15	1,86	1,77	1,81	2,22	N/A	1,95	2,02	1,54
	KF	1,88	1,59	1,75	1,70	1,92	N/A	2,48	1,63	1,59
	PBP	3,09	2,90	3,74	4,03	3,16	N/A	2,85	3,67	3,96
	BP	2,55	3,71	3,08	3,87	2,55	N/A	2,79	3,02	3,18
MPE	PKF	12,9	22,4	10,4	32,8	10,6	N/A	14,7	17,7	28,3
	KF	10,8	15,5	11,3	11,8	11,7	N/A	77,3	9,4	11,8
	PBP	13,4	15,5	20,5	15,2	15,5	N/A	15,5	21,6	18,5
	BP	10,5	20,1	14,5	16,2	11,9	N/A	13,0	16,8	14,4
r^2	PKF	0,790	0,835	0,860	0,817	0,763	N/A	0,808	0,798	0,873
	KF	0,839	0,874	0,858	0,866	0,826	N/A	0,455	0,851	0,878
	PBP	0,613	0,615	0,504	0,276	0,586	N/A	0,662	0,556	0,406
	BP	0,693	0,334	0,655	0,478	0,720	N/A	0,634	0,579	0,590

Forecasting base loads, the Kalman filter approach with input set Z slightly the PCA-Kalman with this same input set. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.26:

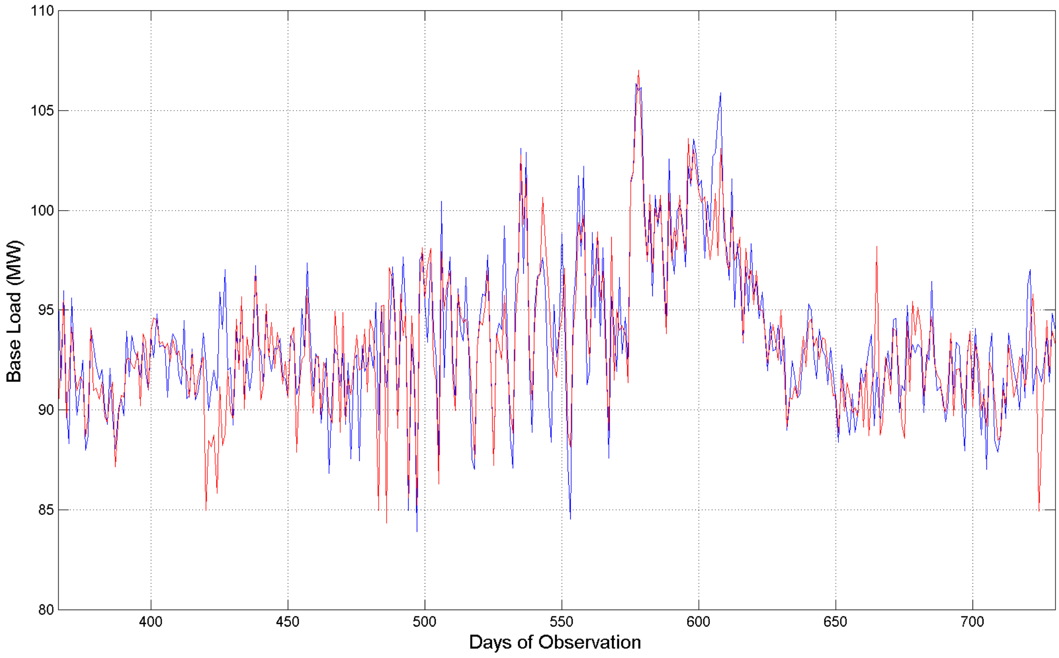


Figure 5.26: Prediction (red line) plotted against the measured base load in Substation S8 (blue line) over 360 days of observation.

Table 5.30: Error metrics for Average load, Substation S8

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	44,9	33,0	48,0	38,9	48,4	N/A	35,8	46,4	27,2
	KF	34,3	26,3	33,6	30,5	37,1	N/A	36,6	31,2	31,7
	PBP	107,3	102,3	137,6	145,5	97,0	N/A	113,3	122,2	103,0
	BP	67,5	156,0	122,0	166,9	50,7	N/A	75,8	89,0	105,5
MAPE	PKF	3,27	2,56	2,88	2,59	3,30	N/A	2,78	3,07	2,17
	KF	2,80	2,27	2,74	2,55	2,87	N/A	2,73	2,48	2,45
	PBP	5,08	4,97	5,65	5,87	5,02	N/A	5,20	5,52	4,99
	BP	4,03	5,92	5,56	6,88	3,51	N/A	4,35	4,66	5,20
MPE	PKF	14,6	19,5	40,7	32,3	16,5	N/A	15,9	26,0	21,5
	KF	14,6	19,0	16,0	19,6	15,2	N/A	22,1	18,1	21,4
	PBP	25,7	31,4	33,1	54,8	24,5	N/A	28,0	24,6	27,7
	BP	17,1	28,3	21,4	29,3	19,8	N/A	24,0	20,1	27,8
r^2	PKF	0,861	0,900	0,854	0,884	0,848	N/A	0,890	0,860	0,919
	KF	0,895	0,921	0,898	0,908	0,886	N/A	0,889	0,906	0,904
	PBP	0,689	0,741	0,670	0,626	0,683	N/A	0,590	0,663	0,672
	BP	0,781	0,408	0,751	0,533	0,855	N/A	0,766	0,713	0,722

For average loads, the Kalman filter approach with input set B slightly outperforms the PCA-Kalman with input set Z. BP with input set E has the best performance among ANN. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.27:

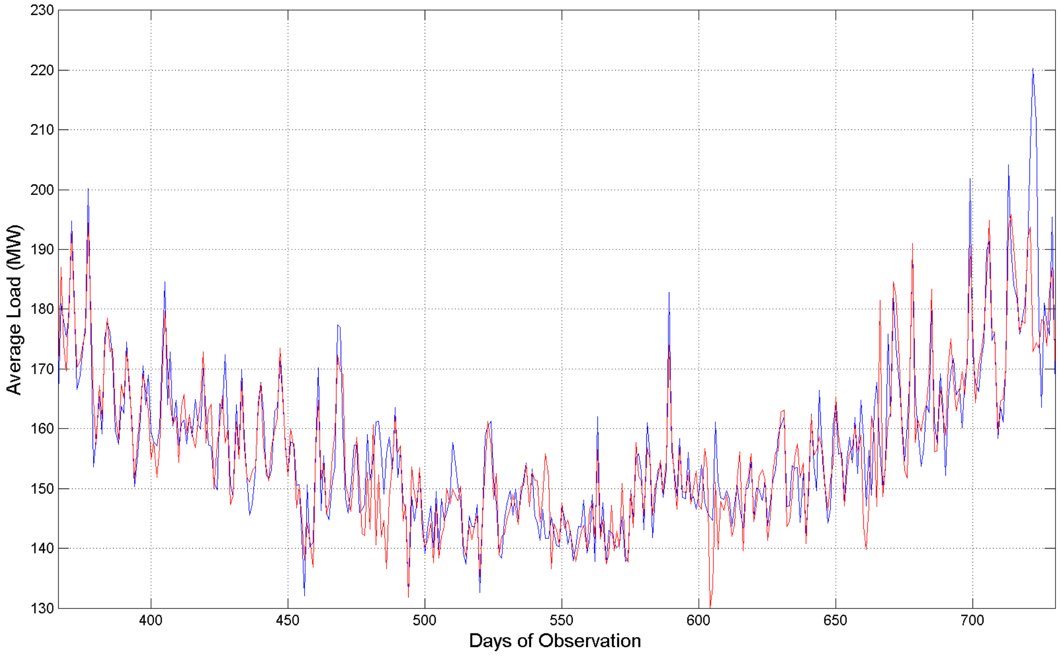


Figure 5.27: Prediction (red line) plotted against the measured average load in Substation S8 (blue line) over 360 days of observation.

Table 5.31: Error metrics for Peak load, Substation S8

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	384,8	330,7	355,2	357,9	392,8	N/A	217,7	379,6	154,3
	KF	201,6	164,6	202,2	216,8	205,4	N/A	970,7	176,8	168,9
	PBP	980,2	1400,8	1488,6	986,1	946,6	N/A	1024,4	1167,8	756,3
	BP	800,5	1351,2	1155,2	1569,2	594,4	N/A	815,3	717,9	785,2
MAPE	PKF	5,60	4,74	5,07	4,65	5,61	N/A	4,26	5,29	3,40
	KF	4,14	3,55	4,08	4,06	4,19	N/A	5,13	3,62	3,51
	PBP	9,85	11,64	11,61	10,17	9,25	N/A	10,04	10,81	8,28
	BP	8,75	12,35	11,27	12,77	8,08	N/A	8,64	8,37	8,44

Table 5.32: Error metrics for Peak load, Substation S8 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	37,5	40,5	47,3	45,6	44,8	N/A	31,4	46,5	31,9
	KF	26,6	30,8	37,4	32,8	26,7	N/A	100,0	29,3	32,9
	PBP	44,6	55,0	75,6	39,8	43,7	N/A	47,4	56,3	37,2
	BP	33,5	48,1	46,9	74,2	31,8	N/A	44,4	38,2	40,3
r^2	PKF	0,883	0,902	0,894	0,895	0,881	N/A	0,936	0,889	0,955
	KF	0,940	0,952	0,940	0,936	0,939	N/A	0,764	0,948	0,950
	PBP	0,679	0,644	0,671	0,660	0,721	N/A	0,695	0,688	0,783
	BP	0,737	0,595	0,674	0,582	0,861	N/A	0,753	0,785	0,756

The PCA-Kalman with input set Z outperforms all methods forecasting peak loads, followed by the Kalman filter with input set B. The better ANN method is the BP, using the input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.28:

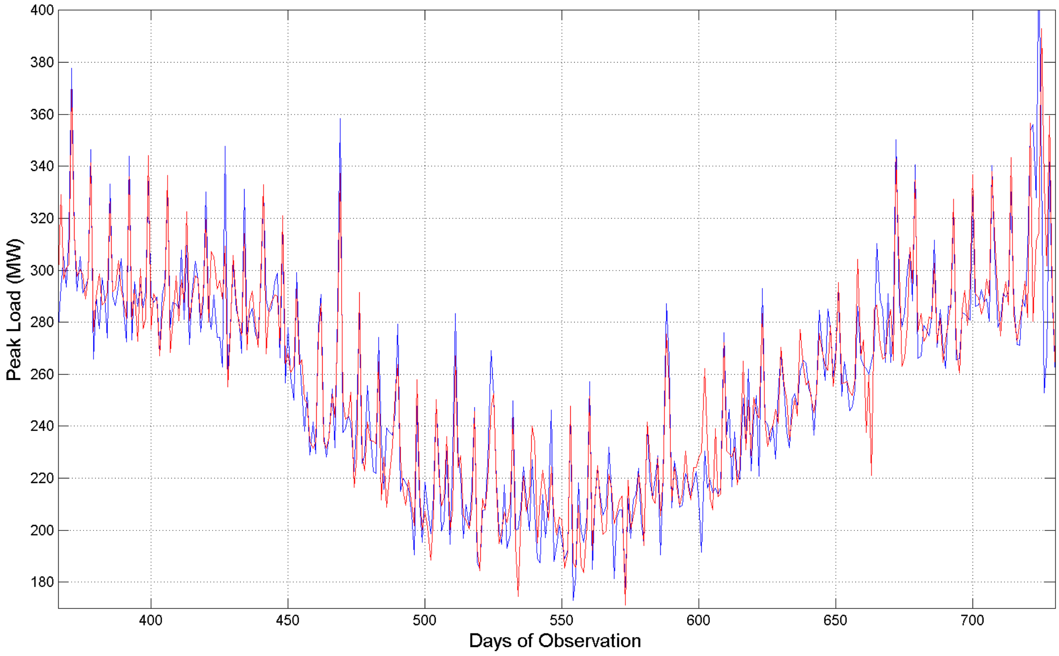


Figure 5.28: Prediction (red line) plotted against the measured peak load in Substation S8 (blue line) over 360 days of observation.

5.1.1.9 All substations combined

In order to evaluate the forescasting of a larger power system, the load of the eight substations is combined by means of simple summation. Tables 5.33, 5.34 and 5.35 present the forecasting results for base, average and peak load, respectively.

Table 5.33: Error metrics for Base load, Substation S9

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	108,9	76,9	68,0	66,0	130,8	N/A	97,5	97,2	57,3
	KF	81,2	61,0	75,7	69,2	89,6	N/A	1149,5	61,2	65,8
	PBP	216,8	289,9	367,2	360,5	311,0	N/A	296,9	349,2	267,4
	BP	129,1	389,6	387,3	278,2	76,0	N/A	542,2	438,2	395,3
MAPE	PKF	1,70	1,35	1,28	1,24	1,82	N/A	1,47	1,51	1,06
	KF	1,43	1,18	1,39	1,32	1,49	N/A	2,06	1,24	1,16
	PBP	2,27	2,66	2,89	2,90	2,80	N/A	2,64	3,00	2,68
	BP	1,83	3,10	3,00	2,54	1,41	N/A	3,30	3,18	3,18
MPE	PKF	8,5	7,3	7,7	9,9	9,6	N/A	10,2	12,3	16,8
	KF	6,5	13,1	6,1	8,0	6,6	N/A	75,2	7,3	6,4
	PBP	14,4	16,5	20,7	22,2	12,3	N/A	16,5	12,7	11,0
	BP	8,2	14,4	19,6	20,5	7,8	N/A	34,4	20,2	14,3
r^2	PKF	0,938	0,957	0,961	0,963	0,924	N/A	0,944	0,946	0,968
	KF	0,953	0,966	0,957	0,961	0,948	N/A	0,662	0,965	0,965
	PBP	0,874	0,839	0,786	0,817	0,814	N/A	0,822	0,803	0,852
	BP	0,926	0,771	0,830	0,857	0,962	N/A	0,728	0,778	0,791

The BP approach with input set E has the lowest MSE when predicting base load. The PCA-Kalman filter with set Z performs better than the classic Kalman at input set B, followed by standard BP with input set E. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.29:

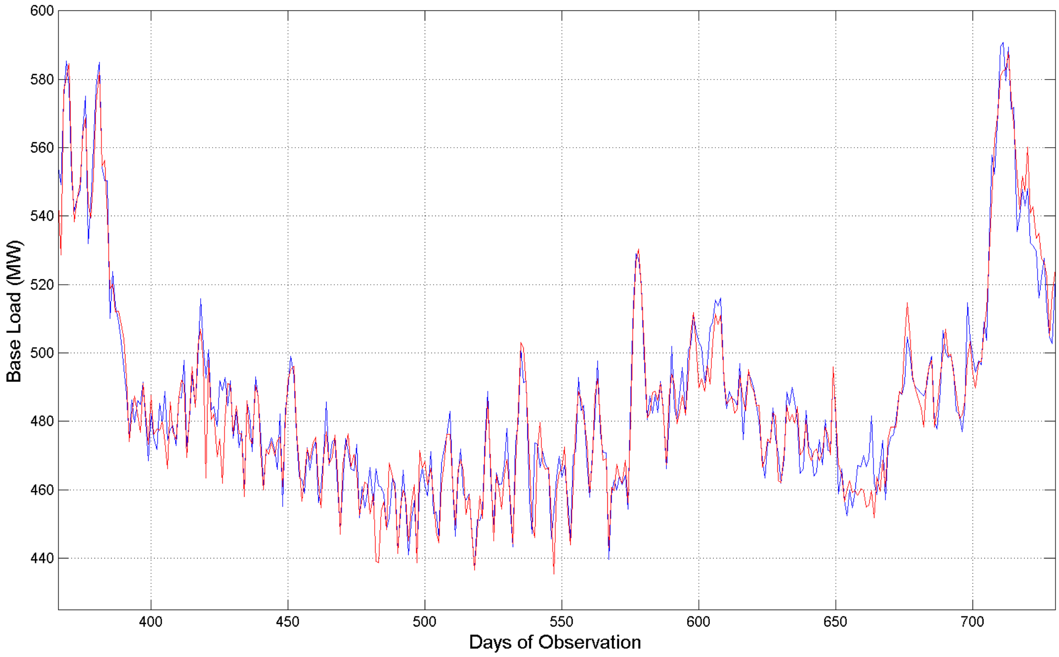


Figure 5.29: Prediction (red line) plotted against the measured base load in Substation S9 (blue line) over 360 days of observation.

Table 5.34: Error metrics for Average load, Substation S9

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	832,3	557,7	962,6	651,7	934,6	N/A	654,2	761,2	422,7
	KF	587,4	432,8	587,5	513,4	667,9	N/A	1097,9	503,1	484,9
	PBP	1689,5	3077,0	3876,6	3363,2	2342,4	N/A	4050,1	2440,4	3327,9
	BP	732,7	2898,0	3616,8	3471,4	731,1	N/A	2113,0	4223,7	2770,3
MAPE	PKF	2,44	1,85	2,13	1,79	2,62	N/A	2,05	2,28	1,48
	KF	2,01	1,64	1,97	1,80	2,12	N/A	2,20	1,80	1,74
	PBP	3,53	4,75	5,26	4,96	4,20	N/A	5,45	4,23	4,80
	BP	2,39	4,70	5,03	5,09	2,15	N/A	3,93	5,47	4,64
MPE	PKF	15,7	16,3	45,9	36,4	17,9	N/A	17,8	16,7	18,7
	KF	12,3	11,0	17,8	15,1	14,9	N/A	32,0	10,9	10,7
	PBP	17,5	24,1	33,7	22,3	20,6	N/A	24,1	22,3	24,0
	BP	11,6	19,4	31,6	26,5	12,5	N/A	18,3	31,1	16,9
r^2	PKF	0,945	0,963	0,937	0,957	0,937	N/A	0,957	0,950	0,972
	KF	0,961	0,972	0,961	0,966	0,956	N/A	0,929	0,967	0,968
	PBP	0,883	0,818	0,752	0,818	0,864	N/A	0,697	0,850	0,794
	BP	0,954	0,805	0,798	0,815	0,967	N/A	0,858	0,801	0,823

The PCA-Kalman approach with input set Z is slightly better than the Kalman filter with input set B, followed by the BP approach with input set E. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.30:

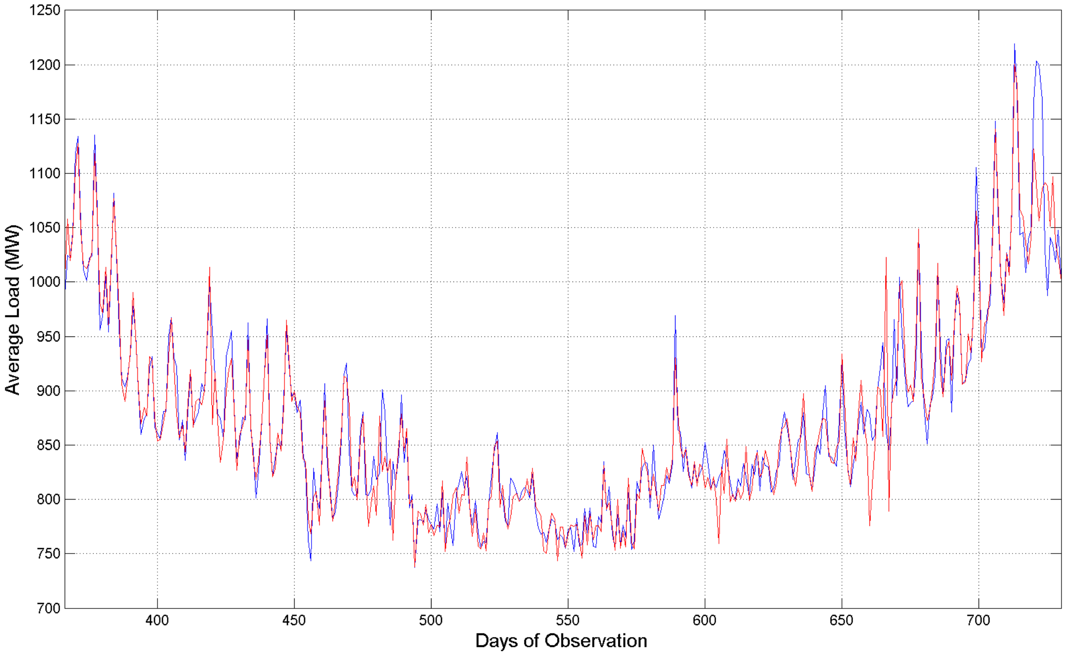


Figure 5.30: Prediction (red line) plotted against the measured average load in Substation S9 (blue line) over 360 days of observation.

Table 5.35: Error metrics for Peak load, Substation S9

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	4205,2	3729,3	3540,2	3390,2	4445,6	N/A	2258,3	3776,8	1498,0
	KF	2100,7	1641,0	2123,1	2223,7	2272,7	N/A	11699,6	1605,6	1508,3
	PBP	8294,1	9237,8	8934,9	9249,0	7270,0	N/A	7876,6	9155,3	8257,6
	BP	4764,5	15478,5	5087,3	9413,8	2680,1	N/A	8314,9	7535,8	11484,1
MAPE	PKF	3,38	2,85	2,98	2,74	3,51	N/A	2,48	3,09	1,85
	KF	2,47	2,03	2,39	2,27	2,57	N/A	2,96	2,09	2,01
	PBP	4,82	5,26	5,01	5,12	4,36	N/A	4,71	5,07	4,93
	BP	3,73	7,32	3,83	5,27	2,62	N/A	4,64	4,56	5,84

Table 5.36: Error metrics for Peak load, Substation S9 (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	21,2	30,6	25,1	30,3	18,8	N/A	23,9	19,8	22,4
	KF	15,9	14,4	20,3	22,6	17,3	N/A	69,6	15,7	14,5
	PBP	25,8	24,4	28,9	24,3	31,8	N/A	31,7	24,1	22,2
	BP	19,8	28,8	18,2	47,9	21,0	N/A	27,6	25,8	24,3
r^2	PKF	0,957	0,962	0,964	0,966	0,954	N/A	0,977	0,962	0,986
	KF	0,979	0,983	0,978	0,978	0,977	N/A	0,891	0,984	0,985
	PBP	0,914	0,904	0,909	0,905	0,924	N/A	0,917	0,906	0,914
	BP	0,952	0,830	0,948	0,900	0,973	N/A	0,915	0,921	0,880

Forecasting peak loads, the PCA-Kalman approach with input set Z slightly outperforms the Kalman filter with the same input set. The better ANN method is the standard BP, when using input set E. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.31:

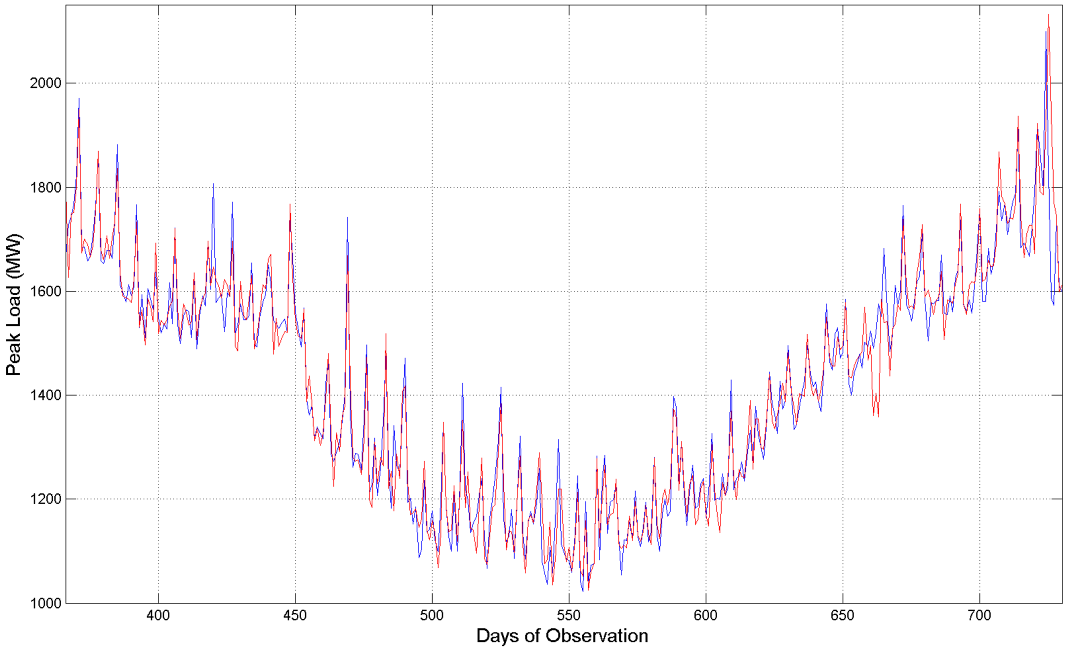


Figure 5.31: Prediction (red line) plotted against the measured peak load in Substation S9 (blue line) over 360 days of observation.

5.1.2 Second load forecasting scenario - Brasilia 2001-2003

Brasilia is the federal capital of Brazil. Located at coordinates 15.78S - 47.83W, it was founded in 1960, purpose built to serve as the new national capital closer to Brazil's geographic center. Currently, Brasilia and its metro area are estimated to be the 4th most populous city in Brazil, and it has the highest GDP per capita among major Latin American cities. The evolution of both population and GDP is shown in Figure 5.32.

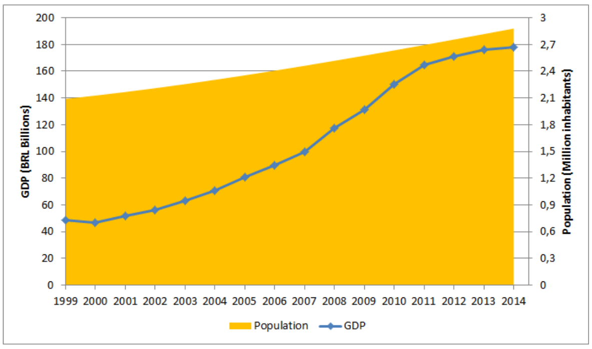


Figure 5.32: Evolution of Brasilia's population and GDP between 1999 and 2014. Credits: CODEPLAN

Besides being the political center, Brasilia is an important economic center, representing 3.76% of the total Brazilian GDP. The main economic activity of the federal capital results from its administrative function, with services accounting for more than 90% of the city's GDP. The public sector is the largest employer, providing around 40% of the city jobs. Besides the government, the city also hosts the headquarters of important companies, such as the two biggest public banks, the Brazilian postal service and a large telecommunications company.

Located in the middle of the Brazilian highlands, Brasilia has a tropical savanna climate with two distinct seasons. The rainy season occurs from October to April, while the dry season spans from May to September. September is also the hottest month, averaging 21.7 Celsius and maximas of 28.3 Celsius. The coldest month is July, averaging 18.3 Celsius and 12.9 Celsius minima. Relative Humidity oftenly drops below 50% between July and September. Average insolation hours vary from 138 in December to 266 in July, mainly determined by the presence of clouds in the sky.

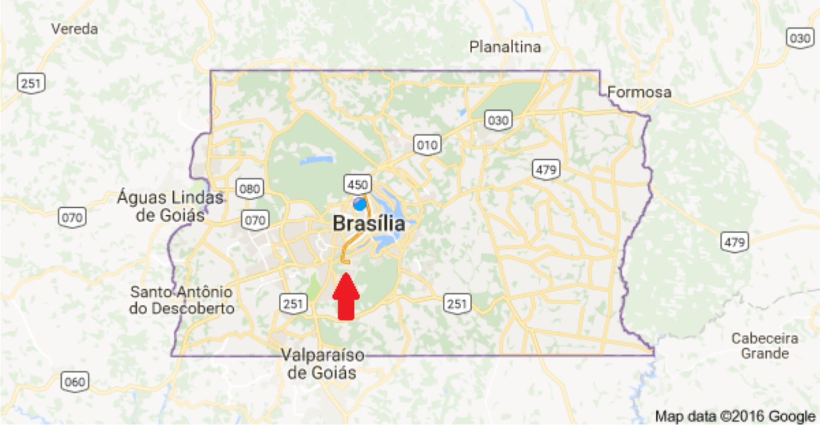


Figure 5.33: Location of Juscelino Kubistchek International Airport relative to Brasilia and the Federal District. Credits: Google Maps

The historical weather data has been collected from the Juscelino Kubistchek International Airport METeorological Aerodrome Reports (METAR), located in a central position relative to the larger load centers as shown in Fig. 5.33. Similarly to what occurred in Leipzig weather measurements, METAR data regarding January 2004 is unavailable. As such, the forecasting of Brasilia electric load has been divided in two scenarios: from July 1st 2001 to December 2003 and from February 2004 to June 2010. Coincidentally, the first period is concurrent with an electricity supply crysis, while the second coincides with a strong economic growth cycle. The first period is analised in this Subsection, while the second period is the third forecasting scenario analised in Subsection 5.1.3. Peak, average and base load in the first period are illustrated in Figure 5.34.

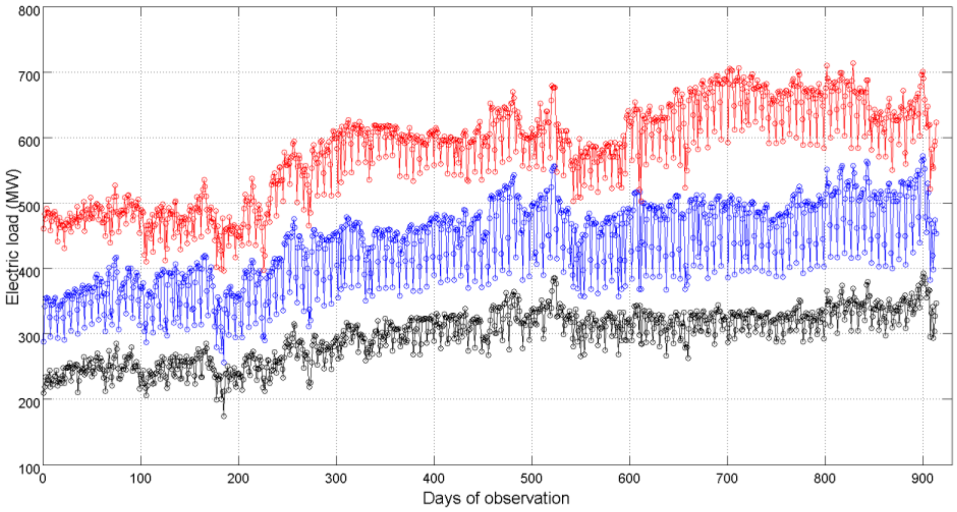


Figure 5.34: Evolution of electric load in Brasilia, from July 2001 to December 2003. Base load is plotted in black, Average load in blue and Peak load in red.

The proposed and the benchmark forecasting methods are used to predict the total load supplied by Brasilia's distribution company. This scenario uses a shorter training period of 182 days, between July 1st 2001 to December 31st 2001, while the prediction period comprises 729 days between January 2002 and December 2003. Error metrics are calculated exclusively for the prediction period.

In order to validate the proposed PCA-Kalman load forecasting system (PKF) performance, similarly to the Leipzig scenario, the load time series have been forecast by concurrent methods of linear and nonlinear natures. A classical Kalman Filter (KF) without PCA and variance estimation represent the linear approaches, while a classical BP double layer Artificial Neural Network (BP) and a PCA enhanced BP ANN (PBP) are employed to showcase the performance of these nonlinear methods. The Kalman filter methods employ an model order estimation in the initialization phase, in this scenario eight is selected as the size of the state vector, as shown in Figure 5.35.

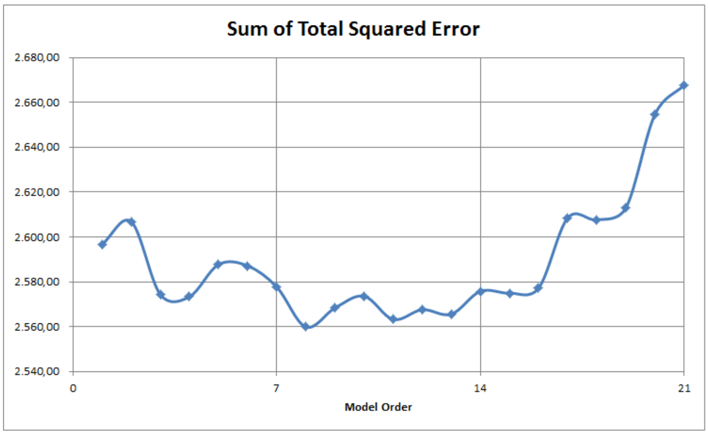


Figure 5.35: Total Squared Error for the second scenario, as a function of Model Order. The minimum is achieved when the Order is set to 8.

The above described benchmark models are used to forecast base, average and peak demand. For each prediction the four error metrics are calculated. Nine input sets are tested, each designated by a capital letter. The input sets have been described in Chapter 4, Table 4.1. Over the results presented in [87], this work expands the scope by adding the input set H, which includes solar resource and natural illumination inputs.

The forecasting period starts at January 1st 2002 and comprises 729 days. Tables 5.37, 5.38 and 5.39, respectively, summarize results for base, average and peak load forescasting.

Table 5.37: Error metrics for Base load, Brasilia first period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	162,8	91,4	114,9	76,4	234,3	189,3	92,1	40,5	39,7
	KF	155,1	92,3	137,3	88,1	251,4	218,7	92,1	322,5	322,5
	PBP	177,8	336,4	517,0	474,0	130,6	110,5	321,4	392,1	354,8
	BP	137,4	348,6	353,4	500,4	165,3	94,3	252,6	388,8	397,7
MAPE	PKF	3,05	2,33	2,63	2,12	3,47	3,29	2,26	1,45	1,45
	KF	2,99	2,30	2,86	2,24	3,53	3,07	2,26	4,45	4,45
	PBP	3,32	4,49	5,64	5,14	2,90	2,69	4,23	4,90	4,55
	BP	2,92	4,57	4,47	5,55	3,31	2,39	3,65	4,74	4,80
MPE	PKF	20,8	11,8	13,1	11,1	24,6	19,6	15,1	10,7	9,8
	KF	20,3	14,5	16,6	14,3	30,4	35,9	15,1	21,5	21,5
	PBP	19,3	18,4	21,6	37,4	13,2	13,2	18,9	20,6	18,5
	BP	11,8	18,8	20,7	23,0	13,0	11,4	17,7	22,7	22,0

The predictions provided by the PCA-Kalman are compared to the real values in figure 5.36:

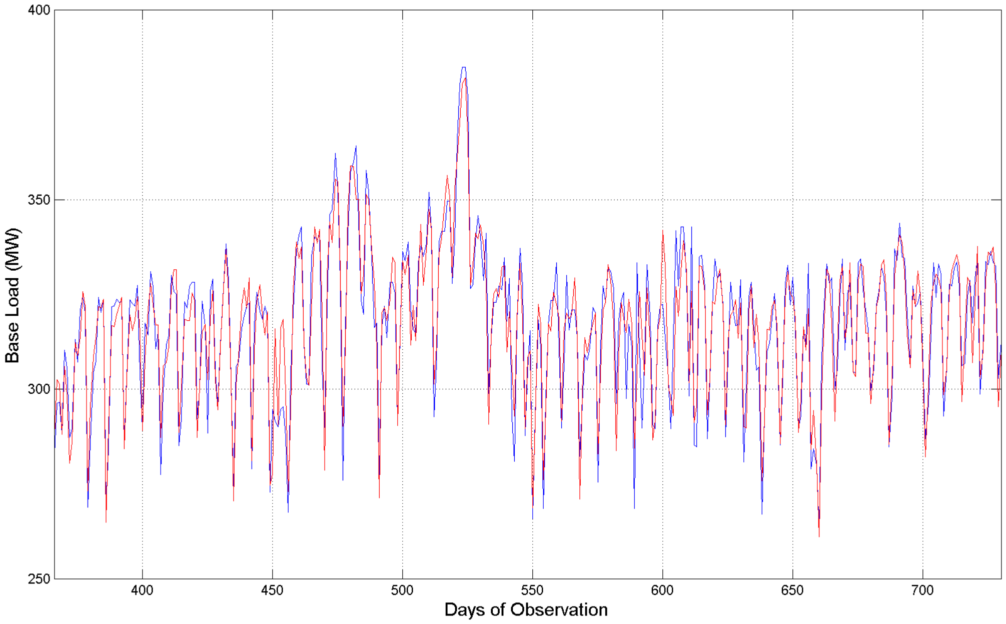


Figure 5.36: Prediction (red line) plotted against the measured base load in Brasilia (2001-2003 period) over 360 days of observation.

Note that all input sets provide reasonable forecasting performance. For the state space approaches, set C slightly outperforms input set A, as D also outperforms B, giving evidence that the performed preprocessing is beneficial to linear predicting algorithms. The ANN methods, however, are negatively affected. Input set F works well with the neural networks. Input set Z combined with the PCA-Kalman load forecasting system provide the best performance.

Table 5.38: Error metrics for Average load, Brasilia first period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	579,0	339,2	454,4	302,2	1403	631,6	204,2	77,4	73,6
	KF	554,2	343,4	484,8	338,7	851,8	640,2	204,2	991,7	991,7
	PBP	1247	1780	1306	1999	1544	1152	1945	1343	1435
	BP	1567	2597	1685	1988	1170	1220	1648	1522	1449
MAPE	PKF	3,62	2,90	3,26	2,59	4,71	3,87	2,14	1,06	1,02
	KF	3,51	2,89	3,37	2,81	4,16	3,57	2,14	3,90	3,90
	PBP	5,79	7,48	6,23	7,72	6,71	5,91	7,33	4,52	4,62
	BP	6,69	9,13	7,09	7,54	6,06	6,02	6,75	4,97	4,64
MPE	PKF	40,8	18,3	23,4	22,8	54,9	37,0	22,1	6,4	6,9
	KF	40,4	18,4	33,6	19,3	39,9	37,0	22,1	38,7	38,7
	PBP	28,2	28,8	21,3	33,7	35,2	29,9	30,0	21,1	31,4
	BP	36,1	26,4	31,0	40,5	30,9	30,8	25,2	27,0	24,2

Overall, the prediction of average load displays the largest error metrics, probably due to the larger quantity of outliers in this particular time series. The only exception is the PCA-Kalman system, as it shows smaller relative errors at the cost of increased maximum error, as compared with the base load prediction problem. ANN do not seem to perform well in this scenario, displaying large error metrics. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.37:

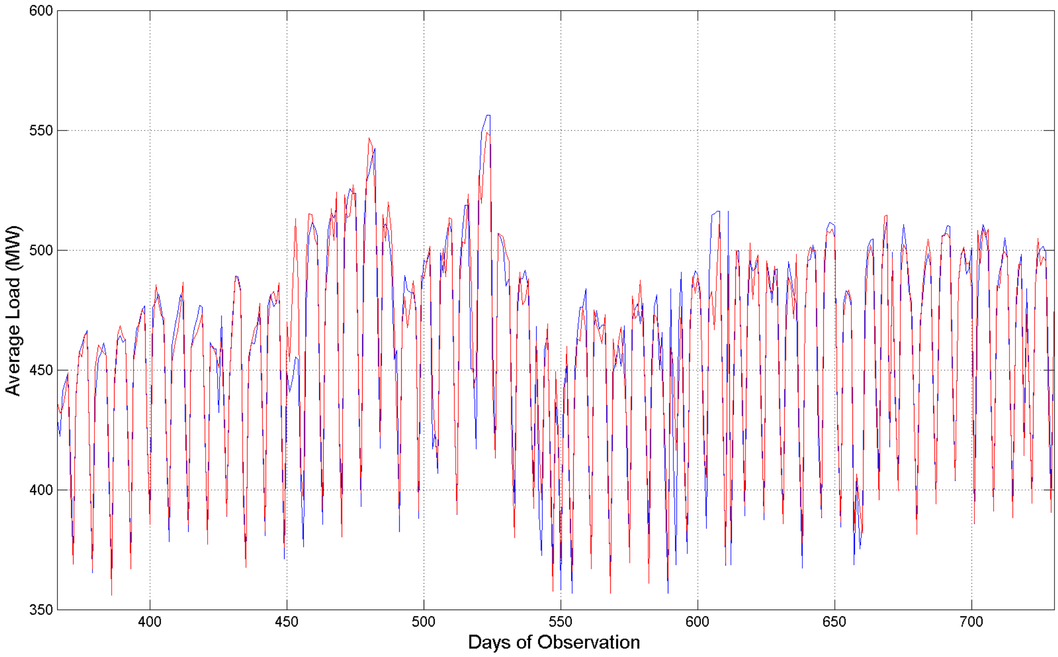


Figure 5.37: Prediction (red line) plotted against the measured average load (blue line) in Brasilia (2001-2003 period) over 360 days of observation.

Table 5.39: Error metrics for Peak load, Brasilia first period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	501,8	294,2	389,2	263,3	1046,3	529,3	189,9	77,4	73,6
	KF	491,0	289,5	413,4	275,4	786,4	494,4	189,9	991,7	991,7
	PBP	646,0	1365	1079	1476	813,3	550,0	1627	1343	1435
	BP	666,1	1988	808,2	1630	587,2	561,5	1205	1522	1449
MAPE	PKF	2,63	2,06	2,30	1,88	3,11	2,72	1,69	1,06	1,02
	KF	2,60	2,06	2,38	1,98	2,98	2,43	1,69	3,90	3,90
	PBP	3,22	4,68	4,20	4,75	3,50	2,99	5,20	4,52	4,62
	BP	3,15	5,74	3,63	5,15	3,06	2,79	4,13	4,97	4,64
MPE	PKF	27,0	15,4	18,3	14,2	39,4	28,0	8,8	6,4	6,9
	KF	27,0	14,8	22,5	14,3	31,5	28,0	8,8	38,7	38,7
	PBP	22,4	25,5	16,5	23,5	19,5	15,2	21,6	21,1	31,4
	BP	24,5	23,0	17,0	26,7	15,2	25,2	25,6	27,0	24,2

The proposed PCA-Kalman based approach vastly outperforms the other methods for peak load prediction. The KF achieves a MSE almost three times larger, yet forecasting with good accuracy. ANN methods produce better results when employing input set F.

Overall, the proposed system displays good forecasting performance, being capable of daily predicting demands with MAPE lower than 2 % in all scenarios. In comparison, the linear and nonlinear predictors employed as benchmark could only achieve MAPE lower than 2.5%, at best. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.38:

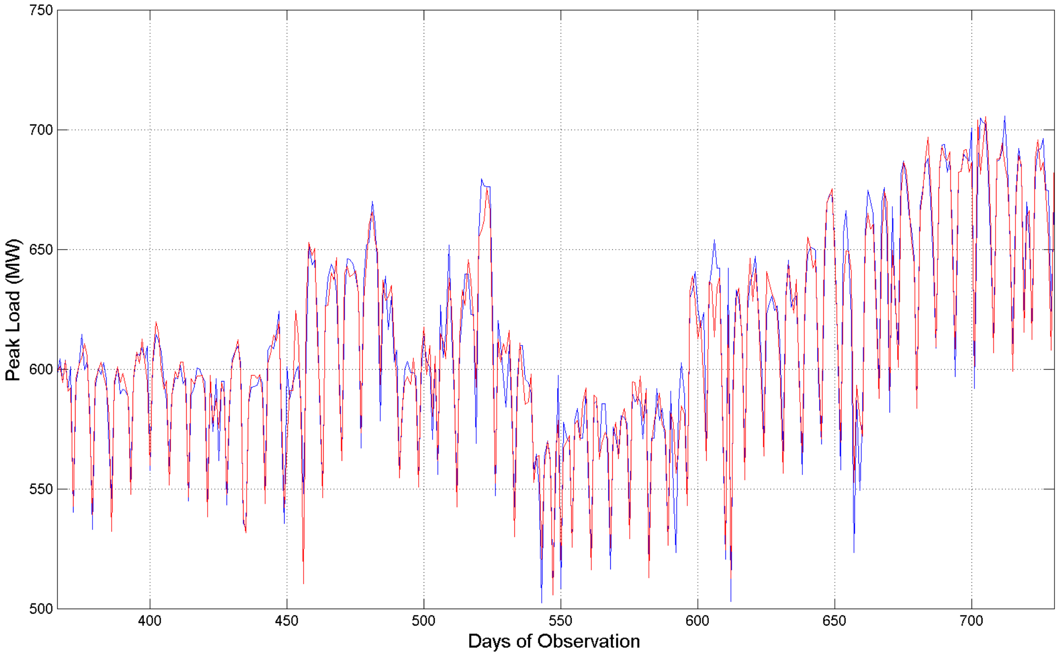


Figure 5.38: Prediction (red line) plotted against the measured peak load (blue line) in Brasilia (2001-2003 period) over 360 days of observation.

5.1.3 Third load forecasting scenario - Brasilia 2004-2010

The third load forecasting scenario is also performed with Brasilia, starting at February 1st 2004. As explained in subsection 5.1.2, this time period in Brasilia is characterized by strong growth in both population and economic output. As a consequence, in contrast with the mild increasing trend shown in Figure, in this scenario the electric loads increase by circa 30% in the time period, as illustrated in Figure 5.39.

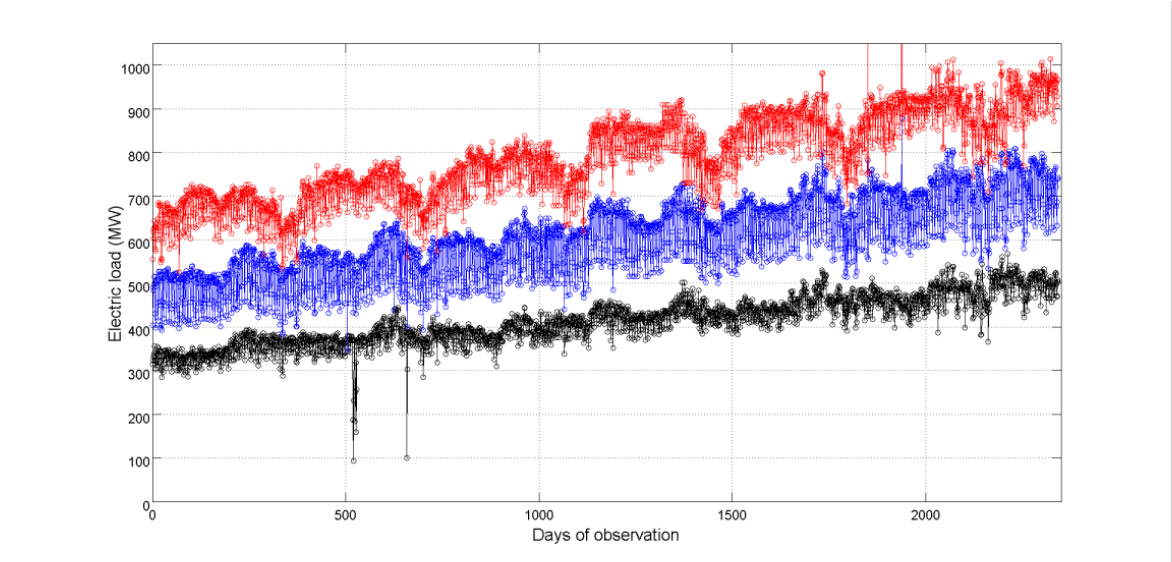


Figure 5.39: Evolution of electric load in Brasilia, from February 2004 to June 2010. Base load is plotted in black, Average load in blue and Peak load in red. Two outliers in the Peak load are not visible in this graph.

The proposed and the benchmark forecasting methods are used to predict the total load supplied by Brasilia's distribution company. This scenario uses a training period of 365 days, between February 1st 2001 to January 31st 2002, while the prediction period comprises 1977 days between Februart 2002 and June 2003. Error metrics are calculated exclusively for the prediction period. Similarly to Leipzig forecasts, seven is selected as chosen as the model order of the Kalman based methods, as shown in Figure 5.40.

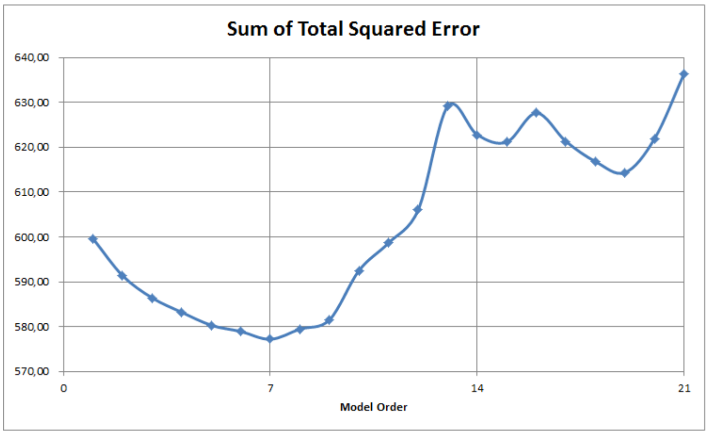


Figure 5.40: Total Squared Error for the second scenario, as a function of Model Order. The minimum is achieved when the Order is set to 8.

The forecasting period starts at February 1st 2002 and comprises 1977 days. Tables 5.40, 5.41 and 5.42, respectively, summarize results for base, average and peak load forescasting.

Table 5.40: Error metrics for Base load, Brasilia second period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	399,6	290,5	307,5	262,8	1427,5	418,2	224,9	364,8	184,8
	KF	373,0	257,6	454,5	236,2	610,7	382,9	226,7	588,8	206,1
	PBP	208,2	501,4	391,5	599,5	257,6	202,5	613,7	291,4	414,6
	BP	203,9	395,4	288,7	680,5	242,4	200,7	551,8	278,0	715,8
MAPE	PKF	3,48	2,85	3,10	2,65	4,73	3,61	2,55	3,26	2,11
	KF	3,39	2,68	3,19	2,60	3,84	3,39	2,56	3,87	2,38
	PBP	2,53	4,19	3,23	4,59	2,89	2,52	4,34	3,13	3,81
	BP	2,36	3,66	2,91	4,55	2,86	2,36	4,07	2,99	4,56
MPE	PKF	258,1	109,4	162,6	97,1	255,3	276,5	232,5	242,3	125,4
	KF	244,9	128,3	176,4	151,5	254,9	185,1	232,5	218,6	232,1
	PBP	204,1	203,2	210,9	155,5	186,8	197,5	281,7	204,9	190,3
	BP	124,3	202,1	220,1	146,2	172,5	159,8	212,9	212,9	259,3
r^2	PKF	0,918	0,942	0,937	0,948	0,764	0,914	0,955	0,925	0,963
	KF	0,924	0,948	0,910	0,952	0,876	0,922	0,954	0,880	0,958
	PBP	0,954	0,898	0,921	0,910	0,949	0,960	0,871	0,943	0,914
	BP	0,957	0,921	0,949	0,862	0,953	0,961	0,884	0,946	0,851

Forecasting base load, the PCA-Kalman filter with input set Z obtains the lowest MSE, followed by the BP ANN with input set F. The predictions provided by the PCA-Kalman are compared to the real values in figure 5.41:

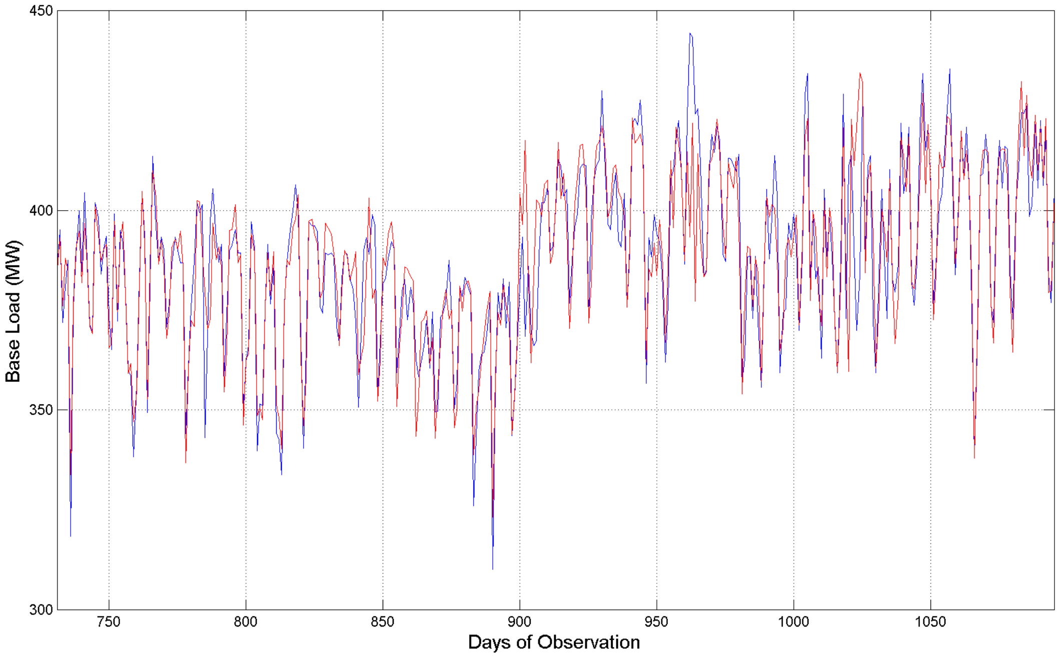


Figure 5.41: Prediction (red line) plotted against the measured base load (blue line) in Brasilia (2004-2010 period) over 360 days of observation.

Table 5.41: Error metrics for Average load, Brasilia second period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	1218,6	918,9	1028,3	842,9	4288	1345,3	387,5	1123,1	313,6
	KF	1167,4	884,7	1454,9	1057,4	5072,7	1369,8	389,2	3190,9	375,8
	PBP	2110	1286	1508	1711	1675	1636	2726	2158	3090
	BP	1990	1484	1914	2733	2162	2083	3116	1160	3156
MAPE	PKF	3,96	3,29	3,67	3,05	5,72	4,20	2,10	3,75	1,87
	KF	3,90	3,22	3,81	3,24	5,44	4,04	2,11	6,34	2,13
	PBP	5,74	4,68	5,03	5,34	5,45	5,36	6,54	5,89	7,17
	BP	5,71	5,04	5,70	6,94	6,01	5,86	7,16	4,48	7,20
MPE	PKF	37,6	36,9	53,2	45,4	118,7	37,9	25,0	51,0	23,5
	KF	37,3	41,0	100,0	80,9	241,3	44,0	25,2	111,0	23,5
	PBP	76,3	33,2	38,6	58,9	40,6	30,6	40,9	63,1	43,8
	BP	42,9	42,0	42,3	50,4	71,0	58,5	40,6	30,6	55,9
r^2	PKF	0,904	0,930	0,920	0,936	0,737	0,893	0,970	0,912	0,976
	KF	0,908	0,932	0,890	0,919	0,711	0,893	0,970	0,754	0,971
	PBP	0,828	0,904	0,880	0,866	0,865	0,868	0,781	0,826	0,749
	BP	0,838	0,884	0,848	0,770	0,822	0,832	0,754	0,918	0,751

For the average load, the PCA-Kalman method vastly outperforms the ANN approaches, also obtaining a 20 % lower MSE than the Kalman filter. The forecasts obtained from the PCA-Kalman are compared to the real values in figure 5.42:

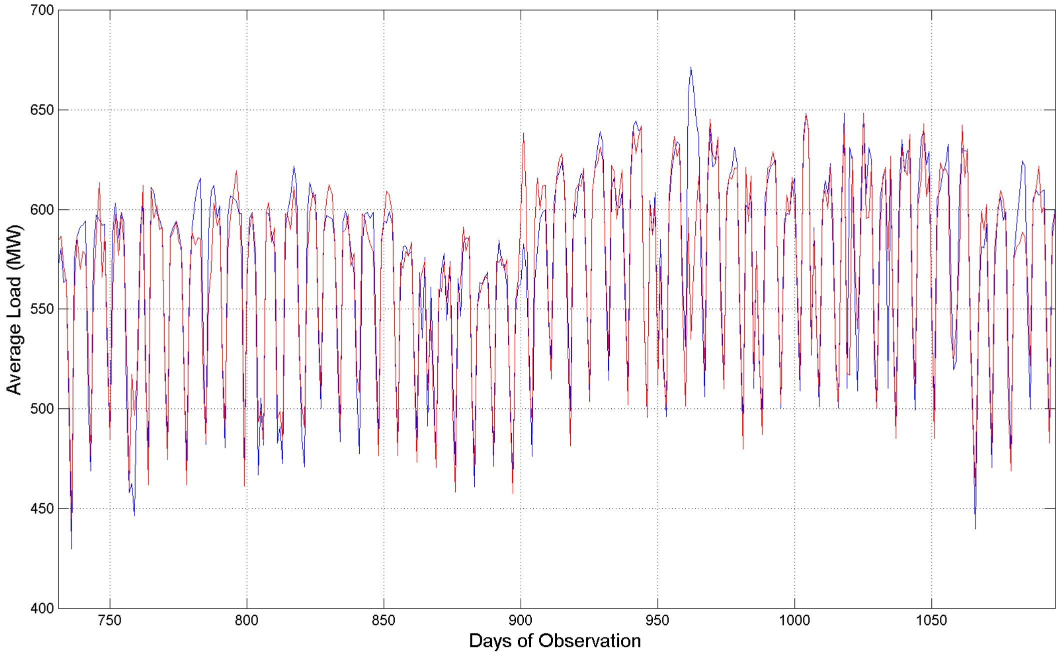


Figure 5.42: Prediction (red line) plotted against the measured average load (blue line) in Brasilia (2004-2010 period) over 360 days of observation.

Table 5.42: Error metrics for Peak load, Brasilia second period

Metric	Method	A	B	C	D	E	F	G	H	Z
MSE	PKF	1218,6	918,9	1028,3	842,9	4288	1345,3	387,5	1123,1	313,6
	KF	1640,1	1186,5	1850,8	1435,4	3363,8	1686,3	1411,3	2581,8	1609,8
	PBP	1950	3751	6340	5500	2421	1860	3292	2705	3130
	BP	2192	5703	3287	3890	2489	2303	3311	3278	4039
MAPE	PKF	3,96	3,29	3,67	3,05	5,72	4,20	2,10	3,75	1,87
	KF	6,02	6,02	6,02	6,02	6,02	6,02	6,02	6,02	6,02
	PBP	5,57	8,02	10,36	9,63	6,41	5,69	7,56	6,77	7,19
	BP	6,09	9,75	7,35	7,91	6,57	5,88	7,14	7,20	8,27

Table 5.43: Error metrics for Peak load, Brasilia second period (continuation)

Metric	Method	A	B	C	D	E	F	G	H	Z
MPE	PKF	37,6	36,9	53,2	45,4	118,7	37,9	25,0	51,0	23,5
	KF	47,0	47,0	47,0	47,0	47,0	47,0	47,0	47,0	47,0
	PBP	73,3	45,0	58,6	56,4	58,7	49,4	42,3	43,0	56,8
	BP	45,1	70,8	49,4	57,2	62,7	110,0	43,2	41,2	57,0
r^2	PKF	0,904	0,930	0,920	0,936	0,737	0,893	0,970	0,912	0,976
	KF	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818
	PBP	0,840	0,683	0,315	0,605	0,811	0,851	0,734	0,780	0,761
	BP	0,823	0,631	0,720	0,724	0,800	0,812	0,747	0,759	0,668

In the peak load forecasting, the PCA-Kalman filter with input set Z achieves the lowest MSE of the comparison. All other methods perform relatively poorly in this case the second best being the classic Kalman. The peak load predictions provided by the PCA-Kalman are compared to the real values in figure 5.43:

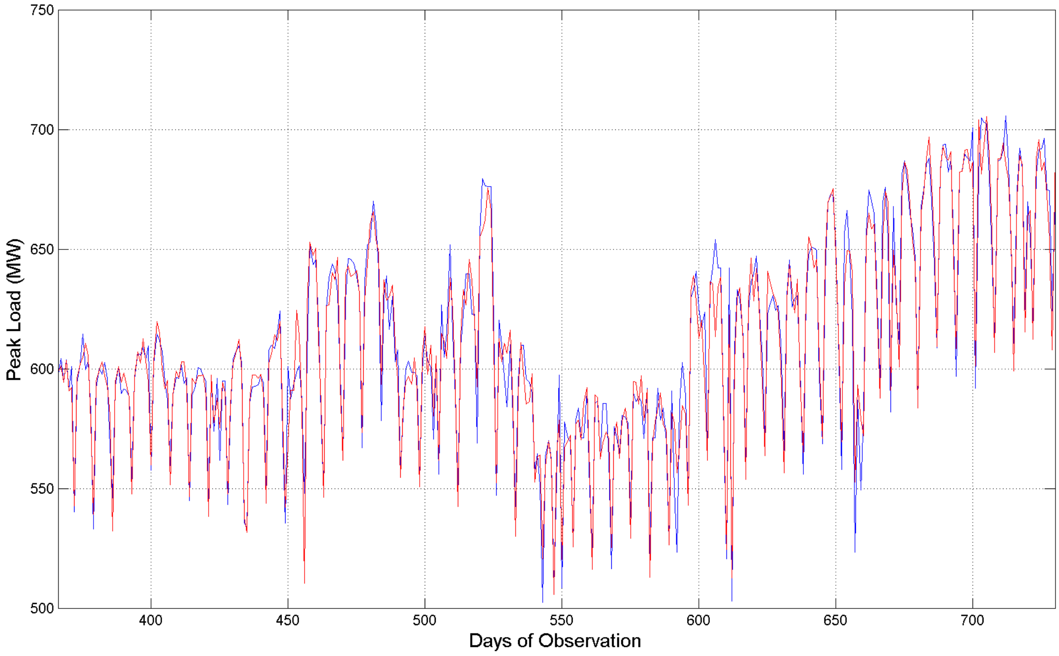


Figure 5.43: Prediction (red line) plotted against the measured peak load (blue line) in Brasilia (2004-2010 period) over 360 days of observation.

5.2 Photovoltaic generation forecasting

The forecasting methodologies are applied to forecast generation in 17 different photovoltaic systems installed in three continents. These PV generators are chosen due to the online availability of production data and proximity to airport weather stations, allowing the analisys of both electricity production and weather time series. The systems capacity range from 0.625 kWp to 24.5 kWp, installed in residential units and directly connected to the distribution grid.

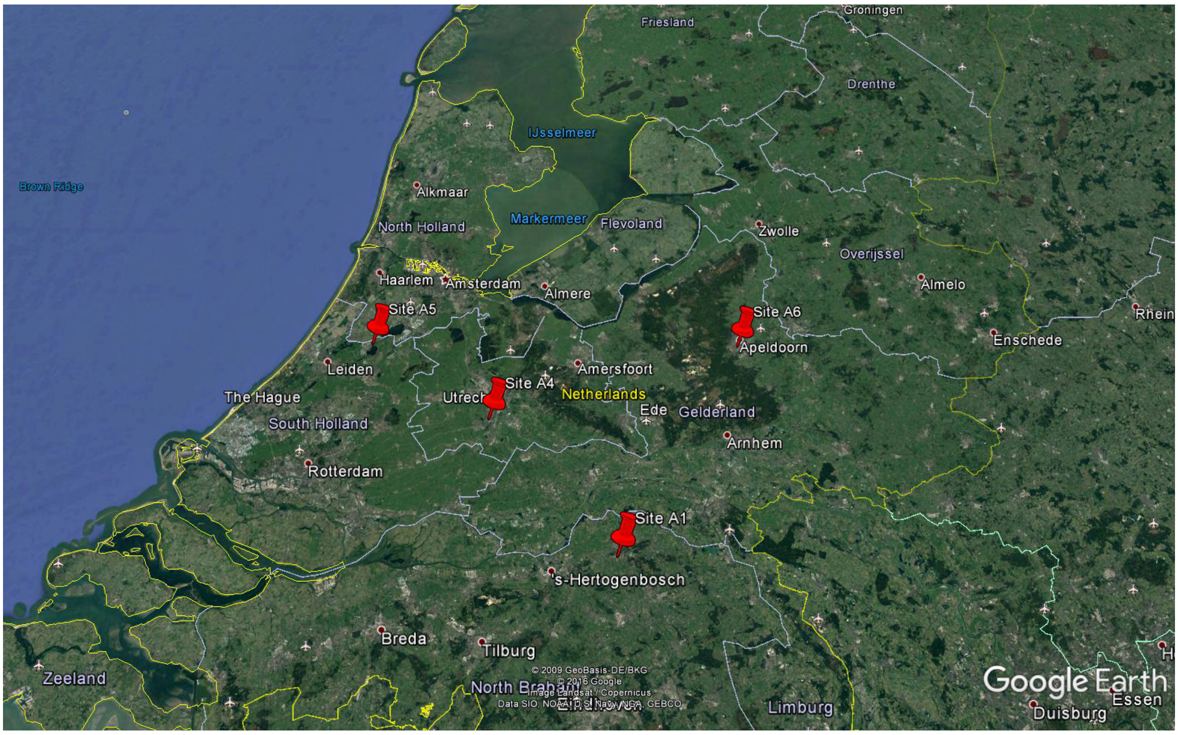


Figure 5.44: The European sites selected for the forecast. Credits: Google Earth.

The proposed PCA-Kalman based forecasting procedure is compared with four different benchmark methods, including a classical State space Kalman filter approach (KF), a autoregressive modified Grey box method (FGM) and Backpropagation artificial neural networks, in a standard implementation (BP) and in a PCA enhanced approach (PBP). Five error performance metrics are employed: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), Maximum Absolute Error (MXE) and Correlation coefficient (r^2), as denoted in equations (5.5), (5.6), (5.7), (5.8) and (5.9), respectively. The absence of relative or percentual error metrics is a consequence of the oftenly occurring "null production" days, which precludes the use of the MAPE and MPE metrics as these relative indicators are ill-defined when the reference approaches zero.

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(y[k] - \widehat{y}[k]\right)^2} \tag{5.5}$$

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|y[k] - \widehat{y}[k]\right| \tag{5.6}$$

$$MBE = \frac{1}{n}\sum_{k=1}^{n}\left(y[k] - \widehat{y}[k]\right) \tag{5.7}$$

$$MXE = \max_{k}\left(\left|y[k] - \widehat{y}[k]\right|\right) \tag{5.8}$$

$$r^2 = \frac{\mathrm{cov}\left(y, \widehat{y}\right)}{\sigma_y \sigma_{\widehat{y}}} = \frac{\sum_{k=1}^{n}\left(\widehat{y}[k] - \overline{y}\right)^2}{\sum_{k=1}^{n}\left(y[k] - \overline{y}\right)^2} \tag{5.9}$$

where $y[k]$ and $\widehat{y}[k]$ respectively denote the measured PV generation and forecasted PV generation for day k, $\overline{y}$ is the time series mean of the PV generation, σ_y and $\sigma_{\widehat{y}}$ the standard deviation from mean in the measurements and predictions.

Other relative indicators are possible, employing plant capacity or typical day generation as reference. However, they are intrisically biased towards the PV system optical and technical parameters, as capacity factors and spectral efficiency do change according to geographical location, instalation geometry, local climate, the type of photovoltaic cells and inverter arrangements employed. As such, absolute and unbiased error performance metrics are widely used when comparing different forecasting methodologies [108].

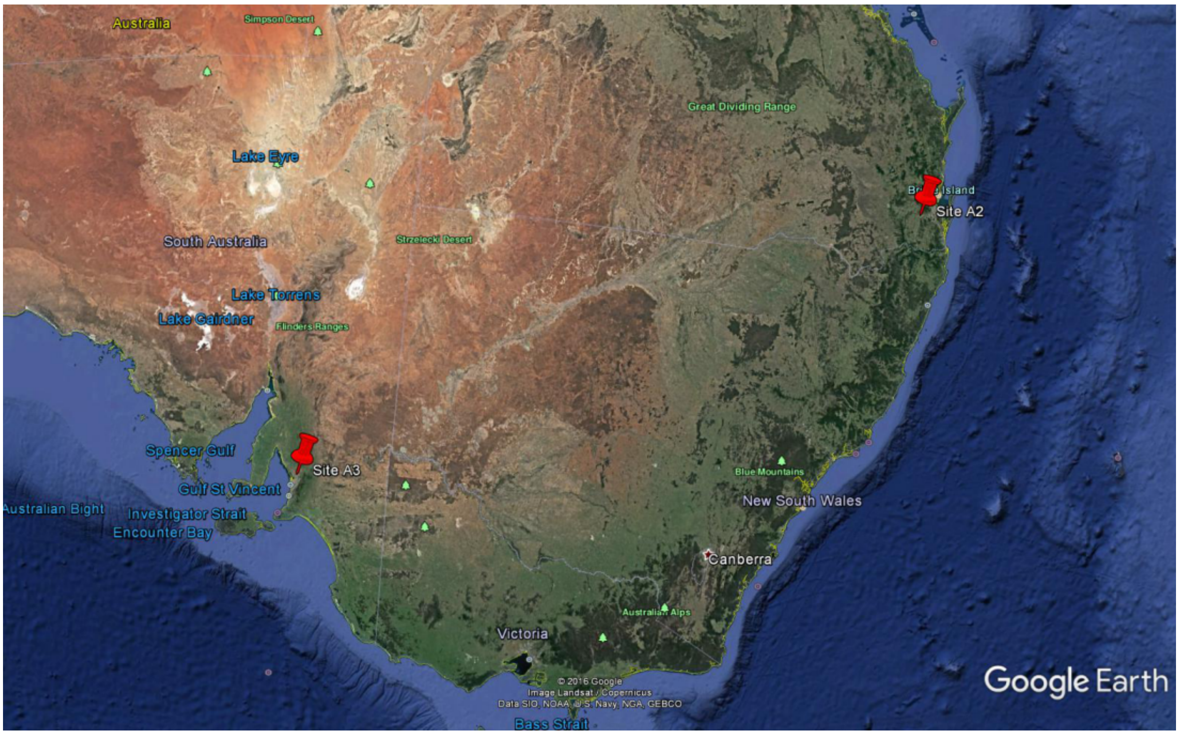


Figure 5.45: Location of the Australian sites selected for the forecast. Credits: Google Earth

The analysed PV systems are grouped in the seven “sites”, named from A1 to A7, regarding their geographical region, presence of other PV systems in a 10 kilometers radius and relative position with respect to weather stations. Their locations are pictured in Figures 5.44, 5.45 and 5.46. Sites A1, A4, A5 and A6 are located in Europe, A2 and A3 in Oceania and site A7 in North America. These sites present varying conditions for PV generation, ranging from semi-arid to subtropical climates, urban and rural landscapes, coastal and inland enviroments.

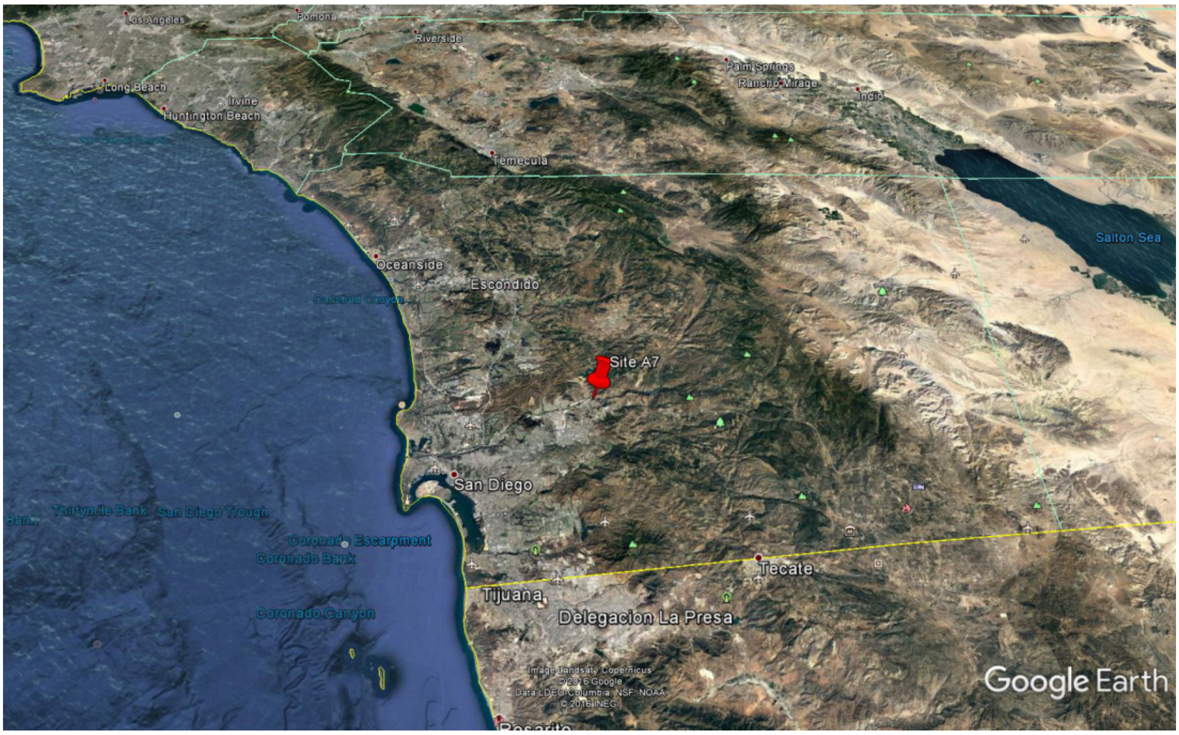


Figure 5.46: Location of the North American generation site selected for the forecast. Credits: Google Earth

The results for each site are presented in Subsections 5.2.1 to 5.2.7.

5.2.1 Site A1 - Oss region, Netherlands

This site is approximately located at coordinates 51.732N - 5.516E and contains four photovoltaic systems, as described in Table 5.44. Site A1 represents a light residential (suburban) enviroment, several kilometers inland and with a humid temperate climate without dry season. It is located ten kilometers south of Oss center, a dutch medium sized city. System capacities range from 0.625 to 7.1 kWp. System A1a is the smallest generator forecasted in this work.

Table 5.44: PV Systems in site A1

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A1a	7	155	35	0.625
A1b	9	180	45	7.100
A1c	8	225	1	2.500
A1d	7	156	45	1.560

As illustrated in Figure 5.47, six weather stations encircle the region and are used to provide both weather and solar resource parameters for the predictions. The weather stations are designated by their IATA codes, the closest being the EHVK airport.

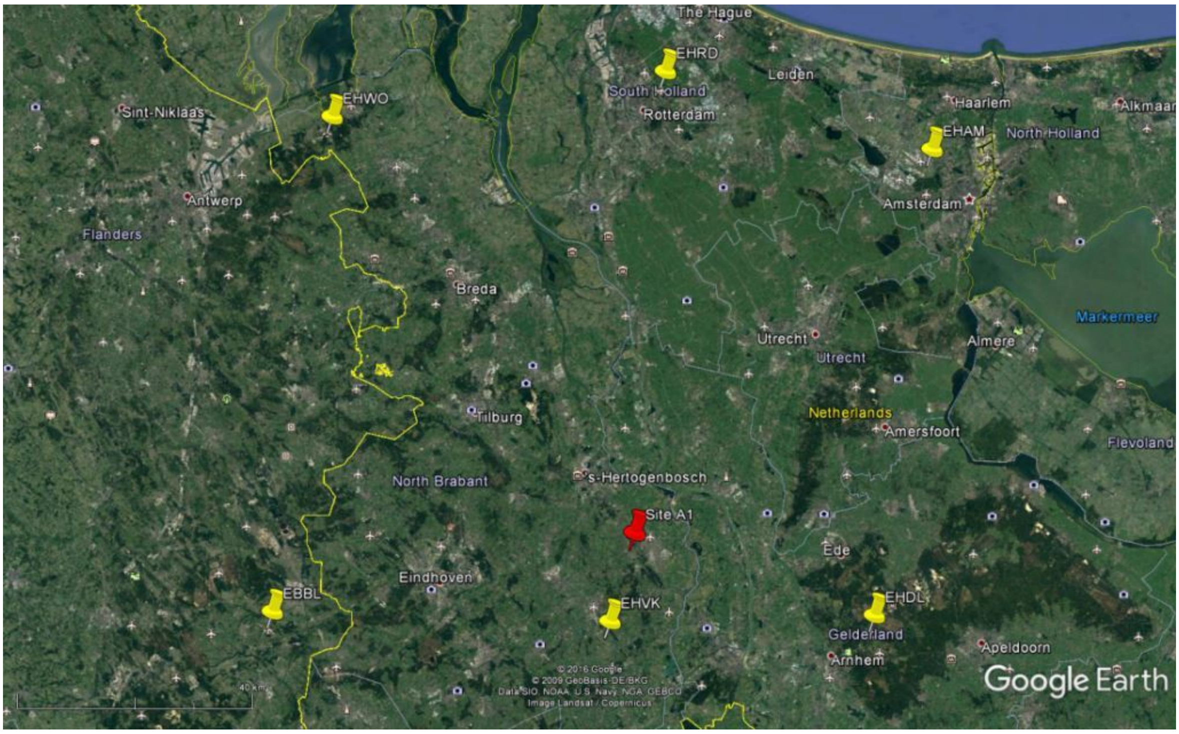


Figure 5.47: Site A1 and the six airport weather stations used for the forecast. Credits: Google Earth

5.2.1.1 Forecasting results, single weather station

Most forecasting methods seem to provide good forecasts to systems A1a, A1b and A1d. Errors are bigger for system A1c. Tables 5.45 and 5.46 list the error performance at different criteria for these systems.

Table 5.45: Results for Systems A1a and A1b, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A1a	RMSE	0,441	0,979	0,383	0,879	0,956
	MAE	0,351	0,757	0,301	0,688	0,744
	MBE	0,018	0,181	-0,005	0,064	-0,053
	MXE	1,380	3,349	1,460	2,700	3,036
	r^2	0,922	0,372	0,943	0,664	0,656
A1b	RMSE	5,413	11,389	4,724	8,671	10,259
	MAE	4,282	8,873	3,736	6,966	8,548
	MBE	1,014	4,596	0,881	-0,371	-0,905
	MXE	16,528	43,397	13,330	25,912	26,009
	r^2	0,925	0,535	0,942	0,729	0,660

Table 5.46: Results for Systems A1c and A1d, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A1c	RMSE	2,867	6,282	2,475	2,697	3,009
	MAE	2,288	3,334	1,953	2,166	2,432
	MBE	0,013	1,123	0,061	0,029	0,519
	MXE	12,606	102,686	10,573	11,158	10,278
	r^2	0,744	0,393	0,797	0,788	0,723
A1d	RMSE	0,520	4,741	0,584	1,978	1,949
	MAE	0,409	2,008	0,445	1,505	1,479
	MBE	-0,011	0,675	0,015	0,274	-0,026
	MXE	1,971	102,211	3,474	8,038	7,845
	r^2	0,978	0,331	0,974	0,696	0,698

RMSE-wise, the PCA-Kalman filter is better method for system A1d, but is slightly outperformed by the classic Kalman filter in the other systems.

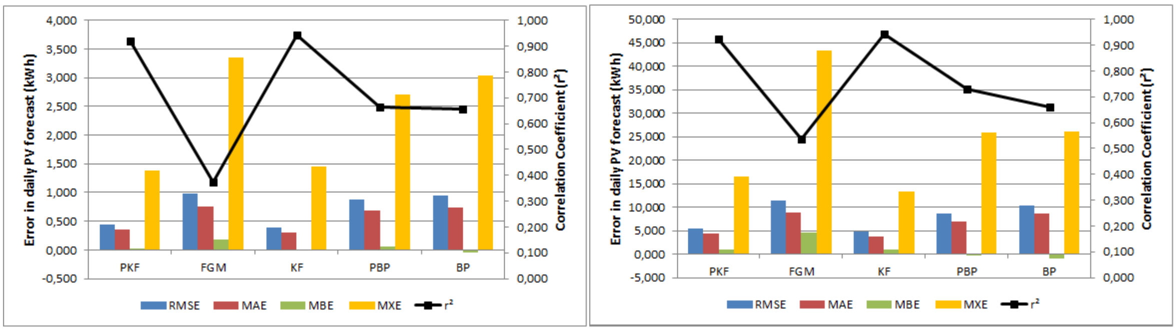


Figure 5.48: Error graphs for forecasts in Systems A1a and A1b, from left to right. Single station.

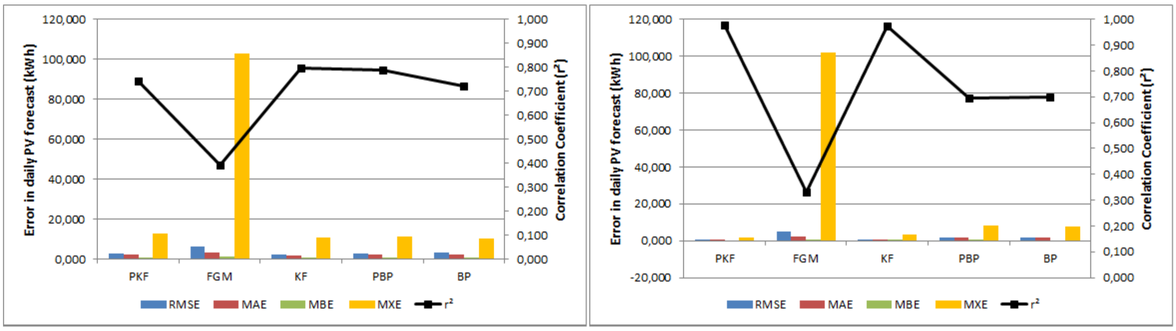


Figure 5.49: Error graphs for forecasts in Systems A1c and A1d, from left to right. Single station

5.2.1.2 Forecasting results, multiple weather stations

Using information from all avaliable stations, a much larger set of inputs for PV forecasting inputs are fed into the prediction methodologies. Table 5.47 lists the results for System A1a and A1b, Table 5.48 for A1c and A1d.

Table 5.47: Results for Systems A1a and A1b, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A1a	RMSE	0,396	0,979	0,661	0,923	0,964
	MAE	0,314	0,757	0,522	0,721	0,757
	MBE	-0,038	0,181	-0,055	0,031	0,013
	MXE	1,460	3,349	2,535	2,587	2,933
	r^2	0,942	0,372	0,865	0,638	0,658
A1b	RMSE	5,795	11,389	8,586	11,917	11,101
	MAE	4,189	8,873	6,690	9,359	8,320
	MBE	1,733	4,596	-0,429	-4,253	3,291
	MXE	17,852	43,397	22,781	29,586	42,021
	r^2	0,933	0,535	0,805	0,501	0,551

Table 5.48: Results for Systems A1c and A1d, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A1c	RMSE	2,450	6,282	2,885	2,666	3,686
	MAE	1,883	3,334	2,238	2,198	2,994
	MBE	0,000	1,123	0,106	-0,038	0,212
	MXE	11,806	102,686	18,423	9,243	11,317
	r^2	0,823	0,393	0,762	0,785	0,554
A1d	RMSE	0,501	4,741	1,585	2,088	2,047
	MAE	0,387	2,008	1,173	1,610	1,577
	MBE	0,001	0,675	0,004	0,176	0,066
	MXE	2,194	102,211	10,616	9,246	10,924
	r^2	0,980	0,331	0,810	0,657	0,655

Using multiple weather stations, the PCA-Kalman provides the lowest RMSE overall at all systems.

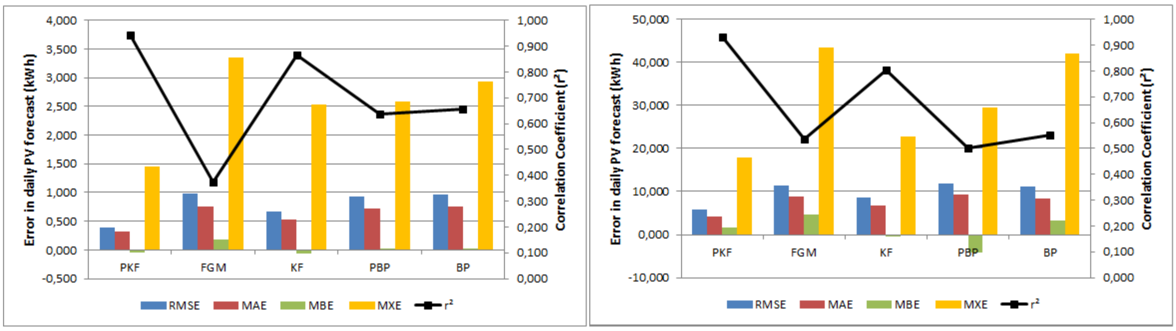


Figure 5.50: Error graphs for forecasts in Systems A1a and A1b, from left to right. Multiple stations.

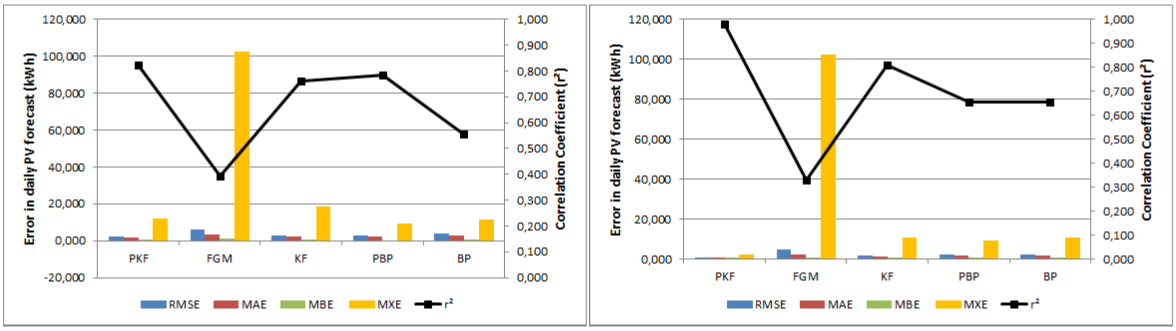


Figure 5.51: Error graphs for forecasts in Systems A1c and A1d, from left to right. Multiple stations.

The predictions provided by the PCA-Kalman algorithm are compared to the measured PV generation values in figures 5.52,5.53, 5.54 and 5.55::

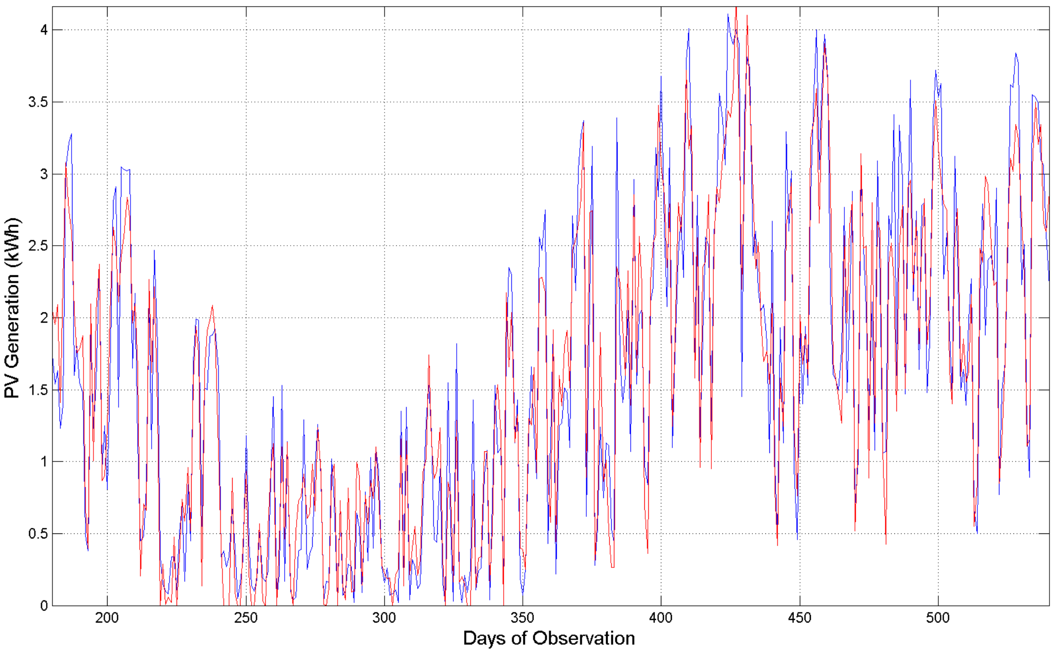


Figure 5.52: Prediction (red line) plotted against the measured PV Generation (blue line) in site A1a over 360 days of observation.

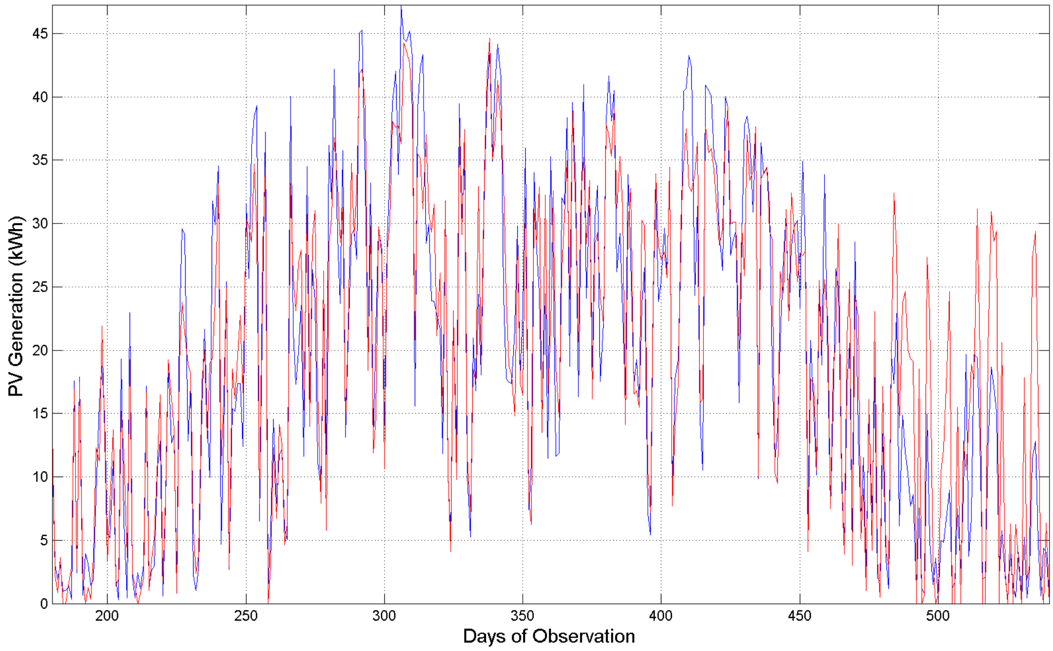


Figure 5.53: Prediction (red line) plotted against the measured PV Generation (blue line) in site A1b over 360 days of observation.

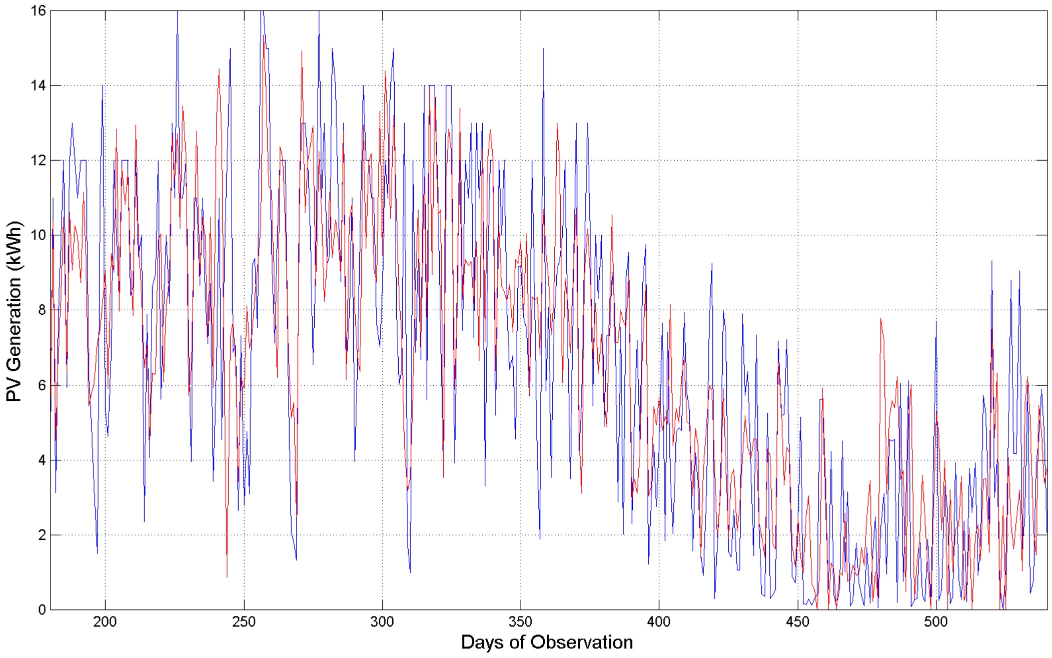


Figure 5.54: Prediction (red line) plotted against the measured PV Generation (blue line) in site A1c over 360 days of observation.

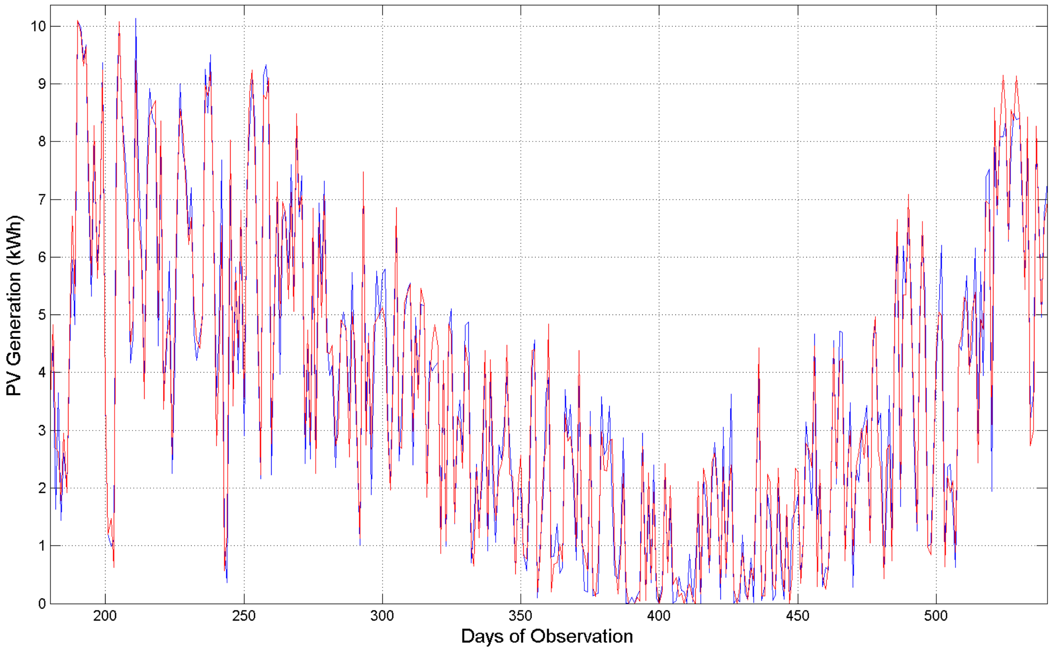


Figure 5.55: Prediction (red line) plotted against the measured PV Generation (blue line) in site A1d over 360 days of observation.

5.2.2 Site A2 - Queensland, Australia

This site is located in the Australian state of Queensland, at coordinates 27.61S - 152.74E. Site A2 represents a residential (urban) enviroment, several kilometers inland and with a humid subtropical climate without dry season. It is located close to Ipswich center, a city with 200.000 citizens southwest of Brisbane metropolitan area. System capacities range from 6 to 7 kWp.

Table 5.49: PV Systems in site A2

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A2a	25	0	15	6.080
A2b	25	357	30	7.000

As illustrated in Figure 5.56, two weather stations in the neighboor regions are used to provide both weather and solar resource parameters for the predictions. The weather stations are designated by their IATA codes, the closest being the YAMB airfield.

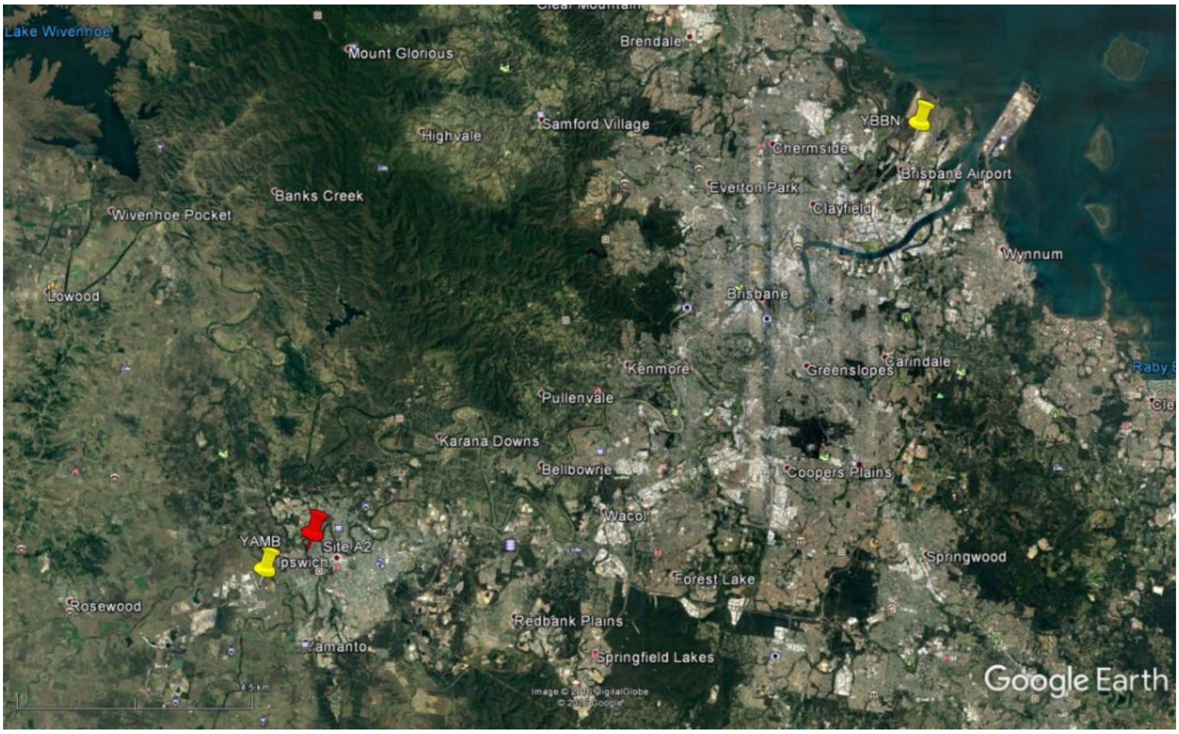


Figure 5.56: Site A2 and the two airport weather stations used for the forecast. Credits: Google Earth

5.2.2.1 Forecasting results, single weather station

Most forecasting methods seem to provide good forecasts to systems A2a, A2b. Tables 5.50 and 5.51 list the error performance at different criteria for these systems.

Table 5.50: Results for Systems A2a and A2b, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A2a	RMSE	3,957	10,950	4,385	7,942	8,093
	MAE	3,062	8,371	3,346	6,288	6,261
	MBE	-0,033	1,484	-0,027	-0,731	0,065
	MXE	16,810	40,684	22,361	27,907	29,916
	r^2	0,922	0,378	0,902	0,655	0,635
A2b	RMSE	3,088	9,539	4,160	6,945	6,294
	MAE	2,283	7,210	2,982	5,477	4,732
	MBE	0,072	0,495	0,030	-1,638	-1,030
	MXE	19,820	38,101	29,523	27,830	30,465
	r^2	0,937	0,434	0,874	0,636	0,707

PCA-Kalman offers the lowest RMSE and MXE, followed by the classic Kalman.

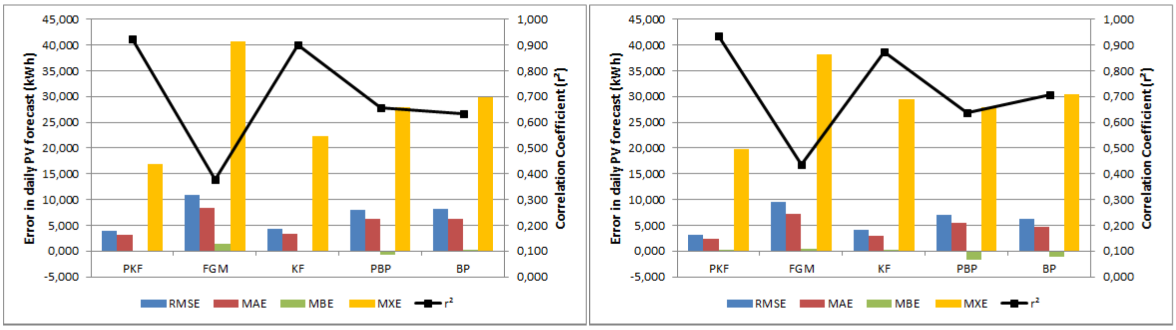


Figure 5.57: Error graphs for forecasts in Systems A2a and A2b, from left to right. Single station.

5.2.2.2 Forecasting results, multiple weather stations

Using information from all avaliable stations, a much larger set of inputs for PV forecasting inputs are fed into the prediction methodologies. Table 5.51 lists the results for System A2a and A2b.

Table 5.51: Results for Systems A2a and A2b, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A2a	RMSE	3,529	10,950	5,620	8,984	9,037
	MAE	2,704	8,371	4,140	6,772	7,066
	MBE	-0,011	1,484	-0,011	0,447	-0,145
	MXE	18,496	40,684	45,717	36,373	34,239
	r^2	0,935	0,378	0,838	0,573	0,482
A2b	RMSE	2,986	9,539	4,984	8,073	7,287
	MAE	2,152	7,210	3,606	6,173	5,657
	MBE	0,052	0,495	0,183	-0,536	-1,294
	MXE	23,331	38,101	32,874	36,355	30,764
	r^2	0,940	0,434	0,848	0,547	0,578

PCA-Kalman improves its results over the single weather station case, and offers the lowest RMSE overall, followed by the classic Kalman.

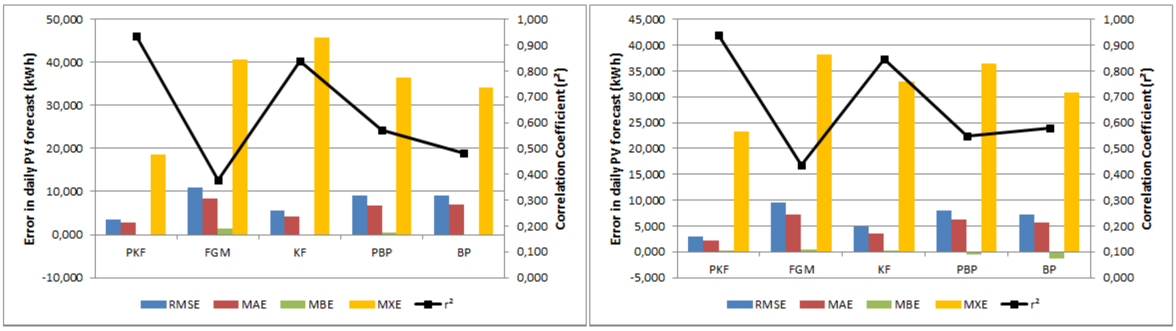


Figure 5.58: Error graphs for forecasts in Systems A2a and A2b, from left to right. Multiple stations.

The predictions provided by the PCA-Kalman algorithm are compared to the measured PV generation values in figures 5.59 and 5.60.

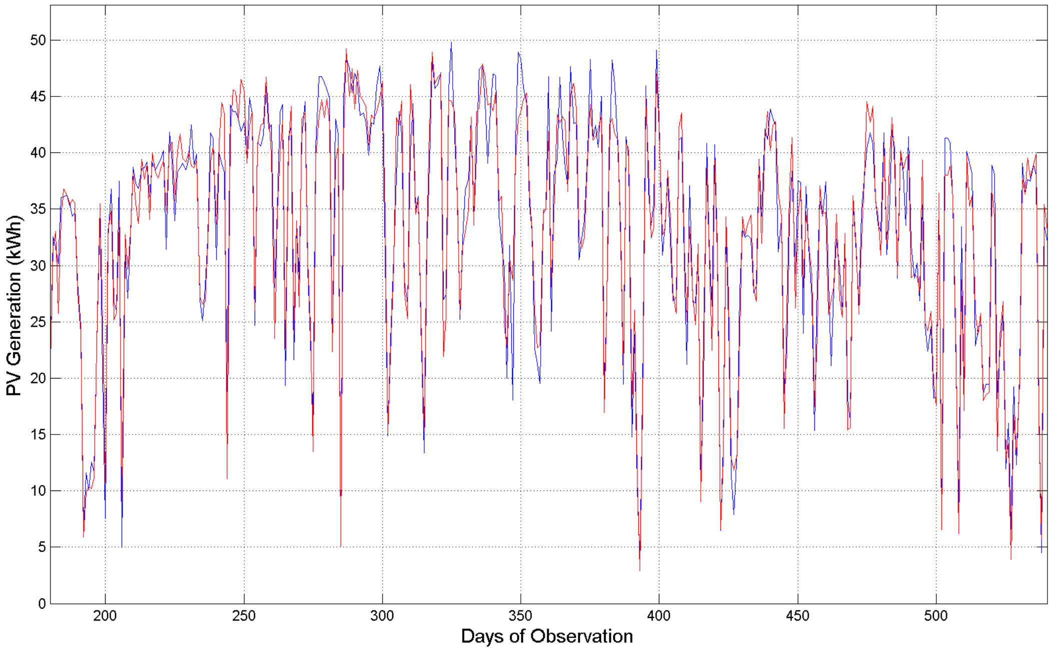


Figure 5.59: Prediction (red line) plotted against the measured PV Generation (blue line) in site A2a over 360 days of observation.

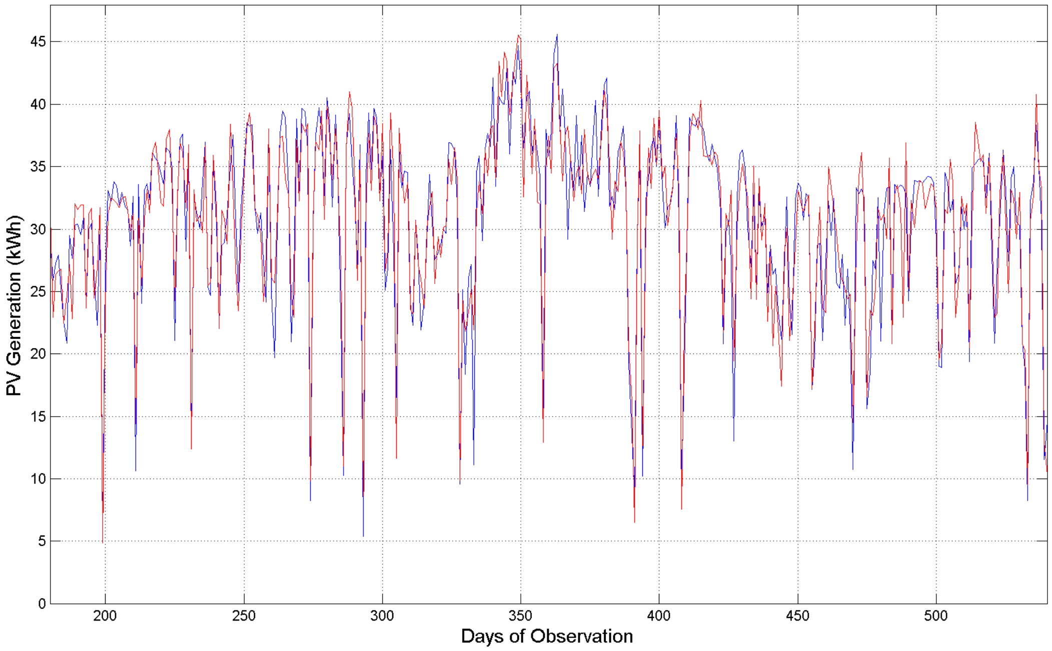


Figure 5.60: Prediction (red line) plotted against the measured PV Generation (blue line) in site A2b over 360 days of observation.

5.2.3 Site A3 - South Australia, Australia

This site is located in South Australia, at coordinates 34.68S - 138.65E, just north of Adelaide metropolitan area. Site A3 represents a rural-suburban enviroment, close to the coast and with a mediterranean climate. It is located close to Blakeview, a city with 4.000 citizens. System capacities in the three analysed generators range from 3 to 6.2 kWp.

Table 5.52: PV Systems in site A3

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A3a	25	45	22	5.280
A3b	25	315	20	3.055
A3c	26	90	24	6.200

As illustrated in Figure 5.61, a sole weather station is used to provide both weather and solar resource parameters for the predictions. The weather stations is designated by its IATA code YPAD, which is Adelaide's international airport.

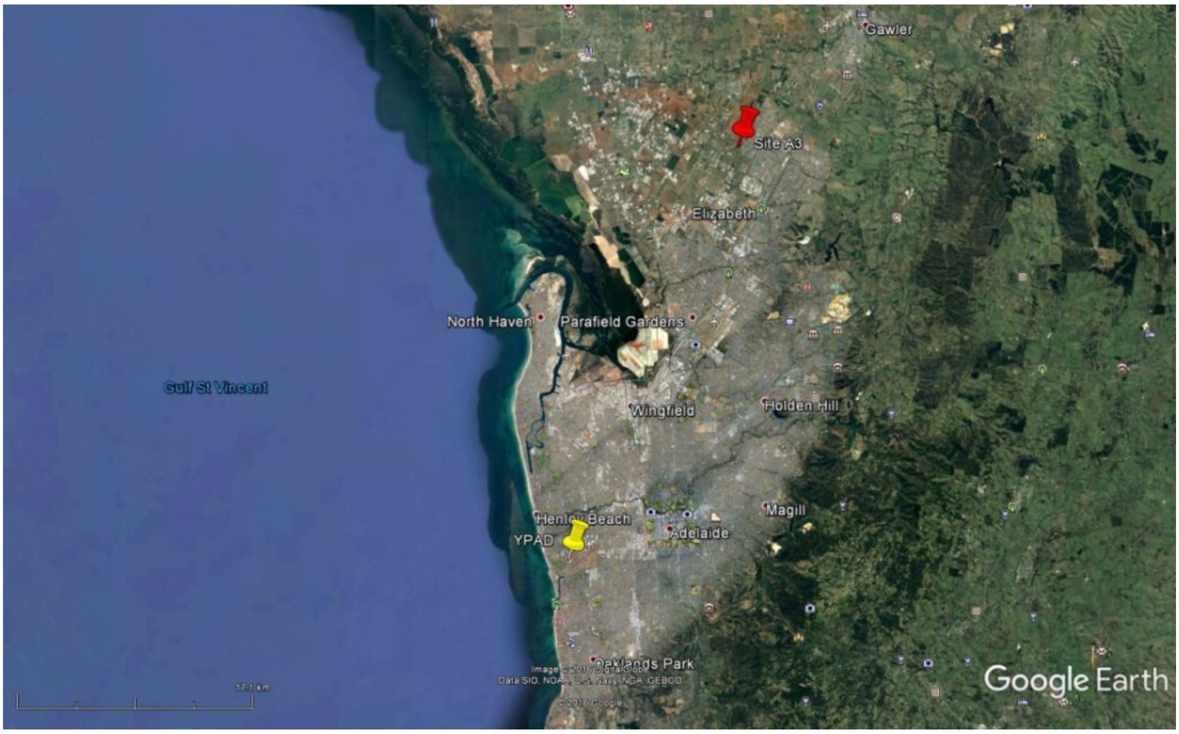


Figure 5.61: Site A3 and the airport weather station used for the forecast. Credits: Google Earth

5.2.3.1 Forecasting results, single weather station

Using information from a single weather stations, performances are better in systems A3b and A3c than in A3a. Table 5.53 lists the results.

Table 5.53: Results for Systems A3a, A3b and A3c, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A3a	RMSE	5,750	8,175	5,339	6,992	7,608
	MAE	4,503	6,090	4,158	5,687	6,330
	MBE	0,001	1,252	0,033	0,335	0,366
	MXE	23,448	35,167	22,279	24,275	27,799
	r^2	0,837	0,703	0,860	0,775	0,733
A3b	RMSE	1,666	4,781	1,689	3,606	3,635
	MAE	1,259	3,659	1,218	2,920	2,913
	MBE	0,008	0,601	-0,132	-0,187	-0,034
	MXE	7,545	17,640	11,624	13,080	15,061
	r^2	0,955	0,635	0,941	0,777	0,768
A3c	RMSE	2,238	7,869	3,001	6,351	6,622
	MAE	1,721	5,923	2,136	5,095	5,127
	MBE	0,021	0,853	-0,310	0,324	0,325
	MXE	11,928	29,937	17,707	24,632	30,885
	r^2	0,969	0,728	0,955	0,806	0,791

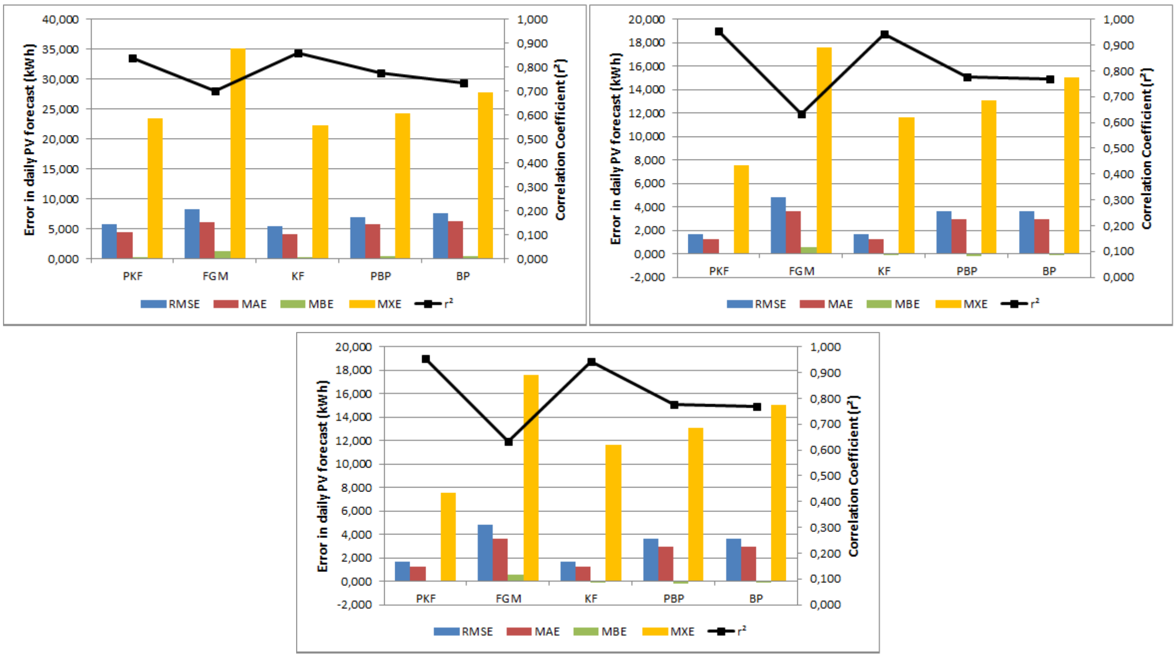


Figure 5.62: Error graphs for forecasts in Systems A3a (top left), A3b (top right) and A3c (bottom). Single station.

PCA Kalman performs slightly better in systems A3b and A3b, but is outperformed in system A3a. The predictions provided by this algorithm are compared to the measured PV generation values in figures 5.63,5.64, 5.65 and 5.55::

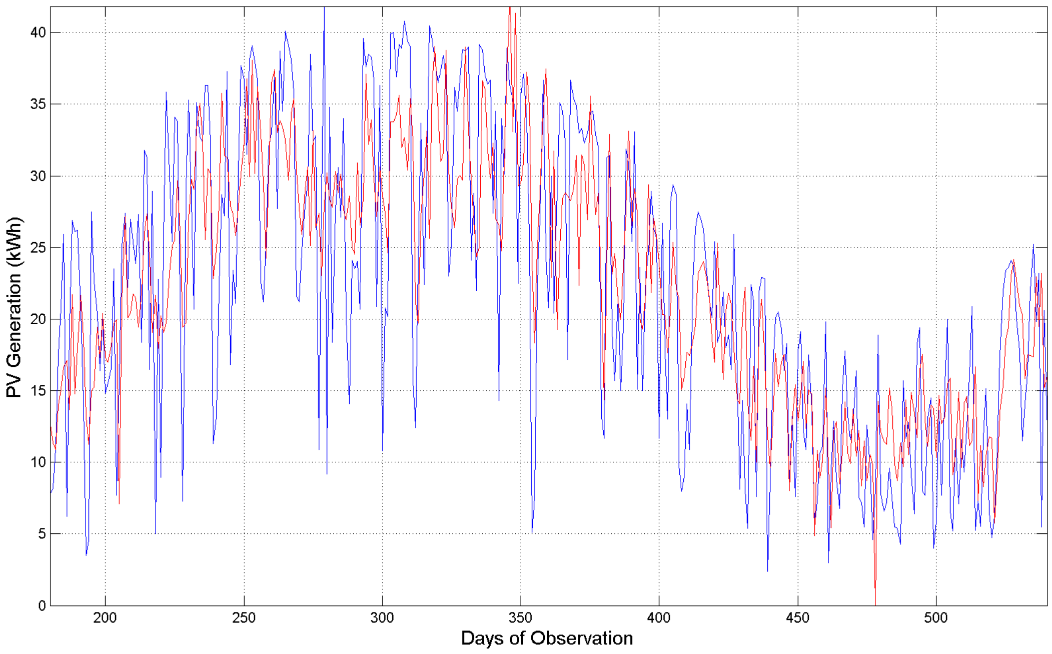


Figure 5.63: Prediction (red line) plotted against the measured PV Generation (blue line) in site A3a over 360 days of observation.

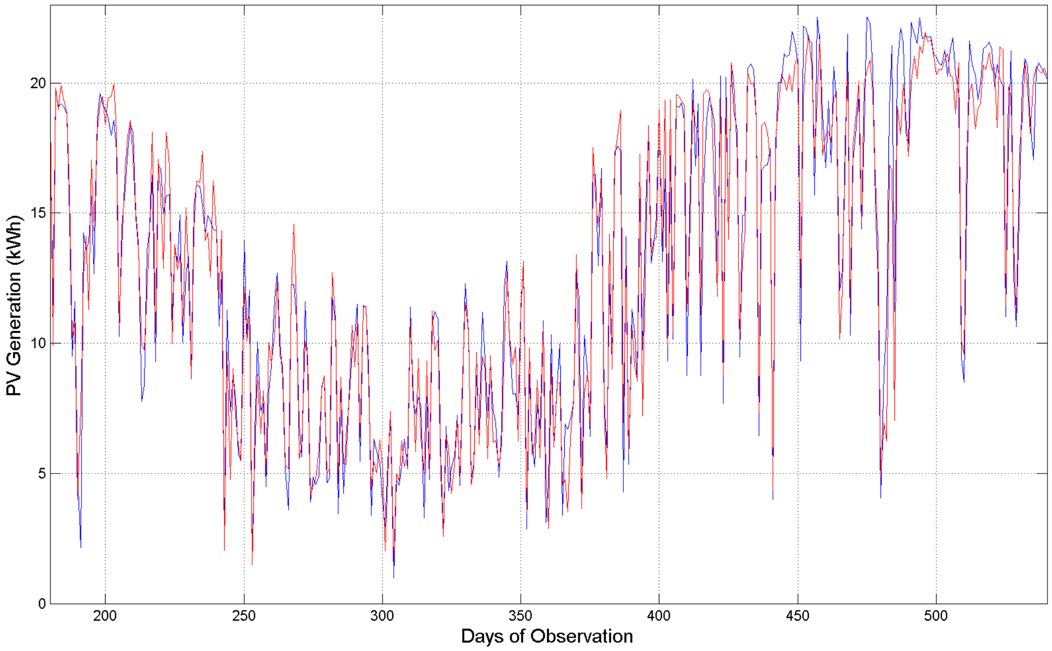


Figure 5.64: Prediction (red line) plotted against the measured PV Generation (blue line) in site A3b over 360 days of observation.

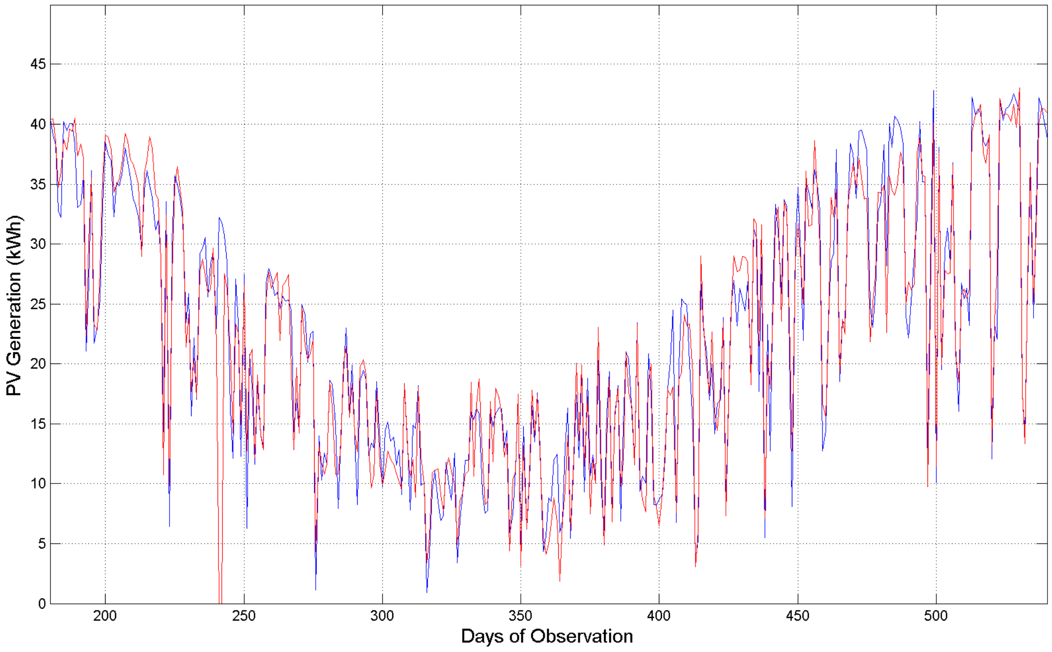


Figure 5.65: Prediction (red line) plotted against the measured PV Generation (blue line) in site A3c over 360 days of observation.

5.2.4 Site A4 - Utrecht region, Netherlands

This site is located in the Randstad conurbation, central Netherlands, at coordinates 52.03N - 5.08E, circa ten kilometers south of Utrecht center. Site A4 represents an urban enviroment, close to the coast and with a oceanic climate without dry season.The sole system in analysed in this site has 4.62 kWp and is a rooftop generator in a residential building.

Table 5.54: PV Systems in site A4

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A4a	3	0	8	4.620

As illustrated in Figure 5.66, a sole weather station is used to provide both weather and solar resource parameters for the predictions. The weather stations is designated by its IATA code YPAD, which is Adelaide's international airport.

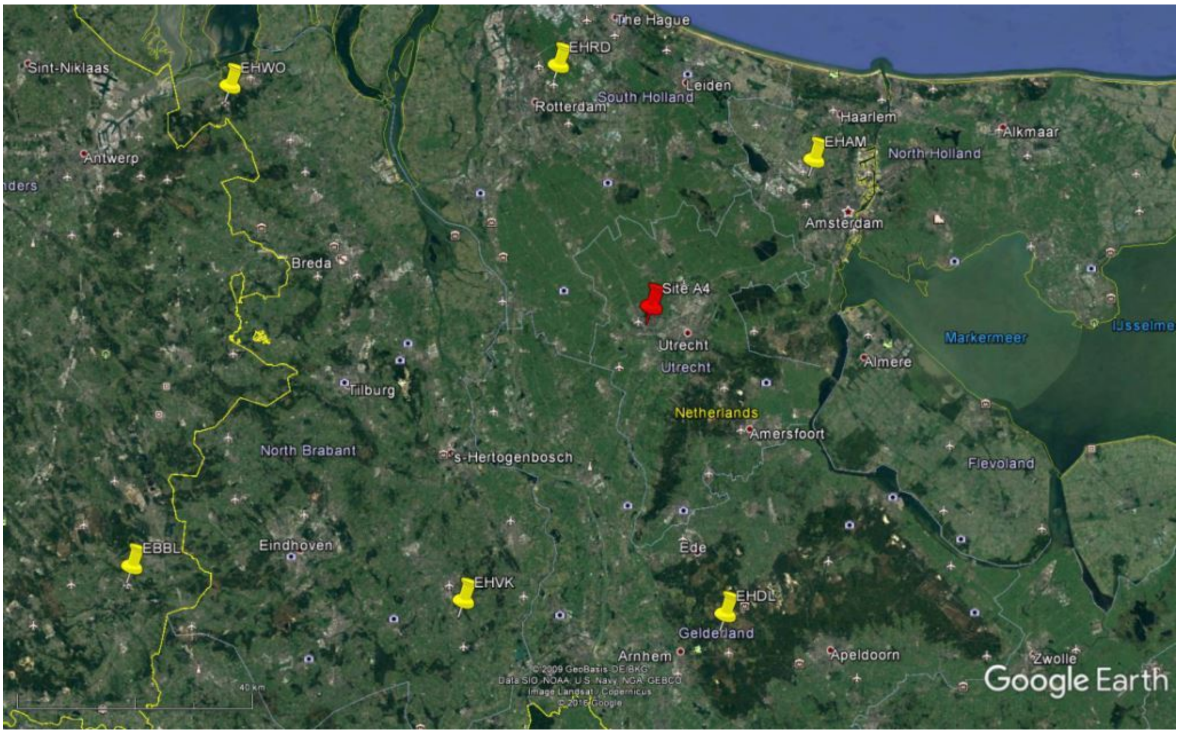


Figure 5.66: Site A4 and the six airport weather stations used for the forecast. Credits: Google Earth

5.2.4.1 Forecasting results, single weather station

Using information from the nearest station, the different methods are applied to forecast the PV generation. Table 5.55 lists the results for system A4a .

Table 5.55: Results for System A4a, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A4a	RMSE	2,583	5,248	1,925	3,862	4,381
	MAE	1,888	3,765	1,401	2,930	3,273
	MBE	-0,020	0,583	-0,215	-0,131	0,041
	MXE	10,898	21,162	10,756	15,451	17,953
	r^2	0,947	0,797	0,969	0,881	0,850

The Kalman filter slightly outperforms the PCA-Kalman method. ANN methods also perform close to the linear filters.

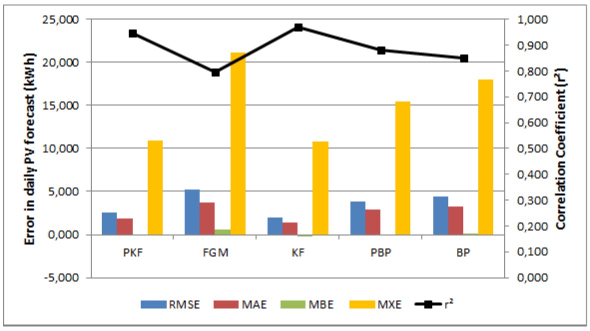


Figure 5.67: Error graphs for forecasts in System A4a. Single station.

5.2.4.2 Forecasting results, multiple weather stations

Using information from all six stations, a much larger set of inputs for PV forecasting inputs are fed into the prediction methodologies. Table 5.56 lists the results for system A4a .

Table 5.56: Results for System A4a, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A4a	RMSE	1,257	5,248	3,653	5,370	5,492
	MAE	0,968	3,765	2,653	4,111	3,946
	MBE	0,074	0,583	0,157	-0,622	0,008
	MXE	5,556	21,162	15,376	19,749	21,658
	r^2	0,984	0,797	0,902	0,753	0,774

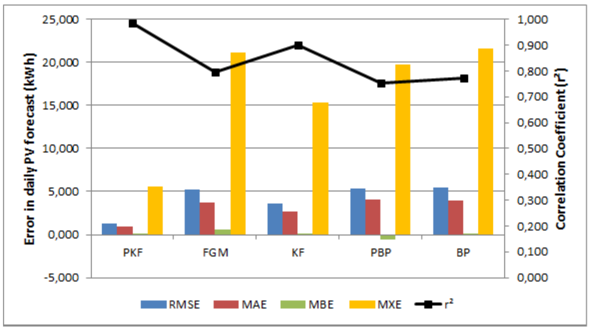


Figure 5.68: Error graphs for forecasts in System A4a. Multiple stations.

The forecasts provided by the PCA-Kalman algorithm with multiple weather stations are the most accurate. A comparison between predictions and the measured PV generation values is shown in figure 5.69.

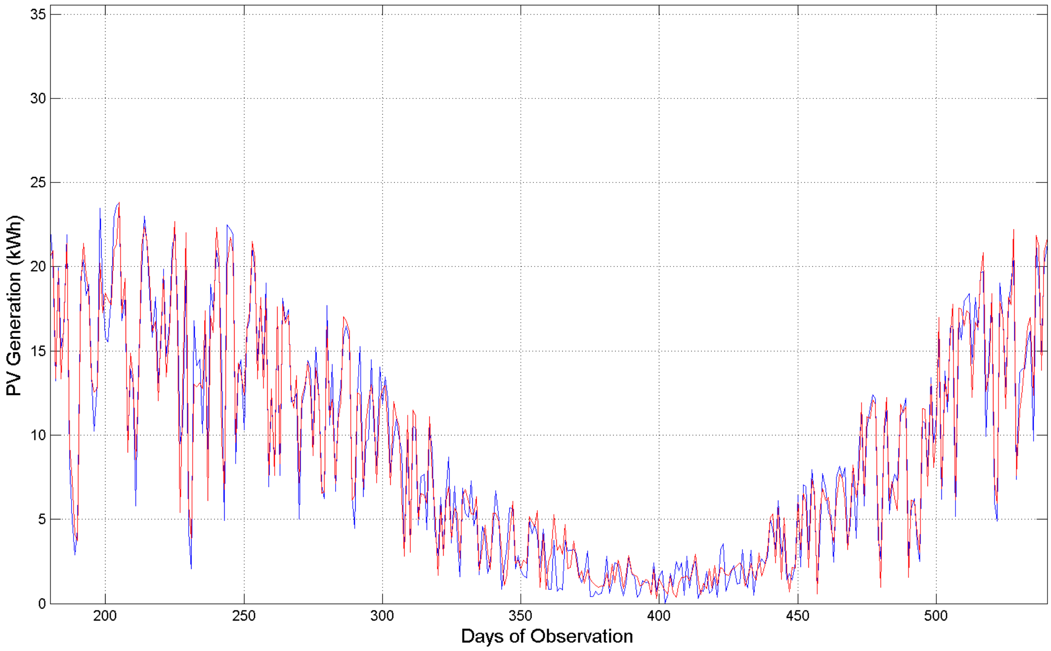


Figure 5.69: Prediction (red line) plotted against the measured PV Generation (blue line) in site A4a over 360 days of observation.

5.2.5 Site A5 - Amsterdam region, Netherlands

Site A5 is located just south of Schipol airport, close to Amsterdam region, at coordinates 52.03N - 5.08E, circa ten kilometers south of Utrecht center. Similarly to A4, site A5 represents a suburban enviroment, very close to the coast and with a humid oceanic climate, without dry season.The sole system in analysed in this site has 3.68 kWp and is a rooftop generator in a residential building.

Table 5.57: PV Systems in site A5

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A5a	-4	220	45	3.680

As illustrated in Figure 5.70, six weather stations is used to provide both weather and solar resource parameters for the predictions. The weather stations are designated by its IATA code, the closest being Schipol's airport EHAM.

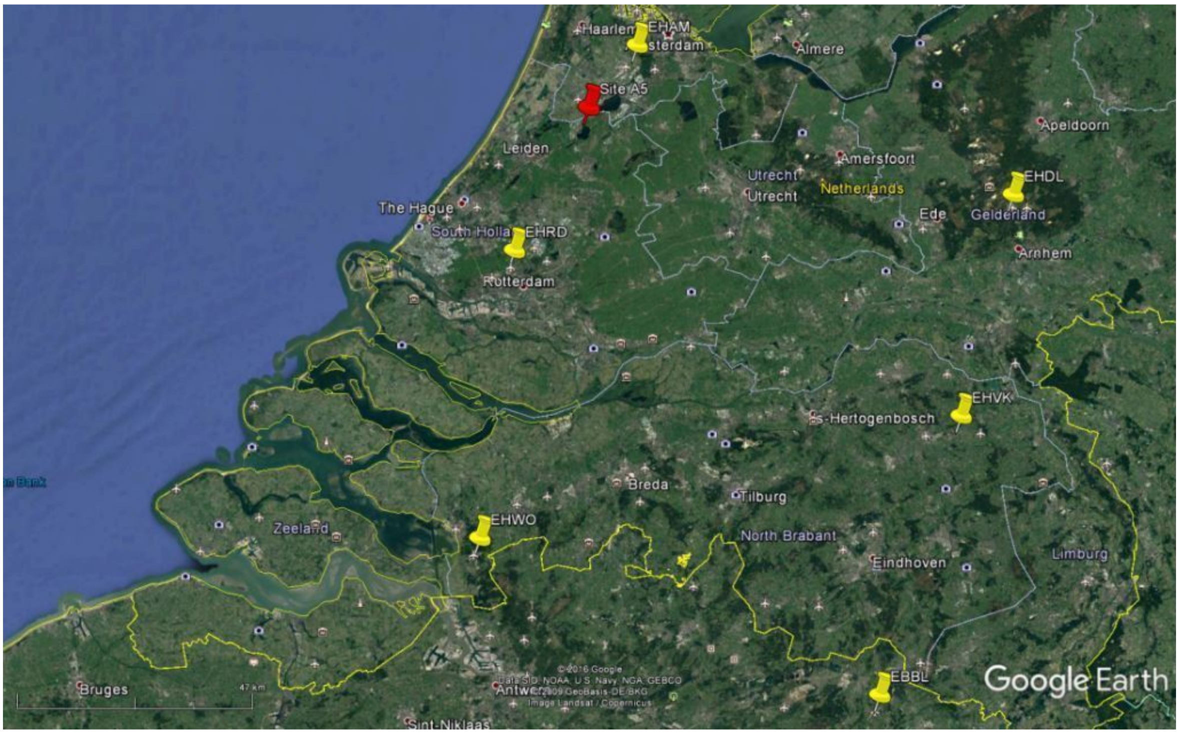


Figure 5.70: Site A5 and the six airport weather stations used for the forecast. Credits: Google Earth

5.2.5.1 Forecasting results, single weather station

Using information from the nearest station, the algorithms provide the performance summarized in Table ,5.58, which lists the results for system A5a .

Table 5.58: Results for System A5a, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A5a	RMSE	3,832	7,582	3,451	3,673	4,064
	MAE	2,788	4,412	2,448	2,823	3,242
	MBE	-0,150	1,385	-0,038	-0,050	0,521
	MXE	16,830	126,610	16,364	12,452	13,179
	r^2	0,863	0,608	0,890	0,876	0,846

Except for the Grey autoregressive model, the methodologies perform very similar at most criteria. PCA-BP has a advantage in maximum error, while classic Kalman has a very slightly better RMSE performance.

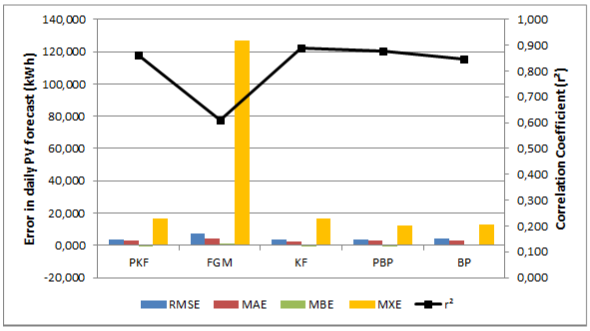


Figure 5.71: Error graphs for forecasts in System A5a. Single station.

5.2.5.2 Forecasting results, multiple weather stations

Using information from all six stations, the prediction methodologies are performed with a larger set of inputs. Table 5.59 lists the results for system A4a .

Table 5.59: Results for System A5a, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A5a	RMSE	3,508	7,582	3,678	4,597	4,779
	MAE	2,481	4,412	2,656	3,630	3,736
	MBE	-0,123	1,385	0,006	-0,131	0,135
	MXE	18,496	126,610	15,541	15,943	17,436
	r^2	0,886	0,608	0,874	0,785	0,765

Employing more information, the PCA-Kalman outperforms the other methods in RMSE criteria, but still does not achieve better performance than the Kalman filter with the single station inputs.

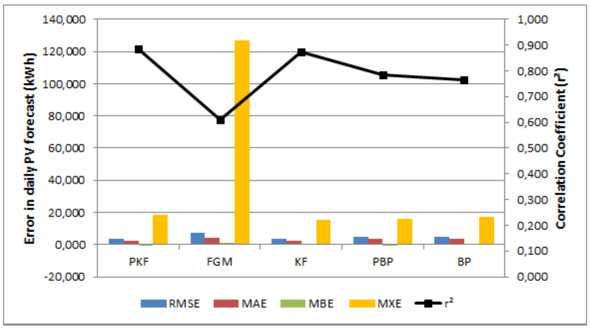


Figure 5.72: Error graphs for forecasts in System A5a. Multiple stations.

A comparison between the PCA-Kalman forecasts and the measured PV generation values is shown in figure 5.73.

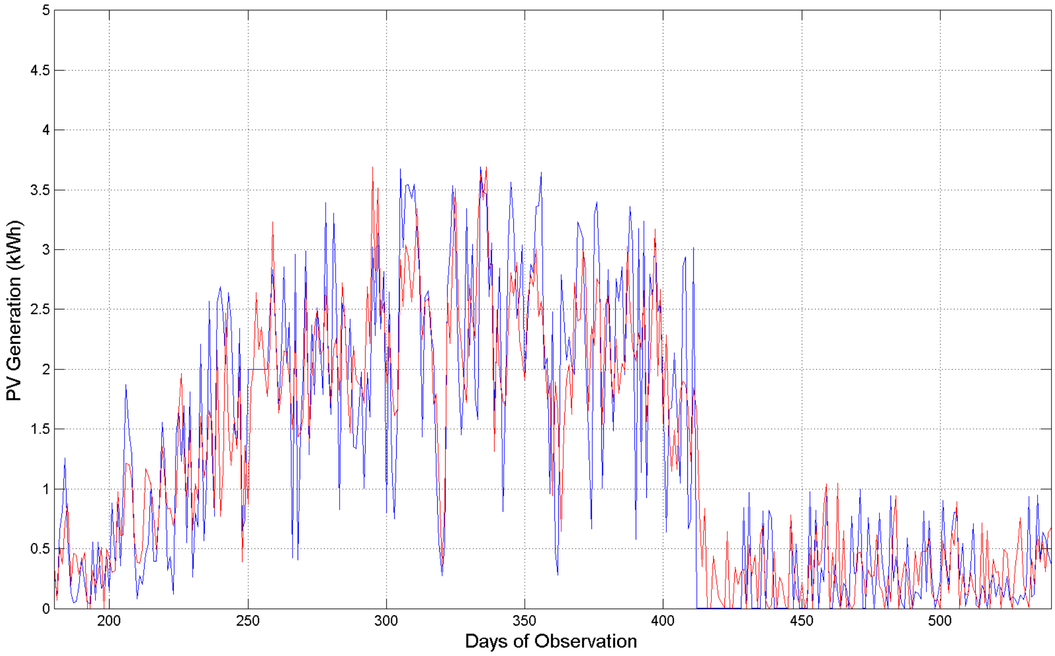


Figure 5.73: Prediction (red line) plotted against the measured PV Generation (blue line) in site A5a over 360 days of observation.

5.2.6 Site A6 - Apeldoorn region, Netherlands

This site is located in Apeldoorn region, at coordinates 52.2N - 5.96E, near Apeldoorn's city center, a medium sized dutch city. Site A6 represents a light residential urban enviroment, several kilometers inland and with a humid temperate climate, without dry season.The two systems analysed in this site have capacities ranging from 1.44 to 3.64 kWp.

Table 5.60: PV Systems in site A6

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A6a	20	180	34	3.640
A6b	22	200	36	1.440

As illustrated in Figure 5.74, six weather stations is used to provide both weather and solar resource parameters for the predictions. The weather stations are designated by its IATA code, the closest being EHDL.

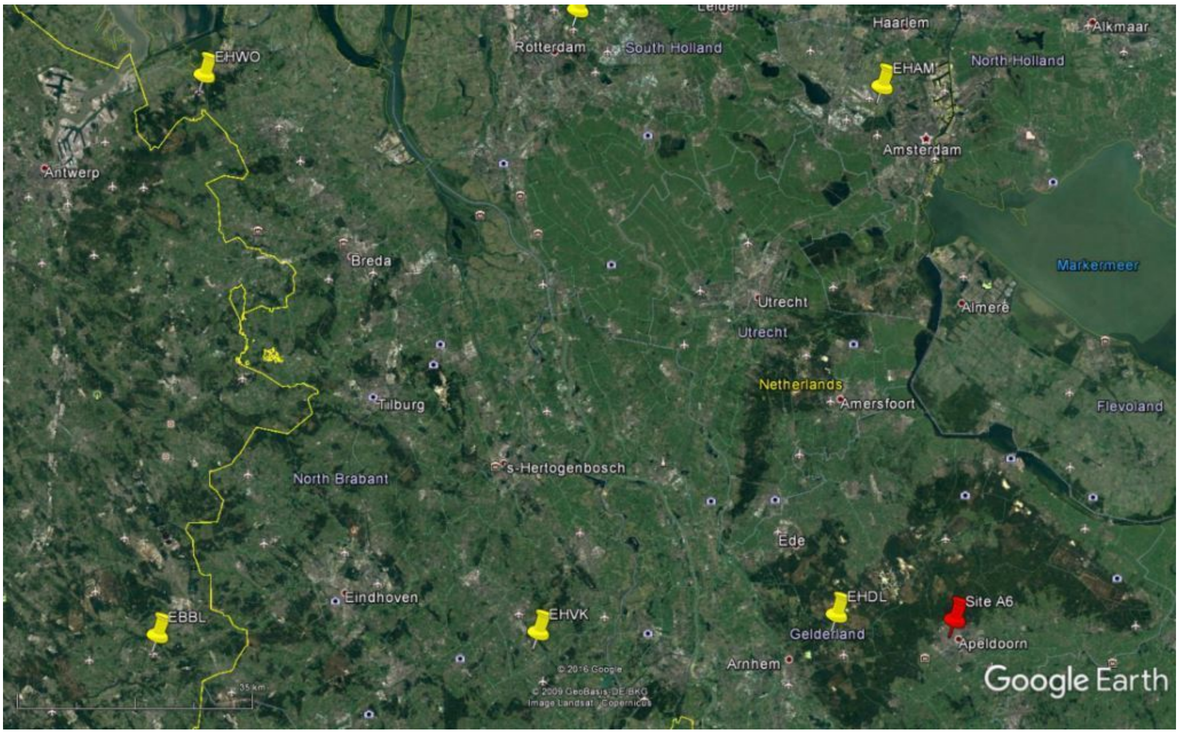


Figure 5.74: Site A6 and the six airport weather stations used for the forecast. Credits: Google Earth

5.2.6.1 Forecasting results, single weather station

Using information from the nearest station, the different methods are applied to forecast the PV generation. Table 5.61 lists the results for system A6a and A6b .

Table 5.61: Results for Systems A6a and A6b,single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A6a	RMSE	2,743	6,117	4,242	4,394	5,670
	MAE	2,002	4,424	3,092	3,405	4,532
	MBE	-0,119	1,177	0,255	0,030	0,624
	MXE	12,252	24,282	25,595	13,954	20,217
	r^2	0,889	0,529	0,724	0,717	0,523
A6b	RMSE	0,747	1,356	0,765	1,020	1,058
	MAE	0,498	0,926	0,522	0,798	0,806
	MBE	0,084	0,202	0,089	-0,030	0,007
	MXE	4,114	5,440	3,889	3,614	4,290
	r^2	0,940	0,769	0,941	0,858	0,845

PCA-Kalman outperforms the other methods, achieving the lowest RMSE.

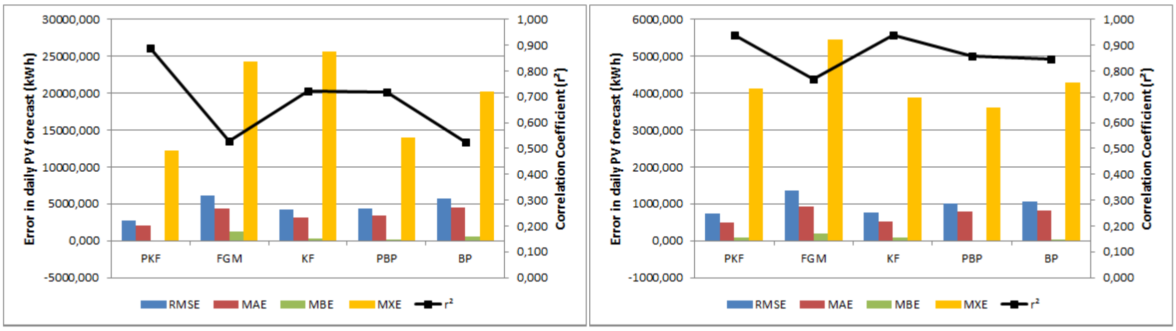


Figure 5.75: Error graphs for forecasts in Systems A6a and A6b, from left to right. Single station.

5.2.6.2 Forecasting results, multiple weather stations

Using information from all six stations, the prediction methodologies are performed with a larger set of inputs. Table 5.62 lists the results for system A6a and A6b.

Table 5.62: Results for Systems A6a and A6b, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A6a	RMSE	2,743	6,117	4,242	4,610	5,414
	MAE	2,002	4,424	3,092	3,619	4,101
	MBE	-0,119	1,177	0,255	0,033	-0,181
	MXE	12,252	24,282	25,595	14,850	21,007
	r^2	0,889	0,529	0,724	0,678	0,583
A6b	RMSE	0,725	1,356	1,407	1,080	1,505
	MAE	0,481	0,926	0,930	0,854	1,137
	MBE	0,111	0,202	0,211	0,121	0,208
	MXE	3,730	5,440	8,997	3,683	6,480
	r^2	0,944	0,769	0,799	0,834	0,716

PCA-Kalman outperforms the other methods, offering slightly improved performance over the single station case.

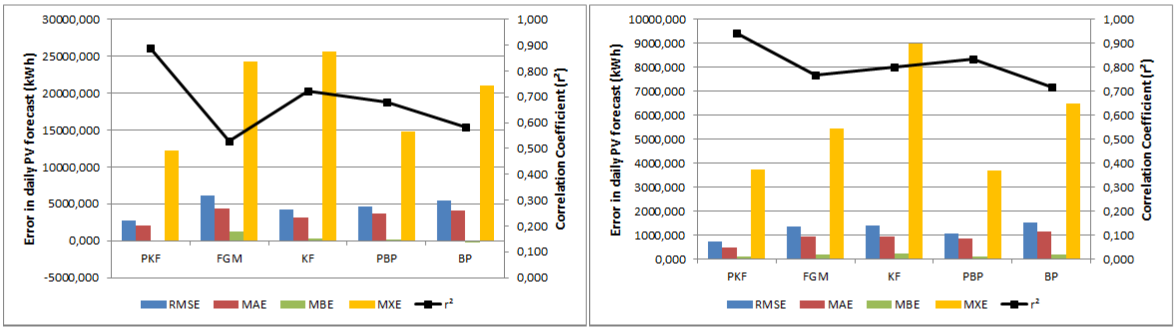


Figure 5.76: Error graphs for forecasts in Systems A6a and A6b, from left to right. Multiple stations.

The predictions provided by this algorithm are compared to the measured PV generation values in figures 5.77 and 5.78:

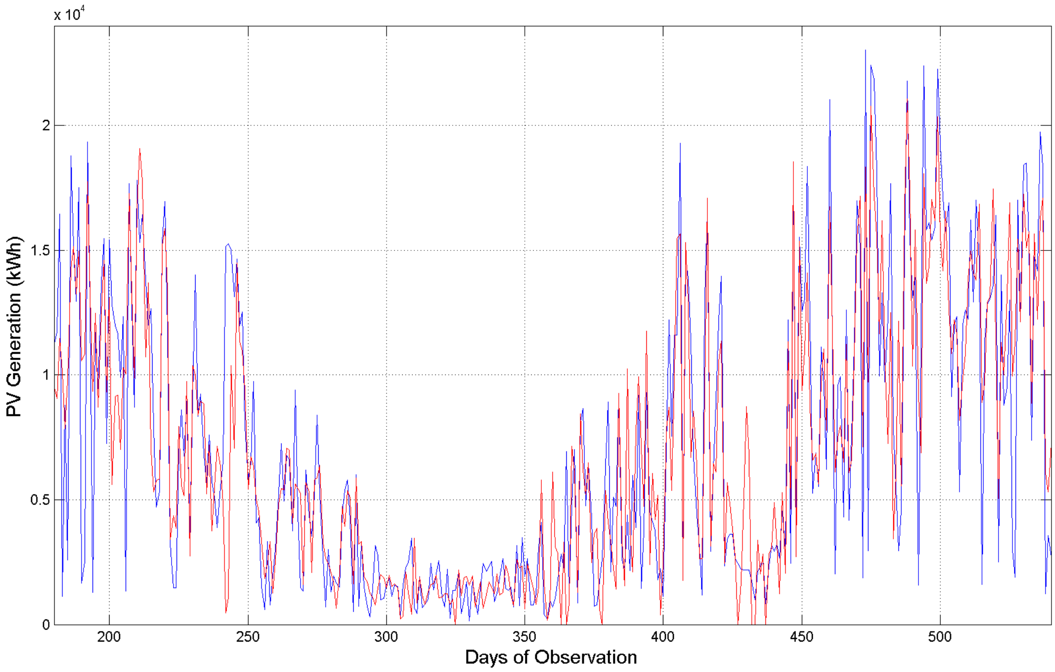


Figure 5.77: Prediction (red line) plotted against the measured PV Generation (blue line) in site A6a over 360 days of observation.

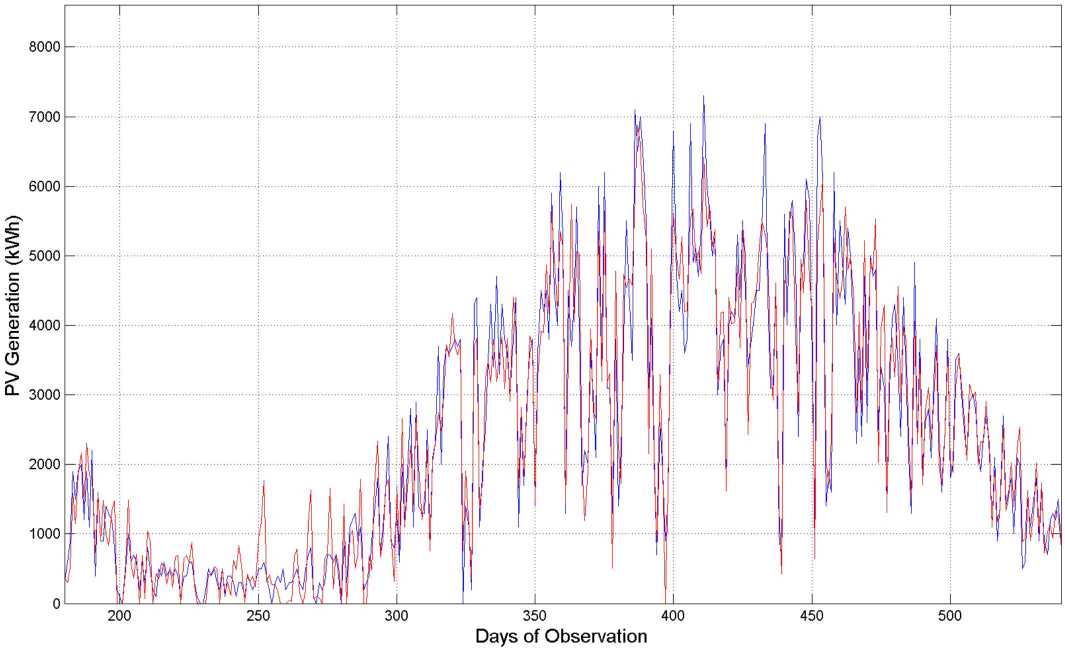


Figure 5.78: Prediction (red line) plotted against the measured PV Generation (blue line) in site A6b over 360 days of observation.

5.2.7 Site A7 - California, USA

Site is located in south California, northeast of San Diego, at coordinates 32.87N - 116.9W, near Lakeside. Site A7 represents a light residential suburban enviroment in a hilly terrain, a few kilometers inland and with a warm summer climate containing dry seasons. The two systems analysed in this site have capacities ranging from 8.16 to 24.5 kWp, the latter being the largest systems forecasted in this work.

Table 5.63: PV Systems in site A7

System	Elevation	Azimuth	Tilt	Capacity
	(m)	(degrees)	(degrees)	(kWp)
A7a	128	270	20	24.150
A7b	200	270	1	8.160

As illustrated in Figure 5.79, two weather stations are used to provide both weather and solar resource parameters for the predictions. The weather stations are designated by its IATA code, the closest being KSEE.

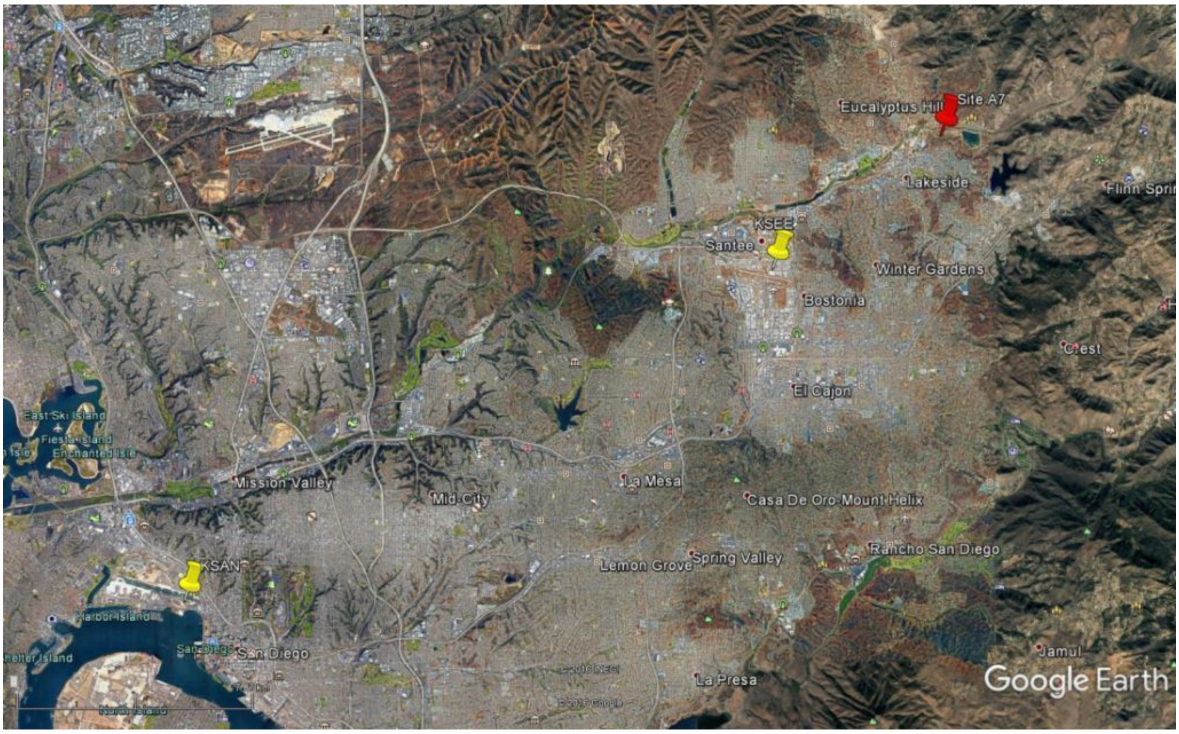


Figure 5.79: Site A7 and the two airport weather stations used for the forecast. Credits: Google Earth

5.2.7.1 Forecasting results, single weather station

Using information from the nearest station, the different methods are applied to forecast the PV generation. Table 5.64 lists the results for system A7a and A7b .

Table 5.64: Results for Systems A7a and A7b, single station

System	Metric\Method	PKF	FGM	KF	PBP	BP
A7a	RMSE	19559,6	34098,8	20755,2	19226,1	20076,7
	MAE	13052,6	24801,0	14953,0	14418,2	14853,7
	MBE	-88,2	1842,4	565,2	-1453,1	-557,5
	MXE	107909,5	222982,0	110496,6	79656,6	100092,7
	r^2	0,846	0,563	0,819	0,852	0,828
A7b	RMSE	2355,642	8268,054	4426,103	6468,202	5597,456
	MAE	1729,417	5767,199	3081,713	4751,885	4121,103
	MBE	-17,045	391,755	34,506	156,529	-103,300
	MXE	13968,052	33237,852	27830,000	32601,750	28100,815
	r^2	0,975	0,750	0,884	0,817	0,869

The PCA-Kalman method performs better for both systems, albeit with a minor performance edge in System A7a

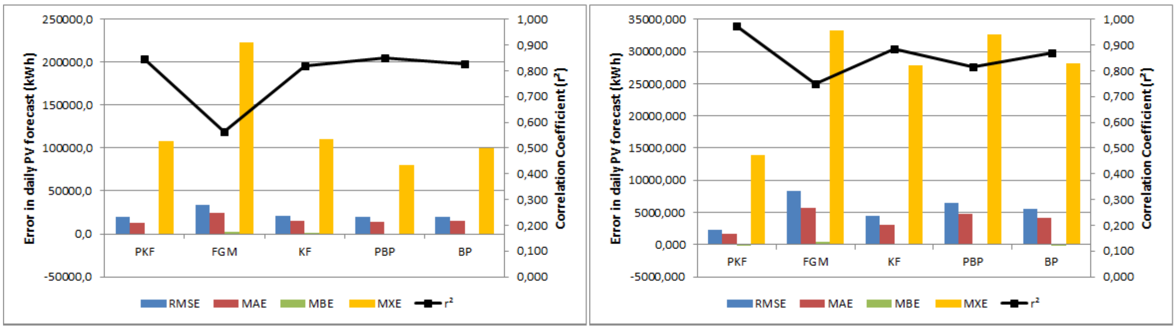


Figure 5.80: Error graphs for forecasts in Systems A6a and A6b, from left to right. Single stations.

5.2.7.2 Forecasting results, multiple weather stations

The forecasting methods are now tried with inputs derived from two weather stations. Table 5.65 lists the results for system A7a and A7b .

Table 5.65: Results for Systems A7a and A7b, multiple stations

System	Metric\Method	PKF	FGM	KF	PBP	BP
A7a	RMSE	16687,0	34098,8	22711,9	20055,6	20978,1
	MAE	11212,7	24801,0	16400,8	15379,7	15723,5
	MBE	90,2	1842,4	899,0	-1196,7	-1496,4
	MXE	103080,4	222982,0	112216,2	84368,9	95141,3
	r^2	0,889	0,563	0,786	0,831	0,804
A7b	RMSE	2257,938	8268,054	5583,472	5827,108	5982,462
	MAE	1523,411	5767,199	3852,952	4246,379	4491,408
	MBE	-84,758	391,755	-78,006	143,069	-244,528
	MXE	20638,529	33237,852	29779,699	28008,626	25235,103
	r^2	0,975	0,750	0,874	0,860	0,849

PCA-Kalman benefits from the second weather station, improving the forecasting performance in all systems, but this effect is more noticeable in system A7a.

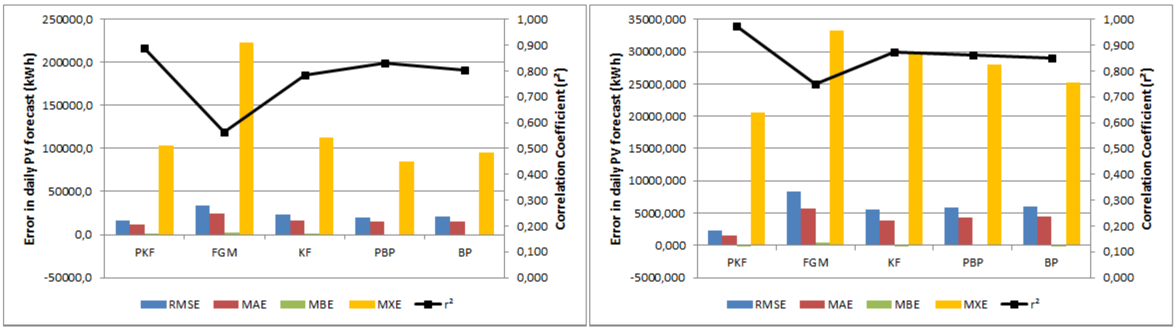


Figure 5.81: Error graphs for forecasts in Systems A6a and A6b, from left to right. Multiple stations.

The forecasts provided by the PCA-Kalman algorithm are compared to the measured PV generation values in figures 5.82 and 5.83:

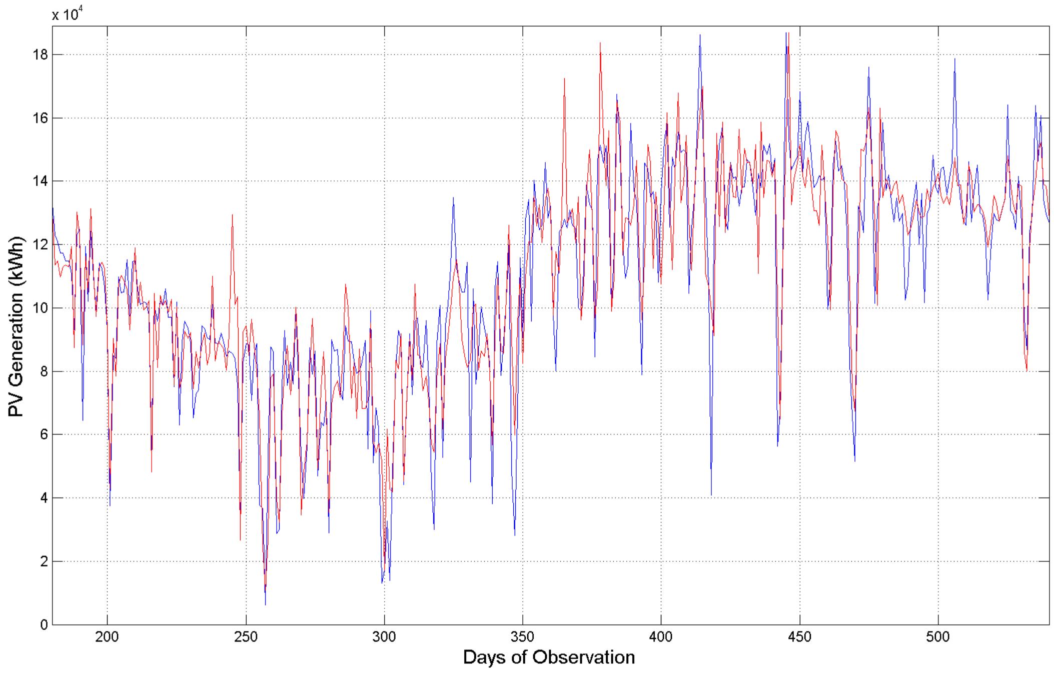


Figure 5.82: Prediction (red line) plotted against the measured PV Generation (blue line) in site A7a over 360 days of observation.

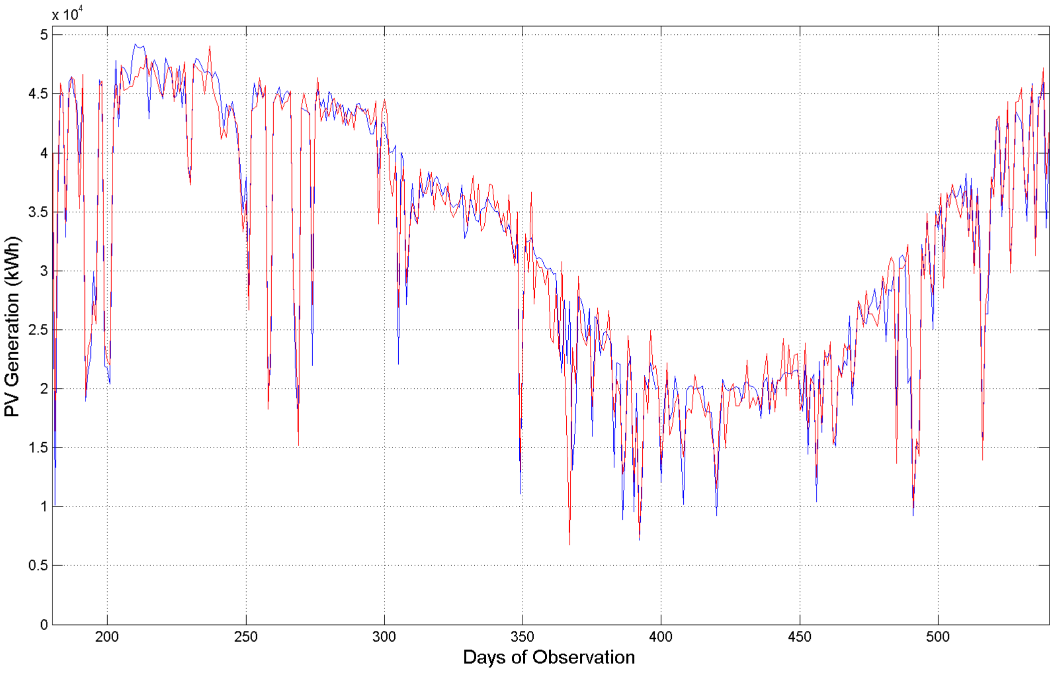


Figure 5.83: Prediction (red line) plotted against the measured PV Generation (blue line) in site A7b over 360 days of observation.

6 CONCLUSIONS

This Chapter is aimed at summarizing the results, point the key findings and conclusions of this work. Section 6.1 deals with the general conclusions. Sections 6.2 and 6.3, respectively, are concerned about specific analysis and commentaries about the load forecasting and photovoltaic generation forecasting. Section 6.4 provides some directions for future research..

6.1 General conclusions

The electric load usually grows due to increasing population or energy intensity. There is also a strong dependence between electrical losses and network reliability with the system load: usually, losses get higher and reliability gets lower with increasing load. As energy prices rise and technology costs decrease, photovoltaic generation increasingly becomes more attractive as an option to provide electricity. The only way to comply with these requirements over time is through a carefully planned network expansion, keeping reliability and quality of service despite increased loads and intermitent, sometimes bidirectional energy flows.

The proposed PCA-Kalman linear model is proven to be satisfactory as being capable of predicting both electric load and PV generation, outperforming a classic Kalman filter and multilayer perceptron (MLP) artificial neural networks and achieving coefficients of correlation oftenly above 90 %. Real data from 33 load forecasting case studies and 15 PV generation case studies has been used to simulate forecasts. Using PCA feature selection, the proposed model benefitted from the additional information provided by a large number of inputs, while the other methods presented loss of performance due to the curse of dimensionality.

Due to their state space mathematical formulation, Kalman filters are intrisically efficient from the computational standpoint. In contrast, backpropagation becomes a cumbersome task when the number of input variables is large. Such theoretical supposition was noticed along the developed analysis, as shown in Table 6.1.

Table 6.1: ANN and Kalman filter processing time ratio.

Scenario	Kalman CPU time	ANN CPU time	Ratio	Input Size
LEJ Set B	46,40%	53,60%	1,155082664	20
PV A6 (Single)	16,29%	83,71%	5,137984129	35
PV A7 (Multi)	20,61%	79,39%	3,851038475	70
BSB Set D	22,93%	77,07%	3,360307076	90
PV A4 (Multi)	7,48%	92,52%	12,36396936	210
PV A5 (Multi)	6,87%	93,13%	13,56342289	210
BSB Set Z	6,74%	93,26%	13,82768111	306

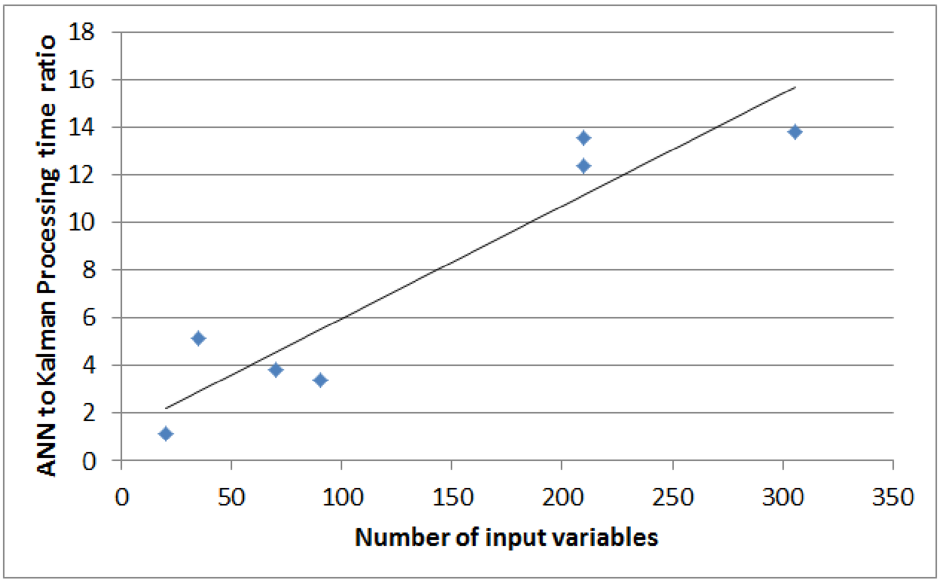


Figure 6.1: Scatter plot of the ANN to Kalman filter processing time ratio, as a function of the input size.

In Figure 6.1, the ANN to Kalman processing time ratio is plotted as a scatter graph. It is noticeable that the ratio gets larger as the size of the input set increases. However, it must be noted that this particular Kalman filter implementation in the MATLAB environment does not have a graphical user interface (GUI), as does the ANN toolbox in the same computational environment. A more fair comparison would require an ANN implementation coded without GUI or assistant wizards.

Considering both the forecasting performance and computational effort, the comparisons performed in this work show that the proposed PCA-Kalman methods compared favourably with the benchmark approaches.

Table 6.2: Summary of results - Leipzig scenario. Best methods per substation and load type.

Load type\Substation	S1	S2	S3	S4	S5	S6	S7	S8	S9
Base load	PKF	BP	BP	PKF	KF	PKF	PKF-KF	KF	BP
Average load	PKF	PKF	PKF	PKF	PKF	PKF	PKF	KF	PKF
Peak load	PKF	PKF	PKF	PKF	PKF	PKF	PKF	PKF	PKF

Table 6.3: Summary of results - Leipzig scenario. Best input sets per substation and load type.

Load type\Substation	S1	S2	S3	S4	S5	S6	S7	S8	S9
Base load	D	A	E	Z	Z-B	Z	Z-B	Z	E
Average load	Z	Z	Z	Z	Z	Z	Z	B	Z
Peak load	Z	Z	Z	Z	Z	Z	Z	Z	Z

6.2 Load forecasting conclusions

The proposed PCA-Kalman forecasting has been compared with other methods in 3 different scenarios, concerning the distribution systems in the cities of Leipzig and Brasilia. Leipzig has nine substations analysed in a time period comprising years 2001-2003, while Brasilia's entire load is forecasted in two different time periods: from year 2001 to 2003, and from 2004 to 2010. Load time series in all scenarios are forecasted at its base, average and peak values.

Summarizing the results presented in section 5.1, Table 6.2 presents the best methods for each substation and type of load time series. PKF represents the proposed PCA-Kalman, KF the classic Kalman filter, PBP the PCA-Backpropagation adjusted multilayer perceptron ANN and BP the standard Backpropagation adjusted MLP ANN.

The proposed method achieves the lowest mean squared error in 21 out of the 27 substation and load type combinations. The classic Kalman filter provides the better performance in 4 out of 27, while the BP is the better method in 3 out of the 27 cases.

Nine different input sets are analysed in this work, as listed in Chapter 4, Table 4.1. Table 6.3 presents the best input sets for each substation and type of load time series.

Despite the increased number of dimensions and the consequent risk of overestimation due to the curse of dimensionality, input set Z provides the better performance overall in 22 out of the 27 cases studies. In almost all of these cases, this input set is paired

Table 6.4: Summary of results - Brasilia scenarios. Best methods per substation and load type.

Load type\Time period	2001-2003	2004-2010
Base load	PKF	PKF
Average load	PKF	PKF
Peak load	PKF	PKF

Table 6.5: Summary of results - Brasilia scenarios. Best input sets per substation and load type.

Load type\Time period	2001-2003	2004-2010
Base load	Z	Z
Average load	Z	Z
Peak load	Z	Z

with the proposed PCA-Kalman method, suggesting that this combination can decrease mean squared errors when compared to smaller inputs sets without feature selection. However, it also must be noted that the BP ANN provided no better performance than the standard BP, even with this input set.

From the analysis of the results obtained from Brasilia's two time periods, Tables 6.4 and 6.5 are constructed.

Similar to the Leipzig scenario, in mean squared error criteria the proposed PCA-Kalman method outperforms all the benchmark approaches at all cases.

The input set Z, again in combination with the proposed method, manages to achieve the lowest mean squared errors in forecasting.

Considering the extensiveness of the case studies performed, this work concludes that the proposed PCA-Kalman load forecasting algorithm realiably outpeforms the benchmark methods. The Mean Average Percentual Error (MAPE) achieved by the proposed model, around 2 % for citysized systems and 4 % for substations, compare favourably with the values given in literature reviews [47], considering the power systems scale. The other methods in the comparison also perform with similar figures to what is presented as state of art, indicating that their performances are adequate to benchmark the proposed forecasting system.

Table 6.6: Summary of results - PV forecasting scenarios. Best methods per system.

PV System\Weather station inputs	Single	Multiple	Better combination
A1a	KF	PKF	KF -Single
A1b	KF	PKF	KF -Single
A1c	KF	PKF	PKF-Multi
A1d	PKF	PKF	PKF-Multi
A2a	PKF	PKF	PKF-Multi
A2b	PKF	PKF	PKF-Multi
A3a	KF	N/A	PKF-Single
A3b	PKF	N/A	PKF-Single
A3c	PKF	N/A	PKF-Single
A4a	KF	PKF	PKF-Multi
A5a	KF	PKF	KF-Single
A6a	PKF	PKF	PKF-Single/Multi
A6b	PKF	PKF	PKF-Multi
A7a	PKF	PKF	PKF-Multi
A7b	PKF	PKF	PKF-Multi

6.3 PV Generation forecasting conclusions

Summarizing the results presented in section 5.2, Table 6.6 presents the best methods for each combination of PV system and number of weather stations. PKF represents the proposed PCA-Kalman, KF the classic Kalman filter, PBP the PCA-Backpropagation adjusted MLP ANN and BP the standard MLP ANN. N/A denotes that the multiweather station approach is not performed.

Interpreting the results by the Root Mean Squared Error criteria, it can be noticed that the proposed PCA-Kalman approach outperform the benchmarks in 12 out of the 15 cases, while the classic KF is the better method in the remaining 3 cases. Comparing input data from single or multiple weather stations, it is noticed that in 9 out of 15 the additional inputs provided by the multiple weather stations is beneficial to the forecasting, while in 4 out of the 15 the single weather station approach is more accurate. In a single case, the results among single and multiple weather stations are practically equal. The proposed PCA-Kalman filter performed better with multiple weather inputs, while the classic Kalman forecasted more accuratelly with single weather inputs.

The proposed method averaged a coefficient of correlation between prediction and measurements above 90 % in most of the cases, while all except one of the benchmark methods averaged between 75-85 %. The autoregressive Gray model, without inputs, averaged less than 70 %. Except for the latter method, both benchmark and proposed methods performed reasonably, considering the plant size and uncertainties in the solar resource [10, 108].

6.4 Directions for future research

Further research must focus on expanding the sets of candidate variables and investigate the possibility to develop universal types of nonlinear transformations, applicable to the full set of candidate variables.

The attempt of combining PCA and multilayer perceptron ANN adjusted by Backpropagation (BP) does not present an advantage in error performances. The probable cause is that the number of dimensions is chosen accordingly to the Kalman filter's performance, while the ANN can have difficulties optimizing all parameters in a high dimensionality scenario. This becomes more clear when thre results show the ANN methods performing better with the smaller input sets, such as A, E and F.

There are opportunities to employ the PCA to also help determine the model order. Figures 6.2 and 6.3 depict the singular values in bar charts over the principal components, as calculated from Leipzig and PV Site 5, respectively. It is noticeable that there are transitions in the components around the chosen model order for each case.

In future works, the model order selection can be obtained by means of Bayesian or Akaike criteria applied to the principal components.

More advanced feature selection procedures either substituting or complementing PCA should be attempted, in order to further reduce complexity and avoid overestimation. A promising candidate is the MinMax technique.

The effect of rapidly growing distributed generation, grid storage and demand response over the performance of this load forecasting system might also be a topic for future work, given availability of applicable time series data.

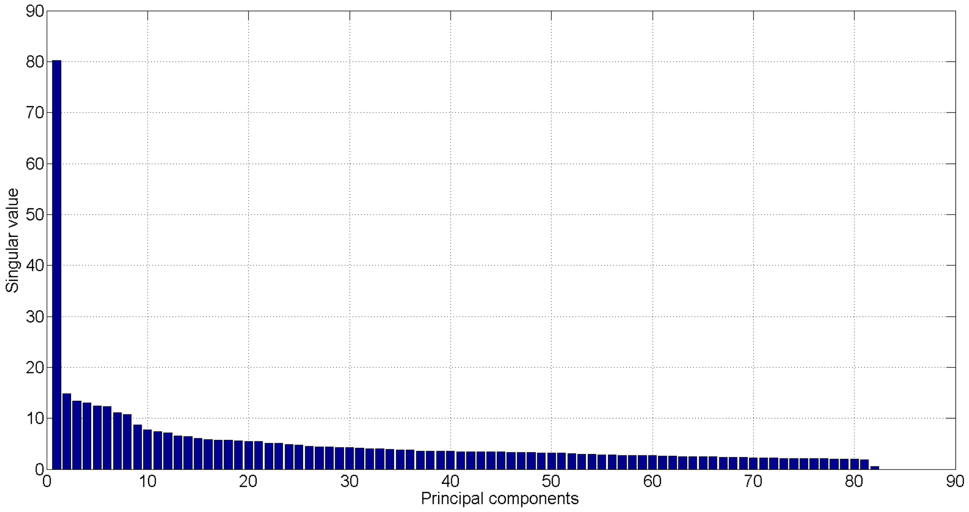


Figure 6.2: Principal components horizontally sorted in decreasing order of singular value. There are noticeable discontinuities around the first, the seventh and eighth component. Model order in this case has been selected as 7.

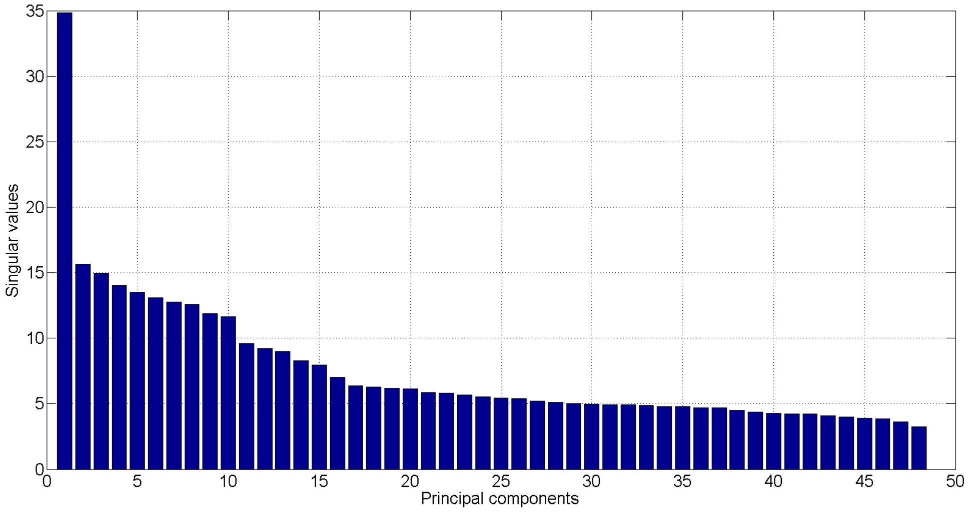


Figure 6.3: Principal components horizontally sorted in decreasing order of singular value. There are noticeable discontinuities around the first and the tenth component. Model order in this case has been selected as 10.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Bird simple spectral model. Available at http://rredc.nrel.gov/solar/models/spectral/. Retrieved in May 1st 2015.

[2] British petroleum statistical review 1951-2011. Available at http://www.bp.com, 2013. Retrieved in April 10th 2015.

[3] The World Bank Report. Available at http://data.worldbank.org/country/germany, 2015. Retrieved in February 20th 2015.

[4] British petroleum statistical review of world energy, june 2015. Available at http://www.bp.com/statisticalreview, 2016. Retrieved in February 12th 2017.

[5] Albadi, M. e El-Saadany, E. Demand response in electric markets: An overview. In *Power Engineering Society General Meeting*, pages 1–5. IEEE, June de 2014.

[6] Albadi, M. H. e El-Saadany, E. F. A summary of demand response in electricity markets. *Electric Power Systems Research*, 78(11):1989–1996, November de 2008.

[7] Altinaya, G. e Karagol, E. Electricity consumption and economic growth: Evidence from turkey. *Energy Economics*, 27(6):849–856, November de 2005.

[8] Amjady, N. Day-ahead price forecasting of electricity markets by a new fuzzy neural network. *IEEE Transactions on Power Systems*, 21(2):887–896, 2006.

[9] Amjady, N. Short-term bus load forecasting of power systems by a new hybrid method. *IEEE Transactions on Power Systems*, 22(1):333–341, February de 2007.

[10] Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Pison, F. M.de , e Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016.

[11] Asafu-Adjaye, J. The relationship between energy consumption, energy prices and economic growth: time series evidence from asian developing countries. *Energy Ec*, 22(6):615–625, December de 2000.

[12] Asbury, C. E. Weather load model for electric demand and energy forecasting. *IEEE Transactions on Power Apparatus and Systems*, 94(4):1111–1116, July de 1975.

[13] Bashari, M., Rahimi-Kian, A., e Farhangi, S. Forecasting generation of a pv power plant with a little data-set using information fusion. *2014 Smart Grid Conference (SGC)*, pages 1–6, 2014.

[14] Bianco, V., Manca, O., e Nardini, S. Electricity consumption forecasting in italy using linear regression models. *Energy*, 34(9):1413–1421, September de 2009.

[15] Bianco, V., Manca, O., Nardini, S., e Minea, A. A. Analysis and forecasting of nonresidential electricity consumption in romania. *Applied Energy*, 87(11):3584–3590, November de 2010.

[16] Bird, R. E. e Riordan, C. Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth's surface for cloudless atmospheres. *Journal of Climate and Applied Meteorology*, 25, Issue 1:87–97, January de 1986.

[17] Blocken, B., Defraeye, T., Derome, D., e Carmeliet, J. High-resolution cfd simulations for forced convective heat transfer coefficients at the facade of a low-rise building. *Building and Environment*, 44(12):2396–2412, December de 2009.

[18] Box, G. E. P. e Jenkins, G. M. *Time Series Analysis, Forecast and Control.* Holden-day, 1970.

[19] Brest, C. L. Seasonal albedo of an urban/rural landscape from satellite observations. *Journal of Applied Meteorology*, 26 Issue 9:1169–1187, September de 1987.

[20] Bruhns, A., Deurveilher, G., e Roy, J. S. A non linear regression model for midterm load forecasting and improvements in seasonality. *Proceedings of the 15th Power Systems Computation Conference*, 2005.

[21] Carpaneto, E., Chicco, G., Napoli, R., e Scutariu, M. Customer classification by means of harmonic representation of distinguishing features. *in Proc. IEEE Power Tech Conf.*, 3, June de 2003.

[22] Chontanawata, J., Huntb, L. C., e Pierse, R. Does energy consumption cause economic growth? evidence from a systematic study of over 100 countries. *Journal of Policy Modeling*, 30(2):209–220, March-April de 2008.

[23] Day, T. *Degree-days: theory and application.* The Chartered Institution of Building Services Engineers, 222 Balham High Road, London SW12 9BS, September de 2006.

[24] Luca, L. A. D.de , Oliveira, C. M.de , e Wazlawick, R. S. Load behavior changes after holidays on thursdays. *Computational Science and Engineering Workshops*, pages 101–106, 2008.

[25] Defraeyea, T., Blockenb, B., e Carmeliet, J. Convective heat transfer coefficients for exterior building surfaces: Existing correlations and cfd modelling. *Energy Conversion and Management*, 52(1):512–522, January de 2011.

[26] DNV GL,. A review of distributed energy resources. Technical report, New York Independent System Operator, September de 2014.

[27] Draper, M. Modeling weather effects on electric energy sales. *IEEE Proceedings of the Southeastcon*, 1:133–136, 1990.

[28] El-Ferik, S. e Malhame, R. Identification of alternating renewal electric load models from energy measurements. *IEEE Transactions on Aut*, 39(6):1184–1196, 1994.

[29] Elias, R. S., Fang, L., e Wahab, M. I. M. Electricity load forecasting based on weather variables and seasonalities: A neural network approach. *8th International Conference on Service Systems and Service Management*, pages 1–6, 2011.

[30] Emmel, M. G., Abadie, M. O., e Mendes, N. New external convective heat transfer coefficient correlations for isolated low-rise buildings. *Energy and Buildings Journal - Elsevier*, 39(3):335–342, March de 2007.

[31] Ernoult, M. e Mattatia, R. Short-term load forecasting: New developments at the edf. *Proceedings of the Eighth Power Systems Computation Conference*, (First Edition), August de 1984.

[32] Espinoza, M., Suyjens, J. A. K., Belmans, R., e Moor, B.de . Electric load forecasting using kernel-based modeling for nonlinear system identification. *IEEE Control Systems Magazine*, pages 43–57, October de 2007.

[33] Feinberg, E. e Genethliou, D. *Applied Mathematics for Restructured Electric Power Systems.* Springer, 2005.

[34] Ferreira, M. J., Oliveira, A. P.de , J. Soares, G. C., Barbaro, E. W., e Escobedo, J. F. Radiation balance at the surface in the city of são paulo, brazil:

diurnal and seasonal variations. *Theoretical and Applied Climatology*, 107, Issue 1-2:229–246, January de 2012.

[35] Forrest, J. The effects of weather on power-system operation. *Journal of the Institution of Electrical Engineers*, 93(64):161–163, April de 1946.

[36] Friedrich, L., Armstrong, P., e Afshari, A. Mid-term forecasting of urban electricity load to isolate air-conditioning impact. *Energy and Buildings Journal - Elsevier*, 80:72–80, September de 2014.

[37] Frisch, H. L., Borzi, C., Ord, G., Percus, J. K., e Williams, G. O. Approximate representation of functions of several variables in terms of functions of one variable. *Physical Review Letters*, 63:927–929, 1989.

[38] Fuwei, Z. e Xuelian, Z. Gray-regression variable weight combination model for load forecasting. *International Conference on Risk Management & Engineering Management*, pages 311–316, 2008.

[39] Green, M. A. e Keevers, M. Optical Properties of Intrinsic Silicon at 300 K. *Progress in Photovoltaics*, 3(3):189–192, 1995.

[40] Gross, G. e Galiana, F. Short-term load forecasting. *Proceedings IEEE*, 75:1558–1573, December de 1987.

[41] Grunewald, P. e Torriti, J. Demand response from the non-domestic sector: Early uk experiences and future opportunities. *Energy Policy*, 61:423–429, 2013.

[42] H., N. e Labs, D. Improved data of solar spectral irradiance from 0.33 to 1.25 micrometers. *Solar Physics*, 74:231–249, 1981.

[43] Hardle, W., Muller, M., Sperlich, S., e Werwatz, A. *Nonparametric and Semiparametric Models.* Springer Series in Statistics. Springer-Verlag Berlin Heidelberg, 2004.

[44] Hassanzadeh, M., Etezadi-Amoli, M., e Fadali, M. S. Practical approach for sub-hourly and hourly prediction of pv power output. *North American Power Symposium 2010*, pages 1–5, 2010.

[45] Hastie, T. J. e Tibshirani, R. J. *Generalized Additive Models.* Chapman & Hall/CRC., 1990.

[46] He, Y., Zhang, J. X., Xu, Y., Gao, Y., Xia, T., e He, H. Forecasting the urban power load in china based on the risk analysis of land-use change and load density. *International Journal of Electrical Power & Energy Systems*, 73:71–79, 2015.

[47] Hernandez, L., Baladron, C., Aguiar, J. M., Carro, B., Sanchez-Esguevillas, A. J., Lloret, J., e Massana, J. A survey on electric power demand forecasting: Future trends in smart grids, microgrids and smart buildings. *IEEE Communications Surveys and Tutorials*, 16(3), Third Quarter de 2014.

[48] Heuklon, T. K. V. Estimating atmospheric ozone for solar radiation models. *Solar Energy*, 22:63–68, 1979.

[49] Hippert, H. S., Pedreira, C. E., e Souza, C. R. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, 16(1):44–51, 2001.

[50] Ilic, M., Krogh, B. H., e Blood, E. A. Electric power system static state estimation through kalman filtering and load forecasting. In *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*. IEEE, 2008.

[51] International Energy Agency,. Technology roadmap solar photovoltaic energy. Technical report, International Energy Agency, 2014.

[52] International Energy Agency,, editor. *Key World Energy Statistics*. International Energy Agency, 9, rue de la Fédération 75739 Paris Cedex 15, 2016.

[53] Iqbal, M. *An introduction to solar radiation*. Academic press Canada, 1983.

[54] Jardini, J., Tahan, C., Gouvea, M., e Ahn, S. Daily load profiles for residential, commercial and industrial low voltage consumers. *IEEE Transactions on Power Delivery*, 15(1):375–380, 2000.

[55] Justus, C. G. e Paris, M. V. A model for solar spectral irradiance at the bottom and top of a cloudless atmosphere. *Journal of Climate and Applied Meteorology*, 1984.

[56] Kalman, R. E. New approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82 (Series D):35–45, 1960.

[57] Karavana-Papadimou, K., Psiloglou, B., Lykoudis, S., e Kambezidis, H. D. *Model for Estimating Atmospheric Ozone Content over Northern Europe for Use in Solar Radiation Algorithms*, chapter Atmospheric Physics, pages 1025–1031. Springer Atmospheric Sciences, 2013.

[58] Kasten, F. A new table and approximate formula for relative optical air mass. *Arch. Meteorol. Geophys. Biochlimatol.*, B14:206–223, 1966.

[59] Kayacan, E., Ulutas, B., e Kaynak, O. Grey system theory-based models in time series prediction. *Expert Systems with Applications*, 37:1784–1789, 2010.

[60] Keogh, E. e Mueen, A. *Encyclopedia of Machine Learning*. Springer, 2010.

[61] Kermanshahi, B. Recurrent neural network for forecasting next 10 years loads of nine japanese utilities. *Neurocomputing*, 23:125–133, 1998.

[62] Kneizys, F. X., Shettle, E. P., Gallery, W. O., Jr, J. H. C., Abrea, L. W., Selby, J. E. A., Fenn, R. W., e McClatchey, R. W. Atmospheric transmittance/radiance: Computer code lowtran5. techreport AFGL-TR-800067, Air Force Geophysics Laboratory, 1980.

[63] Kolmogorov, A. The representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, 114:953–956, 1957.

[64] Kurkova, V. Kolmogorov theorem is relevant. *Neural Computation*, 3(4):617–622, 1991.

[65] Leckner, B. The spectral distribution of solar radiation at the earth's surface: Elements of a model. *Solar Energy*, 20:143–150, 1978.

[66] Leipzig Stadt,. *"Strukturatlas Leipzig 03-04"*. Leipziger Statistik und Stadtforschung, 2004.

[67] Leni, P.-E., Fougerolle, Y., e Truchetet, F. Kolmogorov superposition theorem for image compression. *IET Image Processing*, 6(8):1114–1123, 2012.

[68] Li, K. e Tai, N. Research and application of climatic sensitive short - term load forecasting. *Power & Energy Society General Meeting*, pages 1–5, july de 2015.

[69] Lima, D. A., Perez, R. C., e Clemente, G. A comprehensive analysis of the demand response program proposed in brazil based on the tariff flags mechanism. *Electric Power Systems Research*, 144:1–12, November de 2016.

[70] Liu, Y. e Harris, D. Full-scale measurements of convective coefficient on external surface of a low-rise building in sheltered conditions. *Building and Environment*, 42(7):2718–2736, July de 2007.

[71] Lyde, D. R. *CRC Handbook of chemistry and physics*. CRC Press, 85 th edition, 2004.

[72] Majithia, S., Watson, S. J., e Hor, C. L. Analyzing the impact of weather variables on monthly electricity demand. *IEEE Transactions on Power Systems*, 20:2078–2085, November de 2005.

[73] Milanezi Junior, J. Spatio-temporal prediction of electric power systems including emergent renewable energy sources. Dissertação de Mestrado, University of Brasilia, Brazil, march de 2014.

[74] Moghram, I. e Rahman, S. Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on Power Systems*, 4:1484–1491, November de 1989.

[75] Mohamed, O., Park, D., Merchant, R., Dinh, T., Tong, C., e Azeem, A. Practical experiences with an adaptive neural network short -term load forecasting system. *IEEE Transactions on Power Systems*, 10(1):254–265, 1995.

[76] Mohamed, Z. e Bodger, P. Forecasting electricity consumption in new zealand using economic and demographic variables. *Energy*, 30(10):1833–1843, July de 2005.

[77] Myers, D. Terrestrial solar spectral distributions derived from broadband hourly radiation data. *Optical Modeling and Measurements for Solar Energy Systems III, Proceedings of SPIE*, 7410(01), September de 2009.

[78] Myers, D. Direct beam and hemispherical terrestrial solar spectral distributions derived from broadband hourly solar radiation data. *Solar*, 86(9):2771–2782, September de 2012.

[79] Nakamura, M., Mines, R., e Kreinovich, V. Guaranteed intervals for kolmogorov theorem (and their possible relation to neural networks). *Interval Computations*, pages 183–199, 1993.

[80] NOAA,. *Federal Meteorological Handbook No. 1 - Surface Weather Observations and Reports*. U.S.Department of Commerce / National Oceanic and Atmospheric Administration, fcm-h1-2005 edition, September de 2005.

[81] Owayedh, M., Al-Bassam, A., e Khan, Z. Identification of temperature and social events effects on weekly demand behavior. *Power Engineering Society Summer Meeting*, 4:2397–2402, 2000.

[82] Papalexopoulos, A. e Hesterberg, T. A regression-based approach to short-term load forecasting. *IEEE Transactions on Power Systems*, 5(4):1535–1550, 1990.

[83] Park, J. H., Park, Y. M., e Lee, K. Y. Composite modeling for adaptive short-term load forecasting. *IEEE Transactions on Power S*, 6(2):450–457, May de 1991.

[84] Peng, M., Hubele, N., e Karady, G. Advancement in the application of neural networks for short-term load forecasting. *IEEE Transactions on Power Systems*, 7:250, 257 de 1992.

[85] Photovoltaic Power Systems Programme,. A snapshot of global pv (1992-2015). Technical report, International Energy Agency, 2016.

[86] Photovoltaic Power Systems Programme,. Trends 2016 in photovoltaic applications. Technical report, International Energy Agency, 2016.

[87] Ribeiro, L. D. X., Milanezi Jr., J., da Costa, J. P., Giozza, W. F., Kehrle Miranda, R., e Vieira da Silva, M. V. PCA-Kalman based load forecasting of electric power demand. *2016 IEEE International Symposium on Signal Processing and Information*, pages 63–68, 2016.

[88] Sailor, D. J. Relating residential and commercial sector electricity loads to climate - evaluating state level sensitivities and vulnerabilities. *Energy*, 26:645–657, July de 2001.

[89] Soares, L. J. e Medeiros, M. C. Modeling and forecasting short-term electricity load: A comparison of methods with an application to brazilian data. *International Journal of Forecasting*, 24(4):630–644, October-December de 2008.

[90] Soliman, S. A. e Al-Kandari, A. M. *Electrical load forecasting: modeling and model construction*. Elsevier, 2010.

[91] Soliman, S., Persaud, S., El-Nagar, K., e El-Hawary, M. Application of least absolute value parameter estimation based on linear programming to short-term load forecasting. *Electric Power Energy Systems*, 19(3):209–216, 1997.

[92] Spencer, J. W. Fourier series representation of the position of the sun. *Search*, 2:172, 1971.

[93] Stetson, L. e Stark, G. Peak electrical demands of individuals and groups of rural residential customers. *IEEE Transactions on Industry Applications*, 24(5):772–776, Sept/Oct de 1988.

[94] Stevens, J. e Choo, K. Temperature sensitivity of the body surface over the life span. *Somatosensory & Motor Research*, 15:13–28, 1998.

[95] Supit, I. e van Kappel, R. R. A simple method to estimate global radiation. *Solar Energy*, 63(3):147–160, September de 1998.

[96] T., B. D. e Iqbal, M. Solar spectral diffuse irradiance under cloudless skies. *Solar Energy*, 30:447–453, 1983.

[97] Taha, H. Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. *Energy and Buildings Journal - Elsevier*, 25, Issue 2:99–103, 1997.

[98] Tien, T.-L. A new grey prediction model fgm(1, 1). *Mathematical and Computer Modelling*, 49:1416–1426, 2009.

[99] Tripathi, M., Upadhyay, K., e Singh, S. Short-term load forecasting using generalized regression and probabilistic neural networks in the electricity market. *The Electricity Journal*, 21(9):24–34, November de 2008.

[100] Tsay, R. S. *Analysis of Financial Time Series.* Wiley-Interscience, 2002.

[101] Tuyishimire, B., McCann, R., e Bute, J. Evaluation of a kalman predictor approach in forecasting pv solar power generation. *4th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, pages 1–6, 2013.

[102] UK Meteorological Office,. Method for calculating heating and cooling degree days. *The Weekly Weather Report*, February de 1928. Available at http://ukclimateprojections.metoffice.gov.uk/22715. Retrieved in september 9th, 2015.

[103] Vos, J. J. Colorimetric and photometric properties of a 2-deg fundamental observer. *Color Research & Application*, 3(125), 1978.

[104] Wasserman, L. *All of Nonparametric Statistics.* Springer-Verlag Berlin, 2006.

[105] Wirth, H. Recent facts about photovoltaics in germany. techreport, Fraunhofer ISE, April de 2016.

[106] Wu, L., Liu, S., e Yang, Y. A gray model with a time varying weighted generating operator. *IEEE Transactions on Systems Man and Cybernetics*, 46(3):427–433, March de 2016.

[107] Wyszecki, G. e Stiles, W. *Color Science - Concepts and Methods, Quantitative Data and Formulae.* John Wiley and Sons, 2nd edition, 2000.

[108] Yadav, H. K., Pal, Y., e Tripathi, M. Photovoltaic power forecasting methods in smart power grid. *Annual IEEE India Conference (INDICON)*, pages 1–6, 2015.

[109] Yuill, W., Kgokong, R., Chowdhury, S., e Chowdhury, S. P. Application of adaptive neuro fuzzy inference system (anfis) based short term load forecasting in south african power networks. *International Universities Power Engineering Conference (UPEC)*, (45th):1–5, 2010.

[110] Ziesing, H.-J. Energieverbrauch in deutschland in jahr 2015. Technical report, AG Energiebilanzen e.V., 2016.

APPENDICES

A The Kalman filter

This appendix contains a more complete description and derivation of the Kalman filter, including:

1. An extended overview;
2. State space representation;
3. Filter derivation;
4. Variance tracking.

Each of these topics will be described in sections A.1 to A.4.

A.1 Overview

In 1960, Rudolf E. Kalman published his famous paper describing a recursive solution to the discrete-data linear filtering problem. Since that time, due in large part to advances in digital computing, the Kalman filter has been the subject of extensive research and application, particularly in the area of autonomous or assisted navigation, data fusion and forecasting of stochastic systems. Typical uses of the Kalman filter include smoothing noisy data and providing estimates of parameters of interest. Applications include global positioning system (GPS) receivers, phaselocked loops (PLL) in radio equipment, smoothing the output from touchpads and touchscreens, and many more.

The Kalman filter is over 50 years old but is still one of the most important and common data fusion algorithms in use today, due to its small computational requirement, elegant recursive properties, and its status as the optimal estimator for one-dimensional linear systems with Gaussian error statistics.

Theoretically the Kalman filter is an estimator for the linear-quadratic problem, which is the problem of estimating the instantaneous state of a linear dynamic system perturbed by white noise. In this sentence, state relates to the so-called state-space representation of a dynamic system, a concise mathematical model based on a finite system of differential equations. Precisely, state can be defined as the values assumed by the state variables, which in turn are related to the degrees of freedom presented by system of differential equations. An important property of the state space representation is that given knowledge of the state at the present instant t_0, this information embodies all previous history of the system's states and inputs. The future outputs of the system are entirely determined by the state at t_0 and by the future values of the inputs.

An estimator is a system that calculates as output a parameter or variable of interest, having sequence of observations as inputs. The Kalman filter is a recursive estimator that calculates a minimum variance estimate for a state that evolves in time as a linear function of variables related to this state. Recursive means that only the previous time step state needs to be stored in memory. The Kalman filter is optimum with respect to diverse criteria, provided some specific hypothesis about process and observation noise are true.

A.2 State space representation

In the general case, the state space model of a dynamic system can be derived from two sets of differential equations: the first shown in Eq. (A.1) relating the m input variables u_i to the n state variables x_j at a given instant t, and the latter reffered in Eq. (A.2) relating u_i and $x_j$to the p output variables $y_q$at a given instant t. $\dot{x}_j$ represents the first derivative of x_j.

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} F_1(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \\ F_2(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \\ \vdots \\ F_n(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \end{bmatrix} \tag{A.1}$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{bmatrix} = \begin{bmatrix} H_1(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \\ H_2(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \\ \vdots \\ H_p(x_1(t), x_2(t), ..., x_n(t), u_1(t), u_2(t), ..., u_m(t)) \end{bmatrix} \tag{A.2}$$

Note that this definition does not require a linear relationship between the variables, and that a Multiple Input Multiple Output (MIMO) system is described. Also note that this representation does not handle systems with delays and those defined by partial differential equations. Delayed input variables, however, can be easily added as additional input variables.

If the n functions F_i and p functions H_p are linear, eqs (A.1) and (A.2) can also be expressed as Eqs. (A.3) and (A.4):

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} a_{11}x_1(t) + a_{12}x_2(t) + ... + a_{1n}x_n(t) + b_{11}u_1(t) + b_{12}u_2(t) + b_{1m}u_m(t) \\ a_{21}x_1(t) + a_{22}x_2(t) + ... + a_{2n}x_n(t) + b_{21}u_1(t) + b_{22}u_2(t) + b_{2m}u_m(t) \\ \vdots \\ a_{n1}x_1(t) + a_{n2}x_2(t) + ... + a_{nn}x_n(t) + b_{n1}u_1(t) + b_{n2}u_2(t) + b_{nm}u_m(t) \end{bmatrix} \tag{A.3}$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{bmatrix} = \begin{bmatrix} c_{11}x_1(t) + c_{12}x_2(t) + ... + c_{1n}x_n(t) + d_{11}u_1(t) + d_{12}u_2(t) + d_{1m}u_m(t) \\ c_{21}x_1(t) + c_{22}x_2(t) + ... + c_{2n}x_n(t) + d_{21}u_1(t) + d_{22}u_2(t) + d_{2m}u_m(t) \\ \vdots \\ c_{n1}x_1(t) + c_{n2}x_2(t) + ... + c_{pn}x_n(t) + d_{p1}u_1(t) + d_{p2}u_2(t) + d_{pm}u_m(t) \end{bmatrix} \tag{A.4}$$

These two sets of equations can be expressed in matricial form, by the following groupment of variables in vectors and parameters in matrices:

$$\dot{X}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} ;\ X(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} ;\ Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{bmatrix} ;\ U(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_m(t) \end{bmatrix} ;$$

$$\mathbf{A_c} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} ;\ \mathbf{B_c} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix} ;$$

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{bmatrix}; \mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \cdots & d_{pm} \end{bmatrix};$$

where $X(t) \in \mathbb{R}^n$ is the state vector and $\dot{X}(t)$ is its derivative, $Y(t) \in \mathbb{R}^p$ is the measurements and/or output vector, and $U(t) \in \mathbb{R}^m$ is the input vector. $\mathbf{A_c} \in \mathbb{R}^{n \times n}$ is the system or dynamics matrix (continuous time), $\mathbf{B_c} \in \mathbb{R}^{n \times m}$ is the input matrix (continuous time), $\mathbf{C} \in \mathbb{R}^{p \times n}$ is the output or sensor matrix, and $\mathbf{D} \in \mathbb{R}^{p \times m}$ is the direct transmission or feedthrough matrix.

With this more succint description, one can write Eqs. (A.3) and (A.4) as:

$$\dot{X}(t) = \mathbf{A_c} X(t) + \mathbf{B_c} U(t) \tag{A.5}$$

$$Y(t) = \mathbf{C} X(t) + \mathbf{D} U(t) \tag{A.6}$$

Note that Eqs. (A.5) and (A.6) describe a dynamic system in continuous time. For many practical problems, however, one is only interested in knowing the state of a system at a discrete set of times $t_k \in \{t_1, t_2, t_3, ...\}$. Considering that t_1, t_2 and t_3 are equally spaced in time by a period τ, it is convenient to order the times t_k according to their integer subscripts:

$$\begin{aligned} x(t_1) &= x[1] \\ x(t_2) &= x[2] \\ &\vdots \\ x(t_k) &= x[k] \end{aligned}$$

For problems with discrete time of this type, it suffices to define the state as a recursive relation of difference equations, instead of differential equations. If τ is small relative to the system dynamics and consequently the input remains approximately constant during each timestep (zero order hold), it is possible to employ the approximation of the first derivate presented in Eq. (A.7):

$$\dot{X}(t_{k+1}) \approx \frac{X[k+1] - X[k]}{\tau} \tag{A.7}$$

Substituting Eq. (A.7) in (A.5) yields:

$$\frac{X[k+1] - X[k]}{\tau} = \mathbf{A_c} X[k] + \mathbf{B_c} U[k]$$

$$X[k+1] = X[k] + \tau \mathbf{A_c} X[k] + \tau \mathbf{B_c} U[k]$$

$$X[k+1] = (\mathbf{I_n} + \tau \mathbf{A_c}) X[k] + \tau \mathbf{B_c} U[k]$$

$$X[k+1] = \mathbf{A} X[k] + \mathbf{B} U[k] \tag{A.8}$$

$$Y[k] = \mathbf{C} X[k] + \mathbf{D} U[k] \tag{A.9}$$

where $\mathbf{A} = \mathbf{I_n} + \tau \mathbf{A_c}$ and $\mathbf{B} = \tau \mathbf{B_c}$. $\mathbf{I_n}$ denotes the identity matrix of order n. Notice that $\mathbf{C}$ and $\mathbf{D}$ are not affected in this discretization procedure.

A.2.1 Obtaining a discrete state space representation from a difference equation

It is possible to convert a nth order linear difference equation of a Multiple Input Single Output (MISO) system to a state space model by means of the so-called companion model. Starting with the difference equation (A.10):

$$y[k+1] + \alpha_1 y[k] + \alpha_2 y[k-1] + ... + \alpha_{n-1} y[k-n-2] + \alpha_n y[k-n-1] = \beta_1 u_1[k] + \beta_2 u_2[k] + ... + \beta_m u_m[k] \tag{A.10}$$

Defining the state and output vectors in the companion forms:

$$X[k+1] = \begin{bmatrix} y[k+1] \\ y[k] \\ \vdots \\ y[k-n] \end{bmatrix} ; \; X[k] = \begin{bmatrix} y[k] \\ y[k-1] \\ \vdots \\ y[k-n-1] \end{bmatrix} ; \; U[k] = \begin{bmatrix} u_1[k] \\ u_2[k] \\ \vdots \\ u_m[k] \end{bmatrix}$$

and using the fundamental relationship between the past and present values of $y[k]$, when the timestep k is increased by 1:

$$\begin{aligned} y[k] &= y[k-1] \\ y[k-1] &= y[k-2] \\ &\vdots \\ y[k-n] &= y[k-n-1] \end{aligned}$$

Equation (A.10) can be rewritten in the matricial form:

$$X[k{+}1] = \begin{bmatrix} -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{n-2} & -\alpha_{n-1} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} X[k]{+} \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_m \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} U[k] \quad \text{(A.11)}$$

Defining the system output $Y[k]$ as $y[k]$, one could also write the output matrix equation:

$$Y[k] = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} X[k] \qquad \text{(A.12)}$$

By inspection, it can be noticed that Eqs. (A.11) and (A.12) are equivalent to Eqs. (A.8) and (A.9), respectively. In this case, the direct transmission matrix **D** is equal to zero.

An important observation is that the companion form can lead to a poorly conditioned system in the numerical sense. This is a direct consequence of concentrating all system information in a single row. If the coefficients α_j and β_i differ in value by several orders of magnitude, rounding errors and numerical issues may arise. In order to mitigate these problems, it is advisable to employ preconditioning and robust numerical methods, such as Generalized Minimal RESidual (GMRES) when evaluating or solving these equations.

A.3 Filter derivation

The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error. The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modeled system is unknown.

The Kalman filter addresses the general problem of trying to estimate the state $X[k+1] \in \mathbb{R}^n$ of a discrete-time controlled process that is governed by the linear stochastic difference equation (A.13) with a measurement $Y[k] \in \mathbb{R}^p$ of the systems output given by equation (A.14):

$$X[k+1] = \mathbf{A}X[k] + \mathbf{B}U[k] + W[k] \tag{A.13}$$

$$Y[k+1] = \mathbf{C}X[k+1] + V[k+1] \tag{A.14}$$

The random variables $W[k] \in \mathbb{R}^n$ and $R[k] \in \mathbb{R}^p$ represent the process and measurement noise (respectively). They are assumed to be independent (of each other), white, and with normal probability distributions respectively given by equations (A.15) and (A.16):

$$p(W) \sim N(0, \mathbf{Q}) \tag{A.15}$$

$$p(V) \sim N(0, \mathbf{R}) \tag{A.16}$$

In practice, the process noise covariance $\mathbf{Q}$ and measurement noise covariance $\mathbf{R}$ matrices might change with each time step or measurement, however here we assume they are constant.

Defining $\hat{X}[k+1|k]$ as the *a priori* state estimate at step $k+1$ given knowledge of the process at step k, and $\hat{X}[k+1|k+1]$ the *a posteriori* state estimate at step $k+1$ given measurement $Y[k+1]$. One can then define *a priori* and *a posteriori* estimation errors $e[k+1|k]$ and $e[k+1|k+1]$ as:

$$e[k+1|k] = X[k+1] - \hat{X}[k+1|k] \tag{A.17}$$

$$e[k+1|k+1] = X[k+1] - \hat{X}[k+1|k+1] \tag{A.18}$$

Estimation error covariance matrices for the *a priori* and *a posteriori* estimation errors, respectively $\hat{\mathbf{P}}$ and $\mathbf{P}$, can be obtained by application of the expectation operator in equations (A.19) and (A.20):

$$\hat{\mathbf{P}} = E(e[k+1|k]e^T[k+1|k]) \tag{A.19}$$

$$\mathbf{P} = E(e[k+1|k+1]e^T[k+1|k+1]) \tag{A.20}$$

The goal of the Kalman filter is to find an equation that computes an *a posteriori* state estimate $\hat{X}[k+1|k+1]$ as a linear combination of an *a priori* estimate $\hat{X}[k+1|k]$ and a weighted difference between an actual measurement $Y[k+1]$ and a measurement prediction as shown in equation (A.21):

$$\hat{X}[k+1|k+1] = \hat{X}[k+1|k] + \mathbf{K}(Y[k+1] - \mathbf{C}\hat{X}[k+1|k]) \tag{A.21}$$

The $n \times p$ matrix $\mathbf{K}$ in (A.21) is chosen to be the gain that minimizes the *a posteriori* error covariance. This minimization can be accomplished by first substituting (A.21) into equation (A.18), giving (A.22):

$$\mathbf{P} = E\left\{\left(X[k+1] - \hat{X}[k+1|k+1]\right)\left(X[k+1] - \hat{X}[k+1|k+1]\right)^T\right\} \tag{A.22}$$

Then, the expectation indicated in (A.20) must be performed to obtain (A.23):

$$\begin{aligned} \mathbf{P} &= E\{(X[k+1] - \hat{X}[k+1|k] - \mathbf{K}Y[k+1] + \mathbf{KC}\hat{X}[k+1|k]) \\ &\cdot\ (X[k+1] - \hat{X}[k+1|k] - \mathbf{K}(Y[k+1] + \mathbf{KC}\hat{X}[k+1|k])^T\} \end{aligned} \tag{A.23}$$

Substituting (A.14) in (A.23):

$$\begin{aligned} \mathbf{P} &= E\{(X[k+1] - \hat{X}[k+1|k] - \mathbf{KC}X[k+1] + \mathbf{K}V[k+1] + \mathbf{KC}\hat{X}[k+1|k]) \\ &\cdot\ (X[k+1] - \hat{X}[k+1|k] - \mathbf{KC}X[k+1] + \mathbf{K}V[k+1] + \mathbf{KC}\hat{X}[k+1|k])^T\} \end{aligned}$$

Factoring some common terms yields:

$$\begin{aligned} \mathbf{P} &= E\{((\mathbf{I_n} - \mathbf{KC})(X[k+1] - \hat{X}[k+1|k]) + \mathbf{K}V[k+1])\} \\ &\cdot\ \left((\mathbf{I_n} - \mathbf{KC})(X[k+1] - \hat{X}[k+1|k]) + \mathbf{K}V[k+1]\right)^T \end{aligned}$$

Taking the expectations, substituting equation (A.19) and remembering that the measurement error $V[k]$ is uncorrelated with $e[k+1|k]$, gives:

$$\mathbf{P} = \left((\mathbf{I_n} - \mathbf{KC})\hat{\mathbf{P}}(\mathbf{I_n} - \mathbf{KC})^T\right) + \mathbf{KRK}^T \tag{A.24}$$

As it is desired to minimize the trace of $\mathbf{P}$, which relates to the mean square error of the estimation, it then proceeds to taking the derivative with respect to $\mathbf{K}$, setting that result equal to zero, and then solving for $\mathbf{K}$ gives:

$$\frac{\partial \mathbf{P}}{\partial \mathbf{K}} = 0 = -2(\mathbf{I_n} - \mathbf{KC})\hat{\mathbf{P}}\mathbf{C}^T + 2\mathbf{KR}$$

$$-\hat{\mathbf{P}}\mathbf{C}^T + \mathbf{K}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C}^T + \mathbf{R}\right) = 0$$

$$\mathbf{K} = \hat{\mathbf{P}}\mathbf{C}^T\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C}^T + \mathbf{R}\right)^{-1} \tag{A.25}$$

The matrix **K** is also known as Kalman gain. By inspection of (A.25), it can be noted that as the measurement error covariance **R** approaches zero, the actual measurement $Y[k+1]$ is "trusted" more and more, while the predicted measurement $\mathbf{C}\hat{X}[k+1|k]$ is trusted less and less. Conversely, as the *a priori* estimate error covariance $\hat{\mathbf{P}}$ approaches zero the actual measurement is trusted less and less, while the predicted measurement is trusted more and more.

The expression of the optimum *a posteriori* estimate error covariance **P** when the optimum Kalman Gain K is calculated can be obtained by substituing (A.25) in (A.24):

$$\begin{aligned}\mathbf{P} &= \left((\mathbf{I_n} - \left(\hat{\mathbf{P}}\mathbf{C^T}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C^T} + \mathbf{R}\right)^{-1}\right)\mathbf{C})\hat{\mathbf{P}}(\mathbf{I_n} - \left(\hat{\mathbf{P}}\mathbf{C^T}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C^T} + \mathbf{R}\right)^{-1}\right)\mathbf{C})^T\right) \\ &+ \left(\hat{\mathbf{P}}\mathbf{C^T}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C^T} + \mathbf{R}\right)^{-1}\right)\mathbf{R}\left(\hat{\mathbf{P}}\mathbf{C^T}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C^T} + \mathbf{R}\right)^{-1}\right)^T\end{aligned}$$

This expression can be simplified to:

$$\mathbf{P} = \hat{\mathbf{P}} - \hat{\mathbf{P}}\mathbf{C^T}\left(\mathbf{C}\hat{\mathbf{P}}\mathbf{C^T} + \mathbf{R}\right)$$

$$\mathbf{P} = (\mathbf{I_n} - \mathbf{KC})\hat{\mathbf{P}} \tag{A.26}$$

It is important to notice that while equation (A.26) is valid only for optimum **K**, equation (A.24) represents the general case. This can have implications when there uncertainties due to unknown or time varying $\mathbf{R}$ and/or rounding and numerical errors due to poor conditioning.

Thus, employing equation (A.21) and substituting the calculated Kalman Gain $\mathbf{K}$ given by equation (A.25) leads to the optimum estimation of the system state $\hat{X}[k+1|k+1]$. However, at time step $k+2$, in order to incorporate the measurement $Y[k+2]$ into the state estimation, one will need the values of the estimation$\hat{X}[k+2|k+1]$ and the corresponding updated error estimation covariance $\hat{\mathbf{P}}$. The *a priori* estimation $\hat{X}[k+2|k+1]$ can be obtained from equation (A.13).

$$\hat{X}[k+2|k+1] = \mathbf{A}\hat{X}[k+1|k+1] + \mathbf{B}U[k+1] \tag{A.27}$$

Notice that the process noise $W[k+1]$ is omitted. This variable can be ignored because it has zero mean and its values are uncorrelated in time due to its normal distribution. Henceforth, the estimation error covariance matrix associated with $\hat{X}[k+2|k+1]$ is given by substitution of equation (A.27) into (A.17), and then in (A.19):

$$e[k+2|k+1] = X[k+2] - \hat{X}[k+2|k+1]$$

$$e[k+2|k+1] = (\mathbf{A}X[k+1] + \mathbf{B}U[k+1] + W[k+1]) - \left(\mathbf{A}\hat{X}[k+1|k+1] + \mathbf{B}U[k+1]\right)$$

$$e[k+2|k+1] = \mathbf{A}e[k+1|k+1] + W[k+1]$$

$$\hat{\mathbf{P}} \quad = \quad E(e[k+2|k+1]e^T[k+2|k+1])$$

$$\hat{\mathbf{P}} = E\left((\mathbf{A}e[k+1|k+1] + W[k+1])\,(\mathbf{A}e[k+1|k+1] + W[k+1])^T\right) \tag{A.28}$$

As $W[k+1]$ and $e[k+1|k+1]$ are uncorrelated, after simplifications the equation (A.28) can be rewritten as:

$$\hat{\mathbf{P}} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{Q} \tag{A.29}$$

In its classical rendition, the recursive Kalman filter algorithm is executed by the step-by-step evaluation of equations (A.27), (A.29), (A.25), (A.21) and (A.20). The first two expressions are dubbed as the time update equations, while the latter three are known as measurement update equations.

A.4 Variance tracking

One of the biggest challenges to Kalman filtering schemes is the determination of suitable values for the **Q** and **R** covariance terms. Previous knowledge of these parameters is seldom available, especially when the model does not represent a definite physical system. Phase III addresses this shortcoming in this proposed Kalman based predicting scheme. There is a recursive procedure that estimates the most probable value for **Q** and **R** at every time step.

In the first step, a R variance tracking routine was employed based on the estimation of V[k]. Isolating it in (A.14) gives the following expression:

$$V[k] = Y[k] - \mathbf{C}X[k] \tag{A.30}$$

It is then possible to estimate $V[k]$ by subtracting the predicted output $\mathbf{C}X[k]$ of the measured output $Y[k]$. By definition, **R** is the variance of $V[k]$ from the first to the kth time step. As the measurements are usually consequence of a very high number of stochastic process (errors in measurement systems, reading errors, random fluctuations), one can suppose that abrupt changes in statistical parameters of an isolated process does not necessarily translates into an abrupt change of the statistical parameters of the measurement process. As it is very unlikely that several of those stochastic processes will change in coordination, one can conclude that abrupt variations in the **R** parameter are also improbable. This approximate continuity is modelled in the tracking routine by weighing in the value of **R** estimated for the previous step, as shown in (A.31):

$$\mathbf{R}[k+1] = k^{-1}\mathbf{R}[k] + (k-1)k^{-1}Var(V[k]) \tag{A.31}$$

Where $Var(V[k])$ indicates the variance operator. In the second step, the estimation of the **Q** Covariance Matrix starts by isolating the $W[k]$ vector from its definition:

$$W[k] = X[\mathrm{k}] - \hat{X}[k] \tag{A.32}$$

Also by definition, **Q** is the covariance matrix of the vector $W[k]$. Considering also that **Q** does not change abruptly, a similar weighing routine is employed to determine it. However, $X[k]$ is a function of the Kalman Gain (A.21), which in its turn is a function of **R**. As by definition **Q** and **R** measure different model imperfections, they are thus modelled as independent variables and it is necessary to subtract the **R** variance from **Q** in the innovation $\triangle\mathbf{Q}$:

$$\triangle\mathbf{Q} = \sqrt{(Var(W[k])^2 - I_n \cdot Var(V[k])^2)} \tag{A.33}$$

$$\mathbf{Q}[k+1] = k^{-1}\mathbf{Q}[k] + (k-1)k^{-1}\triangle\mathbf{Q} \tag{A.34}$$

B The SPCTRL2 Radiative Transfer Model and SEDES2 Cloud Cover Modifier

This appendix contains a description and derivation of the Simple Solar Spectral Model (SPCTRL2) for Cloudless Atmospheres' and the SEDES2 empirical Cloud Cover Modifier, including:

1. An introduction;
2. Key concepts;
3. Direct Normal Irradiance;
4. Diffuse Irradiance;
5. Cloud Cover

Each of these topics will be described in sections B.1 to B.6.

B.1 Introduction

This introduction is comprised of some basic information about electromagnetic radiation, solar radiation, solar spectrum in Earth's surface and the relevant factors: sun's position and atmospheric composition.

Electromagnetic (EM) radiation is a form of transmitted energy, whose name arises from the electric and magnetic fields that simultaneously oscillate in planes mutually perpendicular to each other and to the direction of propagation through space, as shown in fig. B.1. Whenever charged particles are accelerated, EM waves are produced and can subsequently interact with any charged particles. EM waves carry energy, momentum and angular momentum away from their source particle and can impart those quantities to matter with which they interact.

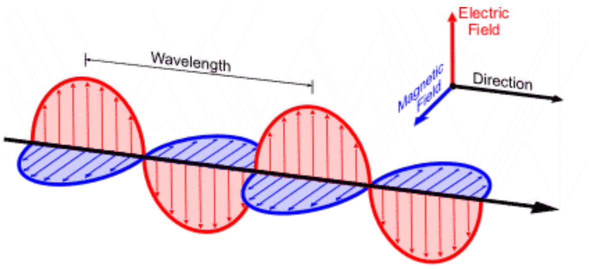


Figure B.1: Electromagnetic wave propagating from left to right. The electric field is in a vertical plane and the magnetic field in a horizontal plane. The electric and magnetic fields are always in phase and at 90 degrees to each other.

Electromagnetic radiation has a propagation speed of approximately 300.000 km/s, known as speed of light, constant and absolute for all referentials according to the theory of relativity. Due to the particle-wave duality, it can can also be described in terms of a stream of photons, massless particles traveling in a wave-like pattern at the speed of light. The larger the amount of photons, larger the energy flux. Also, each photon contains a certain amount of energy, which are related to the wavelenghts and define the different types of radiation. The set of all wavelengths define the electromagnetic spectrum, shown in fig. B.2. Radio waves have photons with low energies, microwave photons have a little more energy than radio waves, infrared photons have still more, then visible, ultraviolet, X-rays, and, the most energetic of all, gamma-rays. The irradiance of an EM wave source is defined as the received power per unit area at all wavelengths.

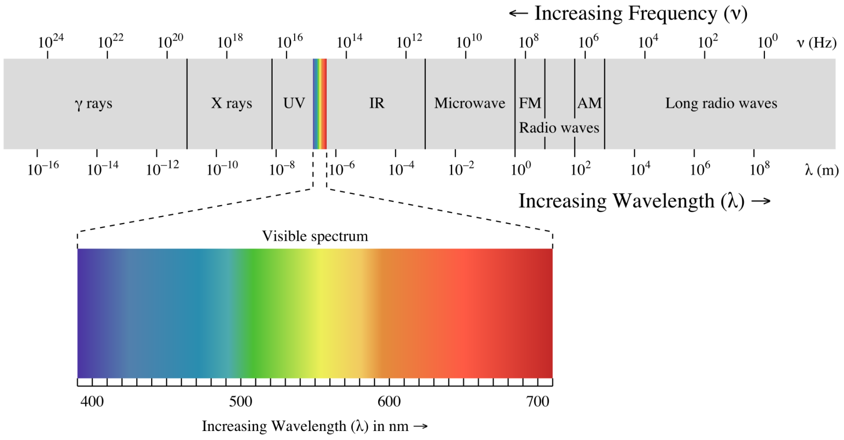


Figure B.2: Electromagnetic spectrum expressed in terms of energy and wavelength. In detail, the visible spectrum perceived by the human eyes as colors.

Most electromagnetic radiation from space is unable to reach the surface of the Earth, as shown in fig. B.3. Radio frequencies, visible light and some ultraviolet light makes it all the way to sea level. The higher the altitude, more rarefied the atmosphere and larger the fraction of the EM spectrum that becomes visible to instruments. This is the reason why many telescopes are built on mountain tops in order to better observe infrared wavelengths. Balloon experiments can reach 35 km above the surface and can operate for months. Rocket flights can take instruments all the way above the Earth's atmosphere, but only for a few minutes before they fall back to Earth. Satellite and spacecraft based instruments can access the entire EM spectrum for long term observations.

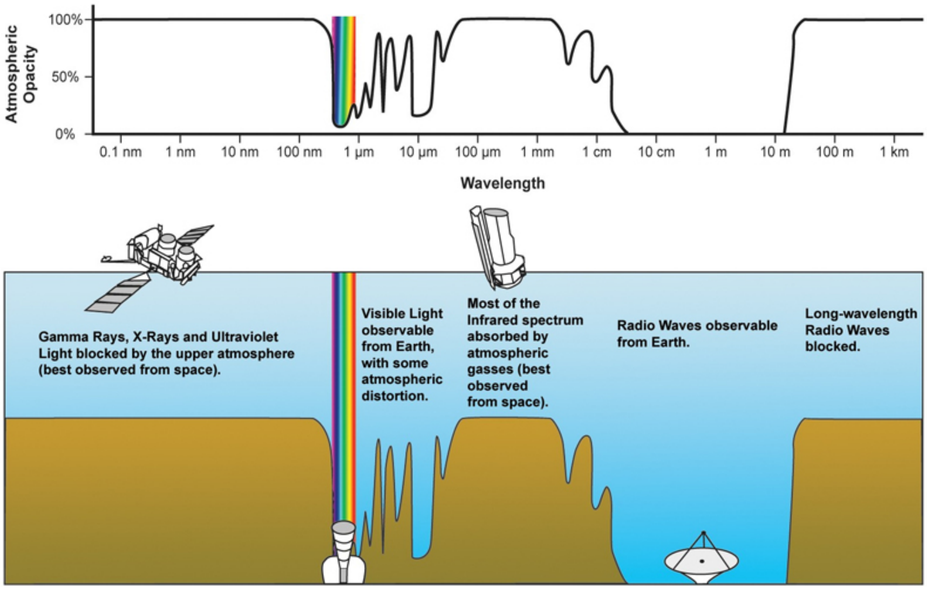


Figure B.3: Atmospheric Opacity as a function of the EM radiation wavelength. Note that the atmosphere is highly transparent to the visible spectrum.

By far the brightest object in the sky, the Sun is the main source of energy in Earth. Almost all energy sources harnessed by the human society have originated from the sunlight, except for the nuclear, geothermal and tidal plants. The sunlight powers photosynthesis, responsible for directly or indirectly feeding most lifeforms present in the planet. When stored in organic compounds such as hydrocarbonates, this energy can be chemically released through combustion, the basis of most thermoelectric units, fossil and biomass fueled. It globally creates temperature gradients that drives the atmospheric circulation, which can be converted in electricity by means of wind turbines. The sunlight also powers the water cycle, causing evaporation, clouds and rainfall, the drivers of hydroelectric generation. Finally, the solar radiation can be directly converted to heat and/or electricity by means of solar heaters, concentrating solar power plants and photovoltaic panels.

Observed in the space, outside the Earth's atmosphere, the sunlight spectrum is similar to that of a blackbody at approximately 5800 K, as shown in fig. B.4. At a distance of 1 A.U. (Astronomic Unit), which is the average radius of Earth's orbit, the sun's full spectrum irradiance at a perpendicular plane has been measured at 1.367 kW/m^2. Both spectrum and irradiance change dramatically when observed in Earth's surface, because the atmosphere scatters and absorbs EM radiation. The presence of clouds can further decrease or in some cases increase the incident irradiation, as they reflect sunlight away from or in direction to the observer.

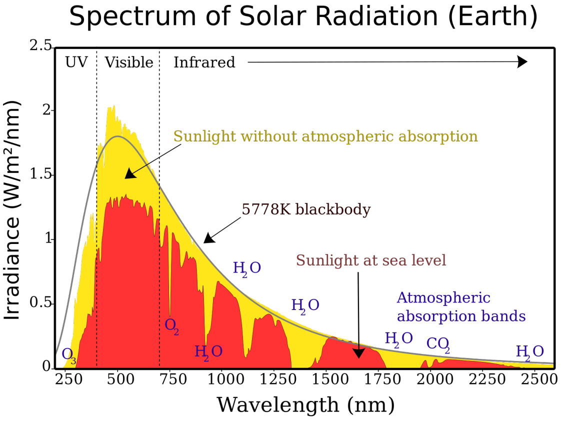


Figure B.4: Solar irradiance at space (yellow) and at sea level (red). For comparison, the gray line corresponds to the blackbody spectrum at 5778 K.

The atmospheric scattering is responsible for dividing the sunlight into two components: the direct and the diffuse solar radiation. As the name implies, the direct radiation is component imparted when the Sun is in line of sight. At zenith in a cloudless sky, this component's irradiance amounts to about 1.05 kW/m^2. The diffuse radiation is the component that does not travel in a straight line: its trajectory is changed after being scattered by molecules or aerosols in atmosphere. The amount of scattering is a function of the atmospheric composition, angle of incidence and wavelength. Indeed, the daytime sky is blue and sunsets/sunrises are red because air scatters short-wavelength light more than longer wavelengths.

As marked in fig. B.4, at some specific wavelenghts the sunlight is strongly absorbed by atmospheric constituents, such as molecular Ozone (O_3), Oxygen (O_2), Carbon Dioxide (CO_2) and Water (H_2O). These molecules absorb the sunlight's energy and later emit EM radiation at a random direction and longer wavelength, further attenuating the direct insolation and changing its spectral components.

As it determines the angle of incidence, the Sun's position in sky is a relevant factor when onde tries to measure the solar radiation's energy imparted over a surface. The

Sun path in the sky is mainly determined by the Earth's orbit, its rotation and axial tilt.

Slightly elliptical, Earth's distance to the Sun changes from circa 152 million kilometers at apoapsis (also known as aphelion) to 147 million at periapsis (perihelion). Due to the free space attenuation, the irradiance is inversely proportional to the distance squared, which corresponds to approximately 6.5% of variation between the orbital extremes.

Earth's axial tilt is responsible for the occurrence of the yearly seasons, as during summer it exposes the northern or southern hemispheres to the sunlight for more than 12 hours in a day, while the converse is true in the winter. At autumn and spring, both hemispheres receive approximately 12 hours of sunlight.

Earth's rotation makes the Sun to rise at the east and to set at the west, giving rise to the days and nights. In figure B.5 these facts are illustrated.

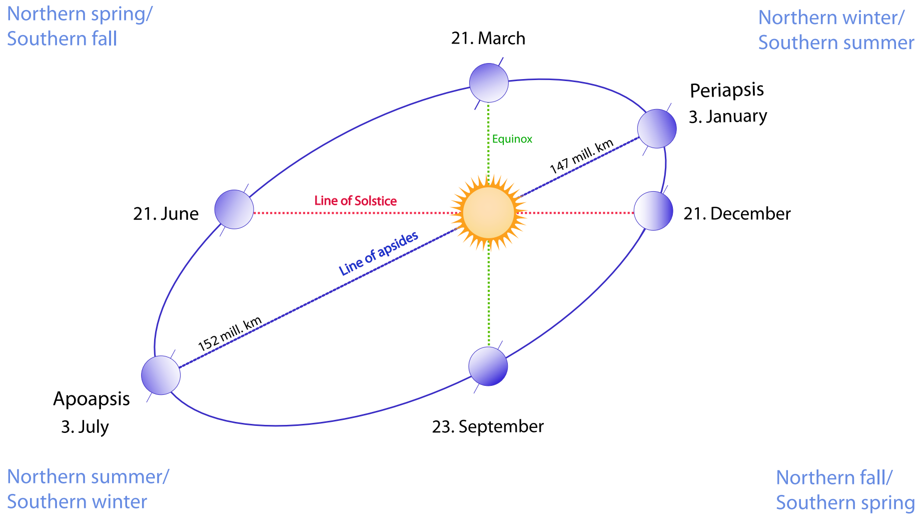


Figure B.5: Dates for seasons, apoapsis and periapsis of Earth's orbit. The elliptical form is exagerated.

Observed from Earth, the path of the Sun across the sky varies throughout the year. The shape described by the Sun's position, considered at the same time each day for a complete year, is called the analemma and resembles a "8" aligned along a North/South axis. While the most obvious variation in the Sun's apparent position through the year is a North/South swing over 47 degrees of angle (due to the 23.5-degree tilt of the Earth with respect to the Sun), there is an East/West component as well. The North/South swing in apparent angle is the main source of seasons on Earth. Figure B.6 plots a graph of the annalemma as seen in the Greenwich observatory.

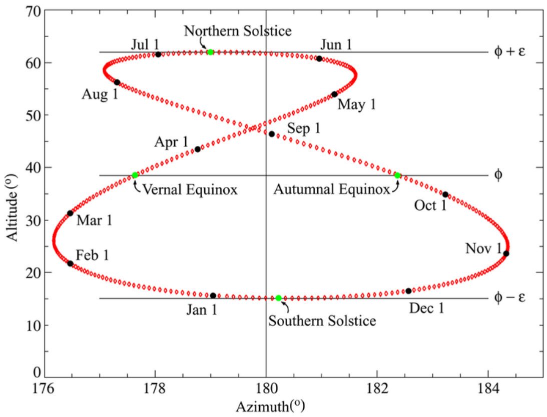


Figure B.6: Analemma plotted as seen at noon GMT from the Royal Observatory, Greenwich (latitude 51.48° north, longitude 0.0015° west).

Atmospheric radiative transfer models have significantly improved in the years. They can be classified as simple, moderately complex, and rigorous, depending on the balance between empirical and theoretical principles incorporated into them.

Complex, rigorous atmospheric transmission models such as MODTRAN are not appropriate for all applications, such as solar energy system engineering. A simpler parameterized or semi-empirical model can usually meet the user needs. Models have been published in the literature [35-40], based on the transmittance model of Leckner [41]. In particular, the SPCTRL2 model developed by Bird and colleagues at SERI/NREL [16], has been extensively distributed and evaluated [44].

SPCTRL2 relies on the product of empirical, closed-form transmission functions for the most important elements of atmospheric extinction: air molecules, ozone, water vapor, uniformly mixed gases, and aerosols. The product of the transmission functions modifies the extraterrestrial spectral direct beam irradiance to produce direct beam radiation. Simple theoretical relations are used to estimate the distribution of sky and ground reflected radiation. The model produces spectral results for 122 irregularly spaced wavelengths from 300 nm to 4000 nm. The equations are simple enough to be entered in personal computer spreadsheets, and can be quickly processed in mobile processors.

B.2 Key concepts

As used in this appendix this section provides definitions and explanations for several key concepts to the understanding of a solar irradiance and atmospheric radiative transfer model.

Albedo(r_g)—The albedo of a surface is the ratio of radiation reflected from the surface to the incident radiation. Its dimensionless nature lets it be expressed as a percentage and is measured on a scale from zero (no reflection) of a perfectly black surface to 1 for perfect reflection of a white surface. Because albedo is the ratio of all reflected radiation to incident radiation, it will include both the diffuse and direct radiation reflected from an object. Figure B.7 shows the annual clear sky and total Earth albedo as measured by the Ceres Probe in 2003 and 2004.

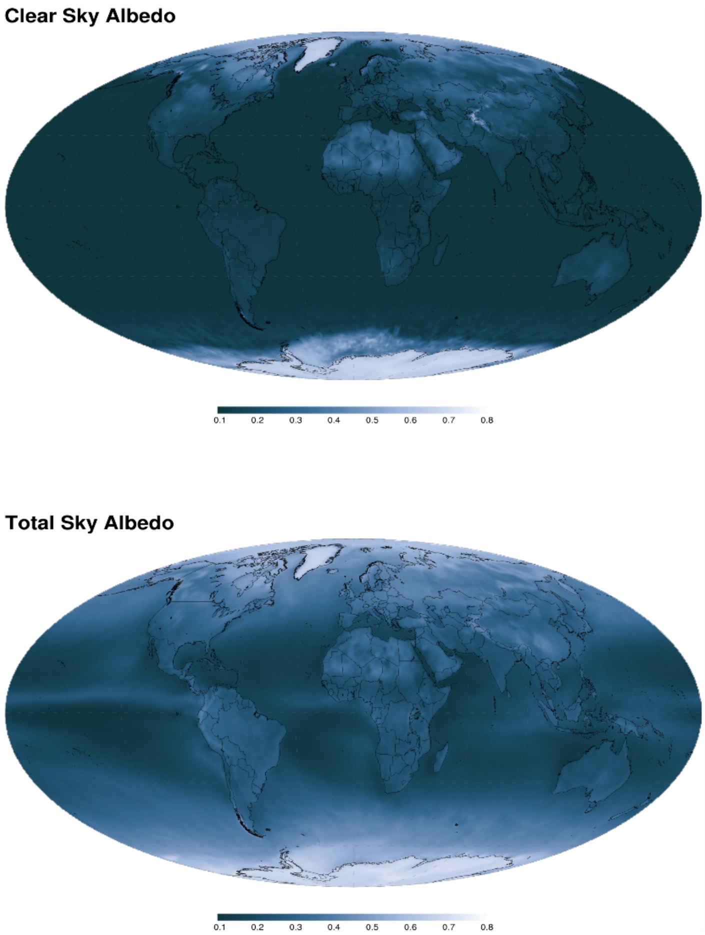


Figure B.7: CERES-Aqua 2003-2004 mean annual clear sky and total sky albedo. Clear sky albedo is the fraction of the incoming solar radiation that is reflected back into space by regions of the Earth on cloud-free days. Total sky albedo include cloudy days. Data source: http://daac.gsfc.nasa.gov/giovanni/

Aerosol Optical Depth (AOD, $\tau_{a\lambda}$)—(also called “optical thickness” or “turbidity”) the wavelength-dependent total extinction (scattering and absorption) by aerosols in the atmosphere. AOD at 500 nanometers (nm) is commonly reported.

Air mass (AM)—Ratio of the mass of the atmosphere in the actual sun-observer path to the mass that would exist if the sun were directly overhead.

Relative air mass (AMR)— AMR is the ratio of the observed path length through the atmosphere to the path length through the atmosphere directly overhead. AMR varies as secant of the zenith angle, Z.

Absolute Air Mass (AMA)— AMA varies with the zenith angle and local barometric pressure, P. Using P_0 to indicate standard atmospheric pressure, it is calculated by (B.1).

$$AMA \approx \frac{P}{P_0} \sec(Z) \tag{B.1}$$

Air mass zero (AM0)—solar radiation quantities outside the Earth's atmosphere at the mean Earth-Sun distance (1 Astronomical Unit).

Azimuth Angle (A)— The azimuth angle is an angular measurement in a spherical coordinate system. The azimuth is the angle formed between a reference direction (usually north or south) and a line from the observer to a point of interest projected on the same plane as the reference direction orthogonal to the zenith. For an observer in Earth surface, the azimuth angle of the Sun defines its direction as projected over the ground plane. An diagram showing Azimuth, Zenith and Tilt angles is shown in figure B.8.

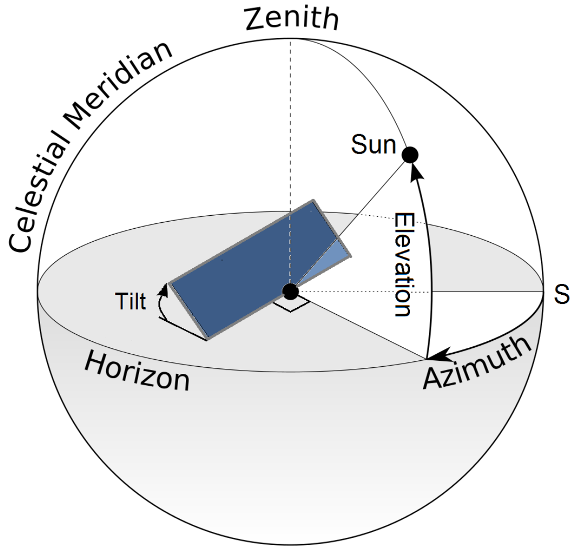


Figure B.8: Solar Azimuth and Elevation (complement of Zenith Angle) angles. Panel Azimuth and tilt angles. Azimuth reference is the geographical south pole.

Circumsolar radiant energy—radiation scattered by the atmosphere from an area of the sky immediately adjacent to the sun, the solar aureole.

Diffuse solar irradiance, diffuse, I_s —downward scattered solar flux received on a horizontal surface from a solid angle of 2π-steradian (hemisphere) with the exception of a conical solid angle with a 100 $mrad$ (approximately 6°) included plane angle centered on the sun's disk Figure B.9 displays a photography that depicts Diffuse, Direct and Global irradiance.

Direct solar irradiance, direct, I_d —solar flux coming from the solid angle of the sun's disk on a surface perpendicular to the axis of that solid angle. Also referred to as "direct normal irradiance". Figure B.9 displays a photography that depicts Diffuse, Direct and Global irradiance.

Global or Hemispherical Irradiance (GHI), I —the solar radiant flux received from within the 2π steradian field of view of a given plane from the portion of the sky dome and the foreground included in the plane's field of view, including both diffuse and direct solar radiation. For the special condition of a horizontal plane the hemispherical solar irradiance is properly termed global solar irradiance, I_H. The adjective global should refer only to hemispherical solar radiation on a horizontal surface. Figure B.9 displays a photography that depicts Diffuse, Direct and Global irradiance.

Integrated irradiance $I_{\lambda1-\lambda2}$—spectral irradiance integrated over a specific wavelength interval from λ_1 to λ_2, measured in Wm^{-2}.

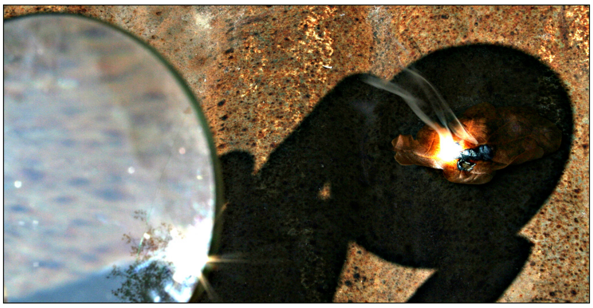


Figure B.9: Burning a dry leaf with a magnifier lens. The bright spot over the smoking leaf is concentrated Direct irradiation of the Sun. The daylit ground is illuminated by the Global irradiance. The shadowed areas are dimly illuminated by the Diffuse irradiation component. Photography credits go to Dave Gough, available at https://www.flickr.com/photos/spacepleb/1505372433 (CC BY 2.0 license)

Rayleigh Scattering— the process of elastic scattering of light or other electromagnetic radiation by particles much smaller than the wavelength of the radiation. The particles may be individual atoms or molecules, and results from the electric polarizability of the particles. It can occur when light travels through transparent solids and liquids, but is most prominently seen in gases. The oscillating electric field of a light wave acts on the charges within a particle, causing them to move at the same frequency. The particle therefore becomes a small radiating dipole whose radiation we see as scattered light.

Solar constant—the total solar irradiance at normal incidence on a surface in space (AM0) at the earth's mean distance from the sun. (1 astronomical unit, or AU = 1.496 x 10^{11} m). The current accepted value of the solar constant is 1366.1± $7Wm^{-2}$ [13]. The AM0 solar flux at the Earth varies by ±3.5% about the solar constant as the earth-sun distance varies through the year, and with the solar sunspot activity.

Spectral solar irradiance, I_λ—solar irradiance I per unit wavelength interval at a given wavelength λ (unit: Watts per square meter per nanometer, $Wm^{-2}nm^{-1}$)

Spectral passband— the effective wavelength interval within which spectral irradiance is considered to pass, as through a filter or monochromator. The convolution integral of the spectral passband (normalized to unity at maximum) and the incident spectral irradiance produces the effective transmitted irradiance. Spectral passband may also be referred to as the spectral bandwidth of a filter or device. Passbands are specified as the interval between wavelengths at which one half of the maximum transmission of the filter or device occurs, or as Full-Width at Half-Maximum, FWHM.

Spectral interval—the distance in wavelength units between adjacent spectral irradiance data points.

Spectral resolution—the minimum wavelength difference between two wavelengths that can be unambiguously identified.

Tilt Angle (T)— The angle between the ground plane and an inclined surface. As such, the tilt angle is zero for a horizontal surface and 90° for a vertical surface. An diagram showing Azimuth, Zenith and Tilt angles is shown in figure B.8.

Total precipitable water—depth of a column of water with a section of 1 cm^2 equivalent to the condensed water vapor in a vertical column from the ground to the top of the atmosphere. (Unit: $atm - cm$ or g/cm^2)

Total ozone— depth of a column of ozone equivalent to the total of the ozone in a vertical column from the ground to the top of the atmosphere. (Unit: $atm - cm$)

Total nitrogen dioxide— depth of a column of pure nitrogen dioxide (NO_2) equivalent to the total of the NO_2 in a vertical column from the ground to the top of the atmosphere. (Unit: $atm - cm$)

Wavenumber— a unit of frequency, ν, in units of reciprocal centimeters (symbol cm^{-1}) commonly used in place of wavelength, λ. The relationship between wavelength and frequency is defined by $\lambda\nu = c$, where c is the speed of light in vacuum. To convert wavenumber to nanometers, λ nm $= 10^7/\nu\ cm^{-1}$.

Zenith Angle (Z)— For an observer in Earth surface, the zenith angle is the angular distance between a point in the sky and the zenith, which is an imaginary point directly above a particular location. It is the complement of the elevation angle. An diagram showing Azimuth, Zenith and Tilt angles is shown in figure B.8.

B.2.1 Local solar position

For an observer in Earth surface, the Sun's position in the sky is completely determined by its zenith and azimuth angles.

The zenith angle Z is calculated from the expression presented in (B.2):

$$Z = \cos^{-1}\left(\cos(\phi)\cos(\delta)\cos(\omega) + \sin(\phi)\sin(\delta)\right) \tag{B.2}$$

where ϕ is the site latitude, δ the Sun declination and ω the true local solar time, all angles in radians. The true local solar time takes into account the local timezone and the difference between the apparent solar time (sundial time) and the mean solar time (equally spaced noons by 24 hours), and is calculated from the equation of time E, the site longitude in degrees ψ_D and the local time t (in hours past midnight and fractions) by means of expression (B.3). The term t_z represent the time zone, in hours to be added to Greenwich Meridian Time (GMT) to obtain the standard local time.

$$\omega = E + \frac{\pi}{12}\left(t + \frac{\psi_D}{15} - t_z\right) \tag{B.3}$$

The equation of time describes the discrepancy between apparent solar time (sundial time) and the mean solar time, and is approximated by equation (B.4):

$$E = a_0 + a_1\cos(\varphi) + b_1\sin(\varphi) + a_2\cos(2\varphi) + b_2\sin(2\varphi) \tag{B.4}$$

The Sun's declination is calculated approximately by the equation (B.5), which is a truncated Fourier series:

$$\delta = a_0 + a_1\cos(\varphi) + b_1\sin(\varphi) + a_2\cos(2\varphi) + b_2\sin(2\varphi) + a_3\cos(3\varphi) + b_3\sin(3\varphi) \tag{B.5}$$

where the constants a_0 to a_3 and b_1 to b_3 for the Equation of Time (E) and Sun Declination (δ) are shown in Table B.1:

Table B.1: Coefficients for the Equation of Time and for the Sun declination

i	Equation of Time (E)	Declination (δ)
a_0	0.000075	0.006918
a_1	0.001868	- 0.399912
b_1	- 0.032077	0.070257
a_2	- 0.014615	- 0.006758
b_2	- 0.040849	0.000907
a_3	0	- 0.002697
b_3	0	0.00148

The day angle φ represent the position of the Sun relative to stars. In radians, φ is a function of the day number d of the year (from 1 to 365), represented by (B.6):

$$\varphi = \frac{2\pi(d-1)}{365} \tag{B.6}$$

The azimuth angle A is a function of the site latitude ϕ, the Sun declination δ and the true local solar time ω. It is calculated by equation ():

$$A = \mathrm{ATAN2}\left(\cos(\omega)\sin(\phi) - \cos(\phi)\tan(\delta), \sin(\omega)\right) \tag{B.7}$$

where ATAN2(x, y) denotes the four quadrant arctangent function, which gives the arc tangent of y/x, taking into account which quadrant the point (x, y) is in.

B.3 Direct Normal Irradiance

The direct irradiance $I_{d\lambda}$ on a surface normal to the direction of the sun at ground level for wavelength λ is modelled by equation (B.8):

$$I_{d\lambda} = H_{0\lambda} D T_{r\lambda} T_{a\lambda} T_{w\lambda} T_{O\lambda} T_{u\lambda} \tag{B.8}$$

The parameter $H_{0\lambda}$ is the extraterrestrial irradiance at the average Earth-Sun distance for wavelength λ, D is the correction factor that accounts for variations in this distance due to the elliptical nature of Earth's orbit, while the other parameters are related to the transmitance factors of the atmosphere at wavelength λ due to five relevant effects. $T_{r\lambda}$ is the transmittance function for molecular (Rayleigh) scattering, $T_{a\lambda}$ for aerosol scattering, $T_{w\lambda}$ is the function for water vapor absorption, $T_{O\lambda}$ for Ozone absorption and $T_{u\lambda}$ for uniformly mixed gas absorption. Thus, equation (B.8) models the direct irradiance for wavelength λ for a surface directly pointed to the Sun.

In order to obtain the direct irradiance I_d on a horizontal surface, one must consider the zenith angle Z as in equation (B.9):

$$I_d = I_{d\lambda} \cos(Z) \tag{B.9}$$

The extraterrestrial spectral irradiance employed is the same used by the SPCTRL2, as illustrated in fig. B.10. It is composed of 122 irregularly spaced wavelengths from 300 nm to 4000 nm. It is based on the standard spectrum presented in [42].

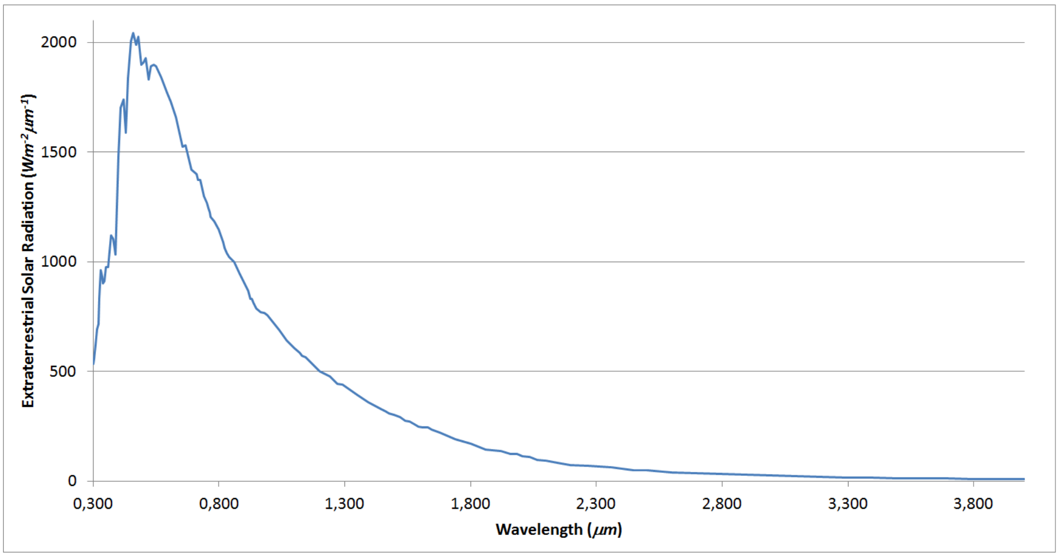


Figure B.10: SPCTRL2 Extraterrestial Solar Radiation.

These values are valid when Earth is exactly at the average orbital distance to the Sun. This only happens twice a year. In order to correct the variation due to the elliptical orbit, reference [92] indicates the following distance factor D, derived from Fourier series approximation, also a function of day angle φ:

$$D = 1.00011 + 0.034221\cos(\varphi) + 0.00128\sin(\varphi) + 0.000719\cos(2\varphi) + 0.000077\sin(2\varphi) \tag{B.10}$$

B.3.1 Rayleigh Scattering

Reference [62] provides an expression to calculate the atmospheric transmittance after Rayleigh scattering:

$$T_{r\lambda} = \exp\left(-M'\lambda^2 \left|115.6406\lambda^2 - 1.3366\right|\right) \tag{B.11}$$

where $M'$is the pressure-corrected air mass, which is a function of the surface atmospheric pressure P and zenith angle Z. Given P_0 as the sea level atmospheric pressure, the relative air mass as calculated by reference [58] is:

$$M' = \frac{P}{P_0\left|\cos(Z) + 0.15(93.885 - Z)^{-1.253}\right|} \tag{B.12}$$

The relative air mass M is obtained if the pressure correction is not applied in (B.12):

$$M = \frac{1}{\left|\cos(Z) + 0.15(93.885 - Z)^{-1.253}\right|} \tag{B.13}$$

B.3.2 Aerosol Scattering and Absorption

The aerosol transmittance is a function of atmospheric aerosol turbidity $\tau_{a\lambda}$ and the relative air mass M, as given by equation (B.14):

$$T_{a\lambda} = \exp\left(-\tau_{a\lambda} M\right) \tag{B.14}$$

where $\tau_{a\lambda}$ is calculated by:

$$\tau_{a\lambda} = \beta_n \lambda^{-\alpha_n} \tag{B.15}$$

In the SPCTRL2 model, the aerosol transmittance is modeled as a piecewise biexponential function. Hence, two α_n are used: $\alpha_1 = 1.0274$ if $\lambda < 0.5\,\mu m$ and $\alpha_2 = 1.2060$ otherwise. Two parameteres β_n are then appropriately chosen for each wavelength in order to match the turbidity values at $\lambda = 0.5\,\mu m$ as calculated by (B.15) with α_1 and α_2.

B.3.3 Water Vapor, Ozone and Uniformly Mixed Gas Absorption

SPCTRL2 adopts the water vapor transmittance expression derived in [65], which has the form:

$$T_{w\lambda} = \exp\left(\frac{-0.2385 a_{w\lambda} W M}{\left(1 + 20.07 a_{w\lambda} W M\right)^{0.45}}\right) \tag{B.16}$$

where W is the precipitable water vapor over a vertical column of atmosphere in cm and $a_{w\lambda}$ is the water vapor absorption coefficient as a function of wavelength. The SPCTRL2 model, however does not employ every value of $a_{w\lambda}$ as tabulated in [65]. They use an adjusted set for this parameter, in order to improve agreement with rigorous atmospheric transfer models [16].

Similarly the expression derived in [65] is used to model Ozone transmittance equation:

$$T_{O\lambda} = \exp\left(-a_{O\lambda} O_3 M_O\right) \tag{B.17}$$

where $a_{O\lambda}$ is the ozone absorption coefficient as a function of wavelength and the ozone mass M_O. Reference [53] gives an expression for determining of M_O as a function of the zenith angle Z and the height of maximum ozone concentration h_O.

$$M_O = \frac{1 + \frac{h_O}{6370}}{\left(\cos^2(Z) + \frac{2h_O}{6370}\right)^{0.5}} \tag{B.18}$$

In absence of direct measurements of the ozone ammount O_3 in $atm{-}cm$, the SPCTRL2 model employs the Van Heuklon models [48]. More recent research has updated some of its parameters [57].

The expression for the transmittance of uniformly mixed gas is given by [65]:

$$T_{u\lambda} = \exp\left(\frac{-1.41 a_{u\lambda} M'}{\left(1 + 118.3 a_{w\lambda} M'\right)^{0.45}}\right) \tag{B.19}$$

where $a_{u\lambda}$ is a combined gaseous amount and absorption coefficient.

B.4 Diffuse Irradiance

The diffuse irradiance is difficult to determine accurately with the simple parameterization methods that were used to calculate direct normal irradiance in the previous section. The SPCTRL2 model uses tabulated correction factors to make the simple formulation for the diffuse irradiance presented in [96] match the results from a rigorous radiative transfer code. The correction factors are adjusted versions of those presented in the formulations shown in [55], which have changed the diffuse formulation and obtained reasonable agreement with rigorous code results without using tabulated correction factors.

The SPCTRL2 simplifies the computation of diffuse irradiance by dividing it in three independent terms: the Rayleigh scattering component $I_{r\lambda}$, the aerosol scattering component $I_{a\lambda}$, and the component that accounts for multiple reflection of irradiance between the ground and the air $I_{g\lambda}$. The scattered (diffuse) irradiance $I_{s\lambda}$ on a horizontal surface is given by the summation of these terms.

$$I_{s\lambda} = \left(I_{r\lambda} + I_{a\lambda} + I_{g\lambda}\right) C_S \tag{B.20}$$

The wavelength dependent correction term C_S employed in SPCTRL2 is given by equation (B.21):

$$C_S = \begin{cases} (\lambda + 0.55)^{1.8}\,; & \lambda \leq 0.45\,\mu m \\ 1.0 & \lambda > 0.45\,\mu m \end{cases} \tag{B.21}$$

B.4.1 Rayleigh scattering term

The Rayleigh scattering term $I_{r\lambda}$is calculated by means of equation (B.22), as a function of the Extraterrestrial irradiation $H_{0\lambda}$, the correction factor D, the zenith angle Z and the atmospheric transmittances defined in Section B.3.

$$I_{r\lambda} = H_{0\lambda} D \cos(Z) T_{w\lambda} T_{O\lambda} T_{u\lambda} T_{aa\lambda} \frac{(1 - T_{r\lambda}^{0.95})}{2} \tag{B.22}$$

where $T_{aa\lambda}$ is the aerosol absorptance transmittance component, determined by equation (B.23):

$$T_{aa\lambda} = \exp\left(-\omega_\lambda \tau_{a\lambda} M\right) \tag{B.23}$$

where in turn, $\tau_{a\lambda}$ is defined in (B.15), M in (B.13), and ω_λ is the aerosol single scattering albedo, given by (B.24):

$$\omega_\lambda = \omega_{0.4} \exp\left(-\omega' \ln^2\left(\frac{\lambda}{4}\right)\right) \tag{B.24}$$

$\omega_{0.4}$ is the single scattering albedo at 0.4 μm wavelength and ω' is the wavelength variation factor, which for the standard rural aerosol model are respectivelly equal to 0.945 and 0.095.

B.4.2 Aerosol scattering term

The aerosol scattering term $I_{a\lambda}$is calculated by means of equation (B.25):

$$I_{a\lambda} = H_{0\lambda} D \cos(Z) T_{w\lambda} T_{O\lambda} T_{u\lambda} T_{aa\lambda} T_{r\lambda}^{1.5} (1 - T_{as\lambda}) F_S \tag{B.25}$$

where $T_{as\lambda}$ is the aerosol scattering transmittance component calculated by (B.26) and F_S is the ratio of forward to total scattering calculated by (B.27).

$$T_{as\lambda} = \exp\left(-(1-\omega_\lambda)\tau_{a\lambda}M\right) \tag{B.26}$$

$$F_S = 1 - \frac{\exp\left((AFS + BFS\cos(Z))\cos(Z)\right)}{2} \tag{B.27}$$

Note that $T_{a\lambda} = T_{as\lambda}T_{aa\lambda}$. The terms AFS and BFS are related to the asymetric nature of aerosol scattering and calculated by (B.28) and (B.29):

$$AFS = ALG(1.459 + ALG(0.1595 + 0.4129ALG)) \tag{B.28}$$

$$BFS = ALG(0.0783 + ALG(-0.3824 - 0.5874ALG)) \tag{B.29}$$

where ALG is a function of the aerosol symmetry factor $ASYM$, whose typical value in rural model is 0.65:

$$ALG = \ln(1 - ASYM) \tag{B.30}$$

B.4.3 Ground and sky reflectance term

The ground and sky reflectance term accounts for multiple reflection of irradiance between the ground and the air. It is modeled as a function of the direct irradiation $I_{d\lambda}$, Rayleigh scattering component, aerosol scattering component, the ground albedo $r_{g\lambda}$ and the sky reflectivity $r_{s\lambda}$, as shown in equation (B.31).

$$I_{g\lambda} = \frac{\left(I_{d\lambda}\cos(Z) + I_{r\lambda} + I_{a\lambda}\right)r_{s\lambda}r_{g\lambda}}{1 - r_{s\lambda}r_{g\lambda}} \tag{B.31}$$

The ground albedo $r_{g\lambda}$ depends on several factors, such as the surface material composition, wavelength, state of motion (if it is a liquid surface), the angle of incidence of the multiple irradiation components, temperature (for some materials), and others. Consequently, it is very hard and usually not feasible to strictly model the albedo coefficient in a given point at all directions and sun positions. The SPCTRL2 employs tabulated values of wavelength independent typical ground albedo as measured in different environments, as listed in Table B.2.

Table B.2: Typical sample values of Albedo for different surfaces/enviroments

Surface	Typical Albedo
Fresh asphalt	0.04
Open ocean	0.06
Worn asphalt	0.12
Conifer forest (Summer)	0.09 to 0.15
Deciduous trees	0.15 to 0.18
Bare soil	0.17
Green grass	0.25
Desert sand	0.40
New concrete	0.55
Urban Enviroment	0.10 to 0.45 (typ. 0.25)
Ocean ice	0.5–0.7
Fresh snow	0.80–0.90

The sky reflectivity $r_{s\lambda}$ is calculated as the sum of the Rayleigh reflectance and the aerosol reflectance, as shown in equation (B.32).

$$r_{s\lambda} = T'_{w\lambda} T'_{O\lambda} T'_{aa\lambda} \left[0.5\left(1 - T'_{r\lambda}\right) + \left(1 - F'_S\right) T'_{r\lambda}\left(1 - T'_{as\lambda}\right)\right] \tag{B.32}$$

The primed atmospheric transmitance terms $T'_{w\lambda}$, $T'_{O\lambda}$, $T'_{aa\lambda}$, $T'_{r\lambda}$ and $T'_{as\lambda}$ are the regular terms evaluated at $M = 1.8$. Likewise, the primed ratio of forward to total scattering F'_S is calculated by equation (B.33):

$$F'_S = 1 - \frac{\exp\left(\left(AFS + \frac{BFS}{1.8}\right)\frac{1}{1.8}\right)}{2} \tag{B.33}$$

B.5 Global irradiance on tilted surfaces

The direct and diffuse componentes calculated in Sections B.3 and B.4 model the global irradiance over a horizontal surface. Using these two irradiation components, the SPCTRL2 model calculates the global irradiance I over a tilted surface for any given Sun position.

The spectral global irradiance on an tilted surface is represented by the expression shown in (B.34):

$$I_\lambda = I_{d\lambda} \cos\left(\theta\right) + I_{s\lambda} \left\{ \frac{I_{d\lambda} \cos\left(\theta\right)}{H_{0\lambda} D \cos\left(Z\right)} + \left[\left(\frac{1 + \cos\left(T\right)}{2} \right) \left(1 - \frac{I_{d\lambda}}{H_{0\lambda} D} \right) \right] \right\} + \frac{\left(I_{d\lambda} + I_{s\lambda}\right) r_{g\lambda} \left(1 - \cos\left(T\right)\right)}{2} \tag{B.34}$$

The angle of incidence θdepends on the solar zenith angle Z, tilt angle T, Sun azimuth A and surface azimuth A_φ, as shown in equation (B.35):

$$\theta = \cos^{-1}\left(\cos(Z)\cos(T) + \sin(Z)\cos(A - A_\varphi)\sin(T)\right) \tag{B.35}$$

B.6 Cloud cover modifiers

The SPCTRL2 radiative transfer model provides a value for global irradiance on tilted surfaces. However, this specific model is only accurate when there are no clouds in the sky. Solar radiation is attenuated, further scattered and even reflected to the surface of interest by the presence of clouds in the sky. This complex and difficult to model process depends on the type of cloud, their thickness, and the number of cloud layers.

The online measurement and prediction of the instantaneous cloud cover effect over the global horizontal irradiance requires the use of expensive and extensive sensors, such as all sky imagers, IR-Visible-UV cameras, weather radars and satellite imaging. This difficulty arises because scattering and reflections are geometrically dependent, which in turn are a consequence of the position, movement, depth and formation rate of all clouds present in the visible sky.

However, the problem is simplified when a average modifier is required to model the cloud effects over a time period, such as a minute, a hour or a day. Parametrical models have been developed to simulate the cloud cover using simpler measurements, such as Clearness index or the related Sky cloud relative coverage, the latter given in the METARs provided by airports stations. A widely used model is the SEDES2. Based on solar resource measurements taken in the SEDES data acquisition center, a remote monitoring station of the Centre for Solar Energy and Hydrogen Research (ZSW) from Germany, this Cloud Cover Modifier (CCM) accounts for the effects of clouds by transforming the clear sky's spectral global irradiance, using empirically determined coefficients [77]. These modifiers use a quadratic equation with the clearness index K_t and six empirically derived constants, as shown in (B.36):

$$I_{C\lambda} = \left[A1_\lambda + \frac{A2_\lambda}{\cos(Z)} + \left(B1_\lambda + \frac{B2_\lambda}{\cos(Z)}\right) K_t + \left(C1_\lambda + \frac{C2_\lambda}{\cos(Z)}\right) K_t^2\right] I_\lambda \tag{B.36}$$

where the clearness index is defined as the ratio between the Global Horizontal Irradiance (GHI) and the extraterrestrial irrandiance H_0 projected over the surface area:

$$K_t = \frac{GHI}{H_0 \cos(Z)} \tag{B.37}$$

The coefficients $A1_\lambda$, $A2_\lambda$, $B1_\lambda$, $B2_\lambda$ $C1_\lambda$ and $C2_\lambda$ are wavelength dependent and have been empirically determined. Their values are shown in Tables B.3, B.4 and B.5.

Table B.3: SEDES2 Coefficients by wavelenght (1st part)

Wavelength (nm)	$A1_\lambda$	$A2_\lambda$	$B1_\lambda$	$B2_\lambda$	$C1_\lambda$	$C2_\lambda$
320	1,28572	0,30679	-0,29613	-0,58516	0,02063	0,20915
330	1,2351	0,26201	-0,28377	-0,53864	0,01073	0,20649
340	1,20617	0,2502	-0,25258	-0,51989	0,00432	0,20461
350	1,13974	0,24268	-0,19222	-0,49821	-0,01184	0,20133
360	1,09164	0,24421	-0,13386	-0,48722	-0,0272	0,20077
370	1,03373	0,2515	-0,07915	-0,48133	-0,04285	0,20297
380	0,99718	0,24386	-0,0655	-0,45039	-0,03607	0,19192
390	0,99795	0,2275	-0,08976	-0,40715	-0,01039	0,17371
400	0,99057	0,2054	-0,12091	-0,35735	0,01808	0,15208
410	0,98402	0,19311	-0,13671	-0,32748	0,0344	0,1407
420	0,97139	0,17787	-0,15584	-0,29288	0,05175	0,12755
430	0,97645	0,1594	-0,18434	-0,25421	0,07213	0,11271
440	0,9732	0,14208	-0,20773	-0,21836	0,08869	0,09857
450	0,97979	0,12932	-0,22806	-0,19197	0,10337	0,08717
460	0,98578	0,11921	-0,24438	-0,1714	0,11745	0,07671
470	0,99861	0,10918	-0,26163	-0,15113	0,1326	0,06607
480	1,00532	0,09968	-0,27866	-0,13004	0,14722	0,05576
490	1,01968	0,08958	-0,30482	-0,10709	0,16626	0,04513
500	1,0244	0,08052	-0,32229	-0,0875	0,17951	0,03647
510	1,03159	0,06907	-0,34795	-0,06441	0,19687	0,02547
520	1,04937	0,05644	-0,38233	-0,04055	0,21881	0,01373
530	1,06394	0,04632	-0,40907	-0,02121	0,23612	0,0042
540	1,07155	0,0383	-0,42769	-0,00587	0,24841	-0,00299
550	1,07039	0,03185	-0,43045	0,00449	0,25183	-0,00768
560	1,06283	0,02634	-0,41879	0,012	0,24665	-0,01046
570	1,04584	0,02469	-0,37226	0,00943	0,22308	-0,00801
580	1,03747	0,02347	-0,33927	0,00897	0,20751	-0,0069
590	1,02608	0,0233	-0,3141	0,00815	0,19573	-0,00518

Table B.4: SEDES2 Coefficients by wavelenght (2nd part)

Wavelength (nm)	$A1_\lambda$	$A2_\lambda$	$B1_\lambda$	$B2_\lambda$	$C1_\lambda$	$C2_\lambda$
600	1,04038	0,01568	-0,34917	0,02434	0,21889	-0,01426
610	1,05082	0,00666	-0,38518	0,04176	0,24156	-0,02411
620	1,05164	0,00029	-0,39171	0,05103	0,24639	-0,02902
630	1,04029	-0,00264	-0,36449	0,05087	0,23063	-0,02769
640	1,04091	-0,00243	-0,35577	0,05171	0,22554	-0,02653
650	1,04068	-0,00316	-0,34746	0,05376	0,22107	-0,02611
660	1,06505	-0,00775	-0,38644	0,0686	0,24625	-0,0347
670	1,08171	-0,0102	-0,40061	0,07729	0,25748	-0,04034
680	1,07724	-0,00697	-0,36968	0,07159	0,24056	-0,03716
690	1,04041	-0,00413	-0,28523	0,05231	0,18754	-0,02455
700	1,01641	-0,00067	-0,23359	0,03604	0,15018	-0,01227
710	1,00652	-0,00416	-0,21335	0,03074	0,13058	-0,00725
720	1,01501	-0,00986	-0,20643	0,03345	0,12001	-0,00709
730	1,11212	-0,03985	-0,3703	0,0868	0,19893	-0,03506
740	1,25964	-0,07938	-0,63633	0,16789	0,33604	-0,08023
750	1,3597	-0,10681	-0,82757	0,2273	0,43503	-0,11411
760	1,36413	-0,10886	-0,84101	0,23364	0,44006	-0,11907
770	1,4135	-0,12491	-0,91952	0,26268	0,4804	-0,13497
780	1,47211	-0,14378	-1,00406	0,29132	0,52458	-0,14918
790	1,46014	-0,14248	-0,96339	0,281	0,49994	-0,14149
800	1,39708	-0,12613	-0,83251	0,24255	0,42831	-0,11892
810	1,30322	-0,09812	-0,64065	0,18469	0,32541	-0,08646
820	1,23119	-0,08347	-0,50422	0,14974	0,25354	-0,06661
830	1,27897	-0,09801	-0,59564	0,17914	0,30194	-0,08288
840	1,3946	-0,12999	-0,82226	0,2486	0,42466	-0,12262
850	1,48684	-0,15767	-1,02211	0,30973	0,53383	-0,15811
860	1,53306	-0,17332	-1,12535	0,34335	0,58958	-0,17738
870	1,54842	-0,17691	-1,14042	0,35056	0,59708	-0,18138
880	1,50916	-0,16271	-1,02979	0,31961	0,53667	-0,1636
890	1,39819	-0,1247	-0,77108	0,24298	0,40087	-0,1215

Table B.5: SEDES2 Coefficients by wavelenght (3rd part)

Wavelength (nm)	$A1_\lambda$	$A2_\lambda$	$B1_\lambda$	$B2_\lambda$	$C1_\lambda$	$C2_\lambda$
900	1,17612	-0,06824	-0,34215	0,12086	0,17627	-0,05349
910	0,98685	-0,01315	0,01589	0,01561	-0,00784	0,00326
920	0,83041	0,03159	0,28469	-0,06127	-0,14019	0,04291
930	0,61123	0,09701	0,6077	-0,15086	-0,28158	0,08258
940	0,36913	0,13744	0,9204	-0,22796	-0,42836	0,12211
950	0,30638	0,13226	1,01793	-0,25108	-0,50619	0,14486
960	0,42764	0,0848	0,85788	-0,20327	-0,46987	0,13276
970	0,65012	0,0345	0,60052	-0,12507	-0,37126	0,09766
980	0,84369	-0,01411	0,35246	-0,04375	-0,26576	0,0582
990	1,01871	-0,05584	0,11521	0,03298	-0,16069	0,01951
1000	1,11071	-0,08242	-0,02662	0,08182	-0,09732	-0,00507
1010	1,15831	-0,09845	-0,10842	0,1117	-0,0598	-0,02013
1020	1,18779	-0,10971	-0,17215	0,13436	-0,02617	-0,03236
1030	1,21662	-0,12039	-0,24681	0,15777	0,01821	-0,04635
1040	1,24295	-0,13007	-0,3248	0,17951	0,06846	-0,06071
1050	1,24295	-0,13007	-0,3248	0,17951	0,06846	-0,06071

The SEDES2 model is an simple and effective approximation for an otherwise exceedingly complex phenomenon and researchs show that reasonable spectral accuracy of about 10% is obtainable. However, the approximation is not exception, and the model performs poorly for weather some events such as snow. Differing cloud climatology and variable albedo and aerosol optical depth atmospheric conditions can lead to spectral model differences in the order of 30-40% [78].

C Principal Component Analysis

Principal Components Analysis (PCA) is a linear transformation that can be used to both reduce dimensionality and crosscorrelation between a given candidate set of input variables. PCA is a well-known technique in statistical data analysis, aimed at expressing the data in such a way that highlight their similarities and differences.

The main goal of a PCA analysis is to identify patterns in data, as it detects the correlation between variables and attempt to reduce the dimensionality. PCA can be interpreted as a method to find the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace, while retaining most of the information.

The remainder of this appendix is divided as follows. Section C.1 deals with the required normalization that the must be given to the input data prior to PCA. Section C.2 presents the Singular Value Decomposition technique which is used the obtain the Principal components, while Section C.3 concerns the transformation of the original dataset into the

C.1 Data standardization

PCA is a statistical technique whose purpose is to condense the information of a large set of correlated variables into a few uncorrelated variables called Principal Components. These components are derived as a linear combination of variables of the data set, with weights chosen so that the principal components become mutually uncorrelated.

Defining a $n \times m$ matrix dataset $\mathbf{M}_0$ as the concatenation of m input vectors u_{0i} of size n, PCA will provide a transformation matrix $\mathbf{T}$ that projects the data to a new coordinate system such that the sucessive coordinates reflects the direction in which there is greater variances. In this projection, the first coordinate, also called the first principal component, carries the greatest variance, the second coordinate the second greatest variance, the pattern repeating until the m-th dimension is reached, in which is contained the smallest amount of variation. Since the first few components contain most of the information, they are retained for further use while the last components can be discarded, thus reducing the dimensionality.

However, it is not advisable to apply PCA directly to $\mathbf{M}_0$. Normalization is important, since it is a variance maximizing exercise. The obtained components will be biased if the mean and variance of the input vectors are not normalized to same values. As such, all input vectors are first normalized to unitary variance and zero mean with help of equation (C.1):

$$u_i = \frac{(u_{0i} - \overline{u_{0i}})}{\sigma_{0i}} \tag{C.1}$$

where u_i is the normalized input vector, $\overline{u_{0i}}$ and σ_{0i} are respectively the mean and the standard deviation of the input vector u_{0i}. The normalized matrix dataset $\mathbf{M}$ is then construted by concatenating the m input vectors u_i, which can then decomposed by Singular Value Decomposition in order to obtain the transformation matrix $\mathbf{T}$.

C.2 Singular Value Decomposition

There are two main methods to perform PCA over a given dataset. The dataset's correlation matrix can be calculated, which is then subjected to eigenvalue decomposition to yield the transformation matrix. Principal components can also be obtained directly from the normalized dataset matrix $\mathbf{M}$, by means of the Singular Value Decomposition (SVD).

SVD is a matrix factoration technique in which the normalized dataset matrix $\mathbf{M}$ is decomposed as a product three matrices, denominated $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$:

$$\mathbf{M} = \mathbf{U\Sigma V}^* \tag{C.2}$$

If $\mathbf{U\Sigma V}^*$ is a singular value decomposition of $\mathbf{M}$, then $\mathbf{U}$ is a $n \times m$ matrix with orthonormal columns, $\mathbf{V}$ is a $m \times m$ orthonormal matrix and Σ is a diagonal matrix with real positive or zero elements, which are called singular values. Columns of $\mathbf{U}$ and $\mathbf{V}$ are respectively called left and right singular vectors. Two positive-definite matrices can be constructed from $\mathbf{M}$: $\mathbf{MM}^*$and $\mathbf{M}^*\mathbf{M}$. Substituting (C.2) yields:

$$\mathbf{MM}^* = \mathbf{U\Sigma V}^* \left(\mathbf{U\Sigma V}^*\right)^* = \mathbf{U\Sigma V}^*\mathbf{V\Sigma U}^* \tag{C.3}$$

$$\mathbf{M}^*\mathbf{M} = \left(\mathbf{U\Sigma V}^*\right)^* \mathbf{U\Sigma V}^* = \mathbf{V\Sigma U}^*\mathbf{U\Sigma V}^* \tag{C.4}$$

As $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, their conjugate product is equal to the identity, i.e. has $\mathbf{U}^*\mathbf{U} = \mathbf{I}$. and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. Substituting in (C.3) and (C.4):

$$\mathbf{MM}^* = \mathbf{U\Sigma^2U}^* \tag{C.5}$$

$$\mathbf{M}^*\mathbf{M} = \mathbf{V\Sigma^2V}^* \tag{C.6}$$

Supposing $n \geq m$, it is possible to show that $\mathbf{MM}^*$and $\mathbf{M}^*\mathbf{M}$ share m eigenvalues, and the remaining $n - m$ eigenvalues of $\mathbf{MM}^*$ are zero. Starting from the decomposition shown in (C.6), the columns of $\mathbf{V}$ and squared diagonal elements of $\mathbf{\Sigma^2}$ can be identified as the eigenvectors and eigenvalues of $\mathbf{M}^*\mathbf{M}$, denoted respectively as V and γ^2. Rewriting (C.6) for a single eigenvector and eigenvalue pair yields:

$$\mathbf{M}^*\mathbf{M}V = \gamma^2 V \tag{C.7}$$

multiplying both sides by $\mathbf{M}$ gives:

$$\mathbf{MM}^*\mathbf{M}V = \gamma^2\mathbf{M}V \tag{C.8}$$

By inspection, it is visible in (C.7) that there is an eigenvector $U = \mathbf{M}V$ and an eigenvalue γ^2 for the matrix $\mathbf{MM}^*$, which proves that $\mathbf{MM}^*$and $\mathbf{M}^*\mathbf{M}$ share m eigenvalues. It remains to be demonstrated that the remaining $n-m$ eigenvalues of $\mathbf{MM}^*$ are zero. Considering an eigenvector-eigenvalue pair $U_\perp$ and δ^2 for $\mathbf{MM}^*$, where $U_\perp$ is non zero and orthogonal to the m eigenvectors $U_i = \mathbf{M}V_i$ already determined. As a consequence $\mathbf{U}^*U_\perp = 0$, and equation (C.9) can be written:

$$\mathbf{MM}^*U_\perp = \delta^2 U_\perp \tag{C.9}$$

Using the decomposition $\mathbf{MM}^* = \mathbf{U\Sigma^2U}^*$ yields:

$$\mathbf{U\Sigma^2U}^*U_\perp = \delta^2 U_\perp$$

$$0 = \delta^2 U_\perp \tag{C.10}$$

As $U_\perp$ is non zero, it is demonstrated that all eigenvalues δ^2 must be zero. Thus, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ can be manually evaluated from the eigenvalue decomposition of $\mathbf{MM}^*$and $\mathbf{M}^*\mathbf{M}$. In practice, more computationally efficient algorithms are employed in analysis software, such as QR decomposition, householder reductions, bidiagonal matrix factoring, and others .

C.3 Projection Matrix

This section concerns about the construction of projection matrix $\mathbf{T}$, which is necessary to transform the original dataset $\mathbf{M}$ to obtain a k-dimensional feature subspace $\widehat{\mathbf{M}}$.

The PCA finds the directions in the data with the most variation, i.e. the eigenvectors corresponding to the largest eigenvalues of the covariance matrix, and project the data onto these directions. The motivation for doing this is that the most variance, i.e. second order information, are in these directions. The choice of the number of directions are often guided by trial and error, but principled methods also exist.

Denoting by $\mathbf{T}$ the matrix of left singular vectors sorted according to its respective eigenvalue, it is possible to perform a transformation $\widetilde{\mathbf{M}}$ in the data by means of a simple multiplication:

$$\widetilde{\mathbf{M}} = \mathbf{T}^*\mathbf{M} \tag{C.11}$$

The eigenvectors, in this case, are called Principal Components. Selecting only the first d rows of $\widetilde{\mathbf{M}}$, one obtains the projection $\widehat{\mathbf{M}}$ of $\mathbf{M}$ in the d-dimensional feature subspace, performing the Principal Component Analysis.

D Artificial Neural Networks

A neural network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Artificial Neural Networks (ANN) are composed of multiple nodes, which mimic biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on it, the result being passed ahead to other neurons. The output at each node is called its activation or node value.

One of the key elements of a neural network is its ability to learn. A neural network is not just a complex system, but a complex adaptive system, meaning it can change its internal structure based on the information flowing through it. Typically, neural networks are trained so that a particular input leads to a specific target output, based on a comparison of the output and the target, until the network output matches the target. Generally, a large amount of input and target data is required to train a network. Typically, this is achieved through the adjusting of weights, a number that controls the signal gain between the two neurons. If the network generates a “good” output according to the training cost function, there is no need to adjust the weights. However, if the network generates a “poor” output, then the system alters the weights in order to adapt and improve subsequent results.

In the last years, Neural networks have been used to perform complex functions in various fields, including time series forecasting, pattern recognition, identification, classification, speech, vision, and control systems. In this appendix, an overview of a Multilayer Perceptron trained via supervised learning by the backpropagation algorithm. Section D.1 explains the perceptrons and the multilayer perceptron structure usually employed to forecast time series, Section D.2 concerns the definitions and parameters applicable to supervised training, and Section D.3 presents an basic backpropagation algorithm.

D.1 Multilayer perceptron

The perceptron is the simplest neural network possible: a computational model of a single neuron. A perceptron consists of one or more inputs, a processor, and a single output.

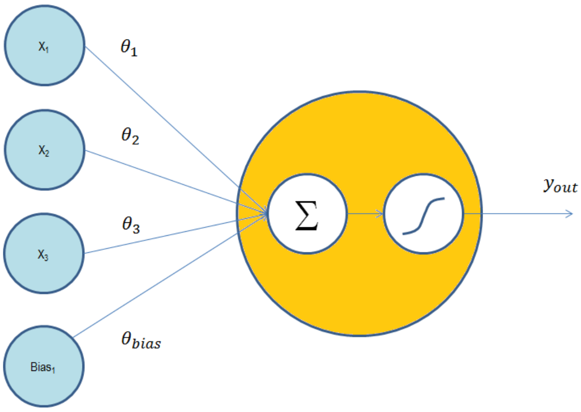


Figure D.1: Perceptron with 3 inputs and bias. From left to right: inputs, weights, summation block, activation function and output.

A perceptron follows the feed-forward model, meaning inputs are sent into the neuron, are processed, and result in an output. In the diagram above, this means the neuron reads from left to right: inputs come in, are weighted and summed, processed by the activation function generating an output. In single perceptrons, the on-off boolean activation function is one of the simplest and most employed. When arranged in networks, the neurons can use other activation functions, usually nonlinear, such as sigmoid function, hyperbolic, radial basis functions, and others.

An array of perceptrons, the Multilayer Perceptron (MLP) can be viewed as a regression classifier where the input is first transformed using a learnt nonlinear transformation, then linearly processed in the output layer. This transformation projects the input data into a space where it becomes linearly separable. This intermediate layer is referred to as a hidden layer. A single hidden layer is sufficient to make MLPs a universal approximator. Figure D.2 illustrates a MLP:

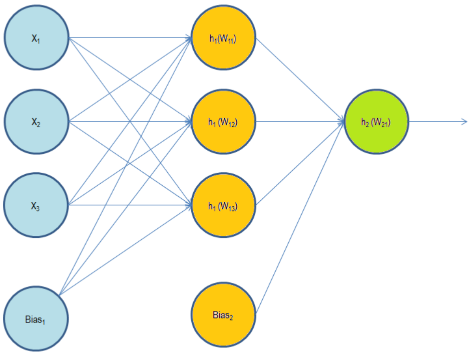


Figure D.2: Schematic of Multilayer Perceptron. From left to right, the input layer (light blue), the hidden layer (yellow) and output layer (green).

D.2 Supervised learning

Learning algorithms can be divided into supervised and unsupervised methods. Supervised learning denotes a method in which some input vectors are collected and presented to the network. The output computed by the network is observed and the deviation from the expected answer is measured. The weights are corrected according to the magnitude of the error in the way defined by the learning algorithm. This kind of learning is also called learning with a teacher, since a control process knows the correct answer for the set of selected input vectors.

When training multilayer networks, the general practice is to first divide the data into three subsets. The first subset is the training set, which is used for computing the gradient and updating the network weights and biases. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error on the validation set typically begins to rise. The network weights and biases are saved at the minimum of the validation set error.

During training, the progress is constantly monitored in order to access the performance, the magnitude of the performance gradient and the number of failures in validation checks. The magnitude of the gradient and the number of validation checks can be used to terminate the training, instead of the raw performance metric. The gradient will become very small as the training reaches a minimum of the performance. A lower threshould can be assigned, and if the magnitude of the gradient decreases below this limit, the training will stop. The number of validation checks represents the number of successive iterations that the validation performance fails to decrease. If this number reaches an also assigned maximum value, the training will stop.

D.3 Backpropagation algorithm

The backward propagation of errors or backpropagation, is a common method of training artificial neural networks and used in conjunction with an optimization method such as gradient descent. The algorithm repeats a two phase cycle, propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network, layer by layer, until it reaches the output layer. The output of the network is then compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output.

Backpropagation requires a known, desired output for each input value in order to calculate the loss function gradient – it is therefore usually considered to be a supervised learning method; nonetheless, it is also used in some unsupervised networks such as autoencoders. It is a generalization of the delta rule to multi-layered feedforward networks, made possible by using the chain rule to iteratively compute gradients for each layer. Backpropagation requires that the activation function used by the artificial neurons be differentiable.

$$\begin{aligned} J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)} log(1 - (h_\Theta(x^{(i)}))_k)\right] + ... \\ +\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2 \end{aligned} \tag{D.1}$$

where K is the number of output elements, i selects ith element, $J(\Theta)$ is the cost function, an inner sum over k output units. Regularization term sums over $\Theta_{ji}^{(l)}$ terms but don't sum over 0^{th}, bias, term.

```
1. Pick a network architecture
    (a) Number of input units: Dimension of features $x^{(i)}$
    (b) Number of output units: Number of outputs
    (c) reasonable default for number of hidden layers: 1, or if ¿1 have same number of hidden units in every layer (usually the more the better but more computationally expensive)
2. Randomly initialize weights
3. Implement forward propagation to get $h_\Theta(x^{(i)})$for any $x^{(i)}$
4. Implement computation of cost function $J(\Theta)$
5. Implement backdrop to compute partial derivatives $\frac{\partial}{\partial\Theta_{jk}^{(l)}}J(\Theta)$
    (a) for i=1:m
        i. Perform forward propagation and backpropagation using example $(x^{(i)}, y^{(i)})$
           (Get activations $a^{(l)}$and delta terms $\delta^{(l)}$for $l = 2, ..., L$)
        ii. compute delta terms
           $\Delta^{(l)} := \Delta^{(l)} + \delta^{(l+1)}(a^{(l)})^T$
    (b) compute derivative terms
        $\frac{\partial}{\partial\Theta_{jk}^{(l)}}J(\Theta)$
6. Use gradient checking to compare $\frac{\partial}{\partial\Theta_{jk}^{(l)}}J(\Theta)$ computed using backpropagation vs. using numerical estimate of gradient of $J(\Theta)$
   Then disable gradient checking code.
7. Use gradient descent or advanced optimization method with backpropagation to try to minimize $J(\Theta)$ as a function of parameters $\Theta$.
   If $J(\Theta)$- is non-convex, it can get stuck in a local minimum.
```

Algorithm 1: Backpropagation - preparation

Minimizing the cost function $J(\Theta)$ evaluating its gradients $\frac{\partial}{\partial\Theta_{ij}^{(l)}}J(\Theta)$, weight matrix element $\Theta_{ij}^{(l)} \in \mathbb{R}$.

1. Apply forward propagation.
   - $a^{(1)} = x$
   - $z^{(2)} = \Theta^{(1)}a^{(1)}$
   - $a^{(2)} = g(z^{(2)})\ \ (add\, a_0^{(2)})$
   - $z^{(3)} = \Theta^{(2)}a^{(2)}$
   - $a^{(3)} = g(z^{(3)})\ \ (add\, a_0^{(3)})$
   - $z^{(4)} = \Theta^{(3)}a^{(3)}$
   - $a^{(4)} = h_\Theta(x) = g(z^{(4)})$
2. Compute gradient by using backpropagation. Then compute the error in the activation of node j in layer l: $\delta_j^{(l)}$.
3. Compute error in last layer: $\delta_j^{(4)} = a_j^{(4)} - y_j$.

   (a) Each of δ, a, y's dimension is equal to the number of output units in the network.
4. Compute δ terms for the earlier terms in the network.

   $\delta^{(3)} = (\Theta^{(3)})^T\delta^{(4)}.*g'(z^{(3)})$

   $\delta^{(2)} = (\Theta^{(2)})^T\delta^{(3)}.*g'(z^{(2)})$
5. Evaluate $g'(z^{(3)}) = a^{(3)}.*(1 - a^{(3)})$

   There is **no** $\delta^{(1)}$term.
6. To calculate backpropagation given a training set $\left\{(x^{(1)}, y^{(1)}, ..., (x^{(m)}, y^{(m)})\right\}$.

   Set $\Delta_{ij}^{(l)} = 0\,(\forall_{l,\,i,\,j})$eventually this will be used to compute the derivative term $\frac{\partial}{\partial\Theta_{ij}^{(l)}}J(\Theta)$.

   Then loop through the training set:

   For $i = 1$to m

   set $a^{(1)} = x^{(i)}$
7. Perform forward propagation to compute $a^{(l)}$for $l = 2,\,3,\,...,\,L$

   Using output label $y^{(i)}$from a specific example, compute the error term $\delta^{(L)} = a^{(L)} - y^{(i)}$ for the output layer L. $a^{(L)}$is what the hypothesis outputs, minus what the target label was, $y^{(i)}$

   Use backprop algo to compute $\delta^{(L-1)}$, $\delta^{(L-2)}$, ..., $\delta^{(2)}$