LOAD HINDCASTING: A RETROSPECTIVE REGIONAL LOAD PREDICTION METHOD USING REANALYSIS WEATHER DATA

A Thesis Presented

by

JONATHAN D. BLACK

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN MECHANICAL ENGINEERING

September 2011

Mechanical and Industrial Engineering

LOAD HINDCASTING: A RETROSPECTIVE REGIONAL LOAD PREDICTION METHOD USING REANALYSIS WEATHER DATA

A Thesis Presented

by

JONATHAN D. BLACK

Approved as to style and content by:

Jon G. McGowan, Chair

James F. Manwell, Member

Christopher V. Hollot, Member

Donald Fisher, Department Chair
Mechanical and Industrial Engineering

ACKNOWLEDGMENTS

Foremost, warm thanks to my committee for suspending their disbelief and asking the right questions as I stumbled through the semidark. Sincerest appreciation goes to all faculty, staff, and students who over the years have carried the torch of the Wind Energy Center (WEC), formerly known as the Renewable Energy Research Laboratory (RERL) – without this tower of shoulders to stand on, this research would not have been possible. Additionally, many thanks to the welcoming camaraderie of the numerous colleagues who became friends over these past years. Special thanks to William Henson, whose vision and mentorship helped pave the way for my timely internship opportunity at ISO New England and also for this research. Loving thanks to Emily Rae, who without complaint weathered all of my incessant rambling, kept the domestic sphere in working order, and single-handedly contributed the majority of line edits to all of this fiction.

ABSTRACT

LOAD HINDCASTING: A RETROSPECTIVE REGIONAL LOAD PREDICTION METHOD USING REANALYSIS WEATHER DATA

SEPTEMBER 2011

JONATHAN D. BLACK

M.S.M.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jon G. McGowan

The capacity value (CV) of a power generation unit indicates the extent to which it contributes to the generation system adequacy of a region's bulk power system. Given the capricious nature of the wind resource, determining wind generation's CV is nontrivial, but can be understood simply as how well its power output temporally correlates with a region's electricity load during times of system need. Both wind generation and load are governed by weather phenomena that exhibit variability across all timescales, including low frequency weather cycles that span decades. Thus, a data-driven determination of wind's CV should involve the use of long-term (i.e., multiple decades) coincident load and wind data. In addition to the challenge of finding high-quality, long-term wind data, existing load data more than several years old is of limited utility due to shifting end usage patterns that alter a region's electricity load profile. Due to a lack of long-term data, current industry practice does not adequately account for the effects of weather variability in CV calculations. To that

end, the objective of this thesis is to develop a model to "hindcast" what the historic regional load in New England would have been if governed by the conjoined influence of historic weather and a more current load profile. Modeling focuses exclusively on summer weekdays since this period is typically the most influential on CV.

The summer weekday model is developed using multiple linear regression (MLR), and features a separate hour-based model for eight sub-regions within New England. A total of eighty-four candidate weather predictors are made available to the model, including lagged temperature, humidity, and solar insolation variables. A reanalysis weather dataset produced by the National Aeronautics and Space Administration (NASA) – the Modern Era Retrospective-Analysis for Research and Applications (MERRA) dataset – is used since it offers data homogeneity throughout New England over multiple decades, and includes atmospheric fields that may be used for long-term wind resource characterization. Weather regressors are selected using both stepwise regression and a genetic algorithm (GA) based method, and the resulting models and their performance are compared. To avoid a tendency for overfitting, the GA-based method employs triple cross-validation as a fitness function. Results indicate a regional mean absolute percent error (MAPE) of less than 3% over all hours of the summer weekday period, suggesting that the modeling approach developed as part of this research has merit and that further development of the hindcasting model is warranted.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

INTRODUCTION

Load hindcasting is a method of extending the record length of electric load data needed to support a *capacity value* (CV) determination for wind power generation (or any other power resource that is weather-driven) that accounts for weather phenomena occurring on interannual to interdecadal timescales. In short, a CV is a measure of how much a generator contributes to the needs of a power system, especially during times of greatest system need. The existing state-of-the-art methods employed in regional wind integration studies typically arrive at a CV using no more than three years of data, which provides only a glimpse of the interannual variability occurring over that time frame (or less). In this sense, load hindcasting will offer an improvement to methods used by the power industry for wind's CV calculation.

The pace of wind power development throughout the world has created a need for understanding the operational effects of increased wind penetrations within regional power systems. Regional balancing area authorities and transmission system operators like ISO New England Inc. (ISO-NE) in Holyoke, Massachusetts are tasked with ensuring the reliable generation and delivery of power to a broad base of customers — or "keeping the lights on." Meeting this responsibility requires an ongoing assessment of whether a region's installed generation capacity is sufficient to meet the forecast load. This presumed sufficiency is put to the test during peak load events, when peak load must be satisfied while also maintaining a fleet of marginal capacity for potential contingency events (e.g., when a transmission line unexpectedly faults and temporarily goes out of service). An integral component of this capacity assessment involves a determination of the capacity value of regional generators, classes of generators (i.e.,

natural gas combustion generators, nuclear steam generators, wind plants), and/or individual generation facilities.

Weather-driven power generation sources – such as wind – introduce a need for a new approach to power system analysis. Save for what can be stored, the bulk of electricity must be generated in relative balance with real-time electricity demand. While conventional power generation is dispatchable in the sense that it can be operated according to the needs of a power system, weather-driven generation is subject to a capricious fuel supply and is therefore non-dispatchable (although one could argue that wind's capability for down-regulation more accurately characterizes it as semi-dispatchable). The two characteristics of wind power that must be accounted for are its variability and its relative unpredictability.

Chapter 1 of this report provides some background information that lends merit to this research, beginning with the concept of capacity value and how it can be calculated. Next, there is a brief discussion of the 2010 New England Wind Integration Study (NEWIS) and the method that ISO-NE currently uses to calculate wind's capacity value. A discussion of weather variability that is relevant to the long-term characterization of wind resources follows, which was reconnoitered mostly from the fields of climatology and atmospheric science. The Chapter finishes with a summary of existing load forecasting techniques, including those currently used by ISO-NE.

Chapter 2 introduces the reader to New England's load profile and to the Modern Era Retrospective-Analysis for Research and Applications (MERRA) reanalysis dataset, which was used to provide weather data. Some analyses of load (regional and subregional) and weather-load relationships in New England follow.

Chapter 3 describes the load response system that will be modeled, as well as how some of the lessons learned from the analysis of the relationships between weather and load will be integrated into the overall modeling approach. Some features of load

hindcasting that distinguish it from load forecasting are presented, as are some other potential applications of the load hindcasting methodology that is under development.

Chapter 4 delves into the Multiple Linear Regression (MLR) method of model building, and includes a discussion of the ordinary least-squares parameter estimation method. A description of the candidate regressors, the stepwise regression and genetic algorithm (GA) based variable selection methods, and the model training methods follows.

Chapter 5 explains the residual analysis and performance metrics that will be used to test the adequacy of the models.

Chapter 6 describes the results of this research, including a comparison of the regression models built by the stepwise and GA-based variable selection methods and their overall performance.

Chapter 7 formulates conclusions about the research conducted and offers suggestions for future research.

CHAPTER 1

BACKGROUND

1.1 The Capacity Value (CV) of Wind Power

The existence of sufficient power generation facilities to serve customer load within a power system is referred to as system adequacy [23]. In general, the capacity value (CV) of an individual generator is a measure of its contribution to system adequacy (i.e., to satisfying the load-generation balance over time), and is expressed as a percentage of the generator's nameplate capacity. As a weather-driven resource, arriving at a CV for wind power is more of a challenge than it is for conventional generation. Due to varying weather, significant fluctuations in a wind generator's CV are common from one year to the next. While there are many ways of calculating CV, there are two general methods [28]:

1. Methods that estimate values based on average power production during times of peak system demand

2. Data-intensive, probabilistically-derived methods that consider the time series system load and assumed availabilities of the conventional generators on the system. These methods treat wind power as a *load modifier* and compare the assumed availabilities of the conventional generation fleet to the time series of the resultant load minus the wind power, referred to as the *net load*. (This research consists of the development of a method of generating a long-term load dataset that supports a robust, data-driven CV determination.)

In December 2007, the North American Electric Reliability Corporation (NERC) created the Integration of Variable Generation Task Force (IVGTF) to identify technical considerations associated with integrating large amounts of variable generation into regional power systems. IVGTF has since been working to develop specific policies, practices, mitigation measures, and requirements needed to ensure bulk power system reliability in the presence of large amounts of variable generation. One of the tasks of the IVGTF is to review the variety of methods currently used to determine the CV of wind, and develop a consistent and accurate method associated with variable generation. Although this work is still in progress, IVGTF has identified a metric called the effective load carrying capacity (ELCC) as the most promising approach [35]. As it applies to a generator (or class of generators), ELCC is the amount of load that can be served by a power system while maintaining the same level of reliability due to the addition of a generator (or class of generators). The reliability level is expressed probabilistically, and is referred to as loss of load expectation (LOLE), which is the number of days per year that load will not be met by available generation. Figure 1.1 illustrates the ELCC concept as it applies to wind power. The ELCC of wind power is the difference in the load carrying capacity of the LOLE curve for the system without wind (in blue) and with wind (in green) at the target reliability level (LOLE of 0.1 days/year), or 400 MW [14]. In general, wind power's CV is a strong function of the temporal correlation in its power output and the system load, especially at times of peak system load.

In an attempt to account for wind's fluctuating CV, common industry practice is to develop an average over multiple years, but rarely more than three. However, the existence of low frequency variations in the wind resource on timescales longer than three years suggests that this practice affords only a glimpse into the inter-annual characteristics of the wind resource [17]. Thus, finding a method of extending the record length of the coincident load-wind dataset could enable a more robust

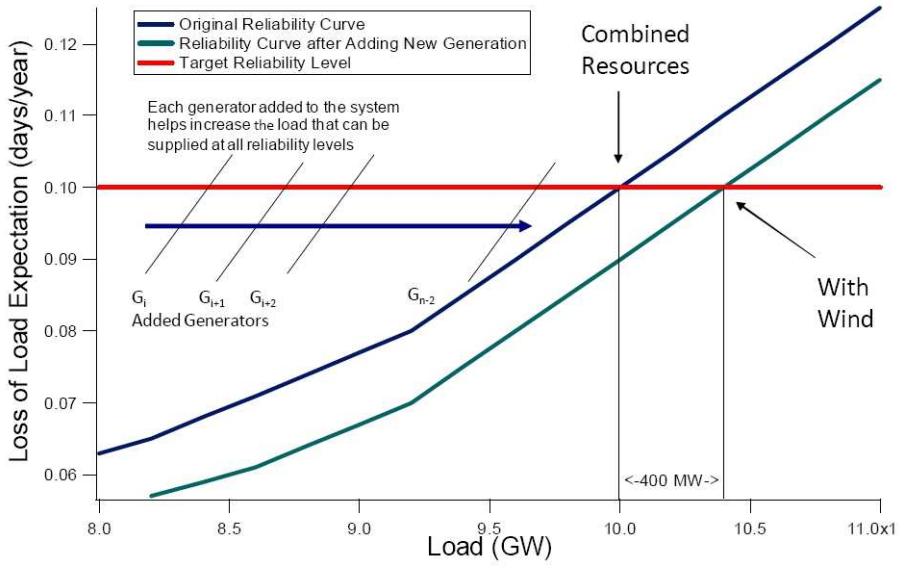

Figure 1.1. A power system's loss of load expectation (LOLE) without wind (blue curve) and with wind (green curve)[14]

CV determination that accounts for lower frequency resource variability, while also making it commensurate with the length of time a wind farm is designed to be in service (typically 20-25 years).

In general, the wind power available in the mass flow of air is proportional to both the cube of the wind speed and the atmospheric air density (itself a function of air temperature and pressure), making it extremely sensitive to changes in the weather. And as we shall see in Chapter 2, the regional load is also heavily influenced by the weather. Since both wind power generation and consumer electricity demand are weather-driven stochastic processes, the effects of the weather simultaneously governing them both must be well-represented in the coincident load-wind dataset that is used in any data-driven CV determination. (This also applies to any other weather-driven generation technology, e.g., photovoltaics).

1.2 The New England Wind Integration Study (NEWIS)

ISO New England (ISO-NE) currently estimates the CV of wind generation using a plant's average capacity factor between 1400 hours and 1800 hours for the months of June through September [13]. The rationale is that since the New England control area is a summer peaking system, generators that contribute firm capacity in typical summer peak load hours are most critical to overall system adequacy.

In 2009, ISO-NE commissioned General Electric International, Inc. (GE) to conduct the New England Wind Integration Study (NEWIS) to assess the operational effects of integrating large-scale wind power development in the region. As part of reliability analyses conducted for the NEWIS, realistic hourly power production simulations were made for a series of hypothetical wind fleets over three calendar years, and data-driven calculations of the capacity value for the regional wind profile were made and then compared to the ISO's approximate method. Based on their evaluation, GE concluded that ISO's approximate method yielded reasonable values, but that because only three years of data were used in the study a comparison between ISO's approximate method and the data-driven method employed for NEWIS should be monitored as wind penetrations increase in the region. A final recommendation emerging from the NEWIS was that ISO-NE should evaluate potential improvements to their method of calculating CV for wind [13].

1.3 Weather Variability

Much like the weather driving it, wind power is variable across all timescales; high frequency variability spanning seconds to minutes is produced by wind gusts and turbulence, whereas low frequency fluctuations on an inter-annual to inter-decadal timescale are produced by low frequency weather cycles. As such, while not its focus, some of the merit of this project rests on a discussion of New England's climatology.

Weather characteristics in the lowest part of the atmosphere (a.k.a. the atmospheric boundary layer) that are most relevant to wind power include atmospheric temperature, wind speed and direction, and atmospheric pressure and humidity (both of which influence air density). Weather variability affects wind power's CV by influencing a variety of wind resource characteristics such as wind speed distribution, turbulence intensity, diurnal profile, and prevailing wind direction. The frequency and duration of extreme temperatures or wind speeds beyond the operating limits specified by a turbine's manufacturer may also reduce the availability of a region's wind fleet.

The field of atmospheric science suggests that 30 years of data is needed to develop long-term estimates of climate, and at least five years of data should give a useful estimate of a location's average annual wind speed [28]. Although shorter wind data sets can be used to extrapolate long-term mean wind speeds, which may be useful for resource assessment or siting purposes, these values are not useful for calculating wind's CV since they ignore the wind's relationship with system load.

The climatology of a particular region gives an indication of how wind resource characteristics may typically vary over the span of years or longer. This climatology is composed of a cascade of inter-related atmospheric and oceanic cycles of varied and varying periodicity, whose sum effect governs climate variability. There is also evidence linking the cycles of the sun to this variability [37].

Weather fluctuations on these timescales are characterized by *modes* of atmospheric variability. A distinct trait of modes is that they manifest as weather anomalies over large geographic areas, with simultaneous variations experienced by two regions separated by thousands of miles often featuring opposing effects. For example, while one region is atypically cool and dry, the other is atypically warm and wet. This phenomenon is referred to as *teleconnection*.

Two of the principal modes of weather variability exhibiting influence across North America are the El Niño-Southern Oscillation (ENSO) in the Pacific Ocean and the North Atlantic Oscillation (NAO) in the Atlantic Ocean. The NAO dictates weather variability over the eastern seaboard of the United States, but its effects are also influenced by the remote forcing of the ENSO [18]. Of particular relevance to wind power is the NAO's strong influence on storm tracks, and its heaviest influence during winter, when the wind resource in New England is typically the strongest [29]. Figure 1.2 shows the differing effects of the NAO between its positive and negative phases. For instance, the meridional path of the jet stream typified by NAO's negative phase often causes a greater frequency of winter storms and snowfall [53].

In spite of a wide range of active climatology research, the fundamental mechanisms behind modes of climate variability like the NAO and ENSO remain unclear, especially on decadal timescales or longer [26]. And while the influence of modes and teleconnection on the variability of weather parameters such as precipitation has been studied in depth [24], their specific effects on near surface wind speeds has scarcely been considered until recently. However, growth in the wind industry over recent years has sparked interest in understanding how shifts in the global climate system translate into variations in near-surface wind regimes [42]. A recent study analyzing near surface wind speed measurements collected between 1973 and 2005 from 157 stations across the continental United States and archived by the National Climate Data Center (NCDC) has indicated a general decline in wind speeds, especially in eastern states [41]. Studies such as these are not yet well understood and/or integrated with respect to climate variability.

Although questions remain about how low frequency weather cycles mesh and interact, and their corresponding relevance to the wind industry, the approximate timescales corresponding to each mode of variability have been established. The period of ENSO-related variability is between two and seven years, whereas the NAO

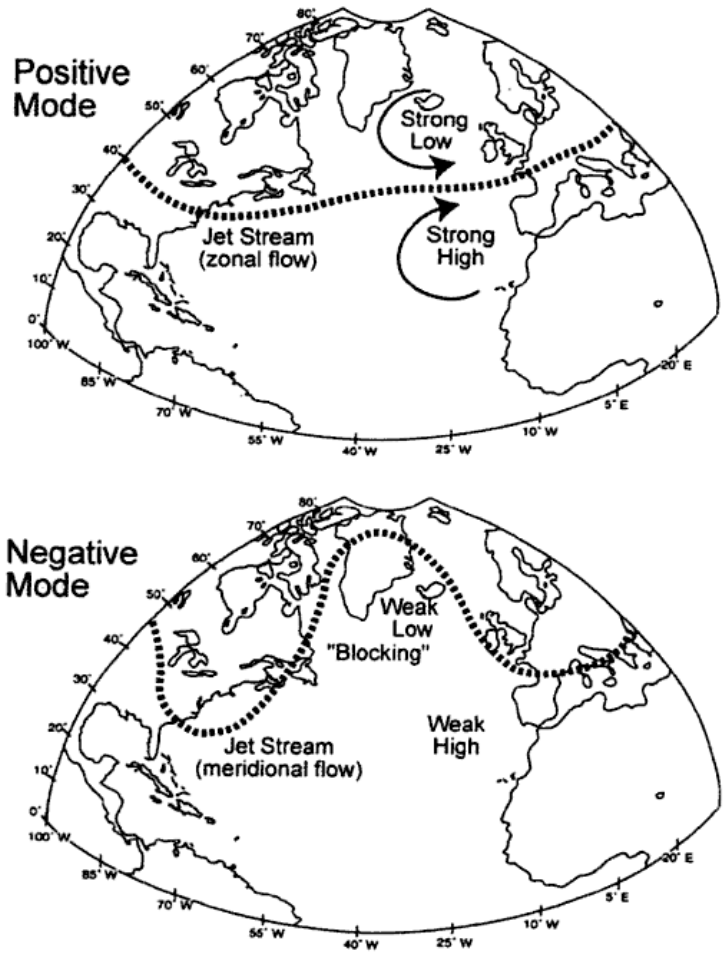

Figure 1.2. The two modes of the NAO and their effects on the trajectory of the jet stream [53]

has a periodicity of six to ten years [52]. If the sun is also a significant forcing mechanism, then the 11-year solar cycle is also relevant. The nature of these cycles and their respective periods suggests a strong relevance to the wind industry, including on the decadal timescale.

While distilling the results of this active research field into quantifiable terms most applicable to wind power in New England may be infeasible at this time, it is clear that low frequency variability exists and that it exerts a strong influence on wind power production. Therefore, methods to capture weather variability on these timescales by realistically extending the data period used in wind's CV calculation are of value since they offer the potential of increased robustness.

1.4 Obstacles to a Robust Capacity Value (CV) Determination

With respect to the data-driven methods (rather than estimation methods such as the one currently used by ISO-NE described above), obstacles to extending the coincident record length required for a robust CV determination for wind power stem from limitations in developing both realistic long-term wind and load data. This research will focus on developing one technique of extending the record length of the load dataset. There may be other methods of making a more robust CV calculation that are not data driven that are beyond the scope of this research.

Although long-term load data already exist [21], the load profile is often subject to fairly rapid change (i.e., over the span of less than a decade) due to changes in population, shifting patterns of consumer activity (e.g., reliance on air-conditioning), and economic factors. Since a CV calculation is inextricably linked to load, these shifts in load limit the utility of the historic load data, since it is no longer representative of more contemporary electricity demand patterns. Some of these changes will be explored in detail in Section 2.2.

It should be noted that methodologies are currently being developed to extend the length of shorter-term wind datasets to more fully characterize the inter-annual variability of the wind resource. One such method is suggested by Henson, McGowan and Manwell [17], and involves using reanalysis datasets (including the Modern Era Retrospective-Analysis for Research and Applications dataset, which will be used for this research) to extend the period of simulated wind data. As methodologies such as these continue developing, they will be increasingly complementary to load hindcasting.

1.5 A Summary of Load Forecasting Models

As previously stated, load hindcasting is a method of extending the record length of electric load data needed to support a CV determination for wind power generation that accounts for weather phenomena occurring on interannual to interdecadal timescales. This retrospective load prediction method is potentially a new application of some of the techniques used for load forecasting. In general, load forecasting is grouped by forecast horizon into short-term load forecasting (STLF), intermediate-term load forecasting (ITLF), and long-term load forecasting (LTLF). STLF predicts the load up to a week in advance to ensure that day-to-day operation of the power system is planned efficiently and cost-effectively [47]. Intermediate and long-term forecasting are concerned with forecast horizons on the order of months and years, respectively. Of the three, STLF typically incorporates weather variables directly due to its relatively short forecast horizon. Longer horizons preclude the incorporation of accurate weather forecasts, but typically involve some kind of statistically-based weather patterns derived from historic data. Given that STLF is more weather-based, some of the methodologies used for STLF may be applicable to this research.

Despite the numerous STLF methods that have been developed over the last few decades, continued research in this field remains active [12]. In general, STLF tech-

niques can be divided into two broad categories: conventional or classical approaches and artificial intelligence-based techniques. Conventional approaches include time series models, Kalman filtering techniques, and regression models. (A form of multiple linear regression (MLR) will be used for this research and will be described in Section 4.) Artificial intelligence-based techniques include artificial neural networks (ANN), fuzzy logic, neural models, expert systems and support vector machines [27]. Although artificial intelligence-based techniques are heavily relied upon in the field of load forecasting (especially ANN) these techniques were not considered for this research due to their 'black box' nature.

Given the dynamic nature of load, the model for the load response must include some form of lagged weather (i.e., weather from previous time steps). An example of a method of using lagged weather is represented by the weather window shown in Figure 1.3, which was suggested by Soliman et al [2]. The weather window illustrates that the current load may be a function of lagged weather from the following time steps:

1. The previous hour (shown as 'Day i+2, hour 1')

2. The previous day (shown as 'Day i+1, hours 2, 3, and 4', which represent time steps 23, 24, and 25 hours prior)

3. Two days prior (shown as 'Day i, hours 5, 6, and 7', which represent time steps 47, 48, and 49 hours prior)

This weather window approach is intended to capture the load's dependence on certain weather variables in previous hours or days.

1.5.1 ISO New England Load Forecasting

The load forecasting methods used by ISO-NE were reviewed to develop a sense of how weather is currently used to predict the load response in New England. ISO-

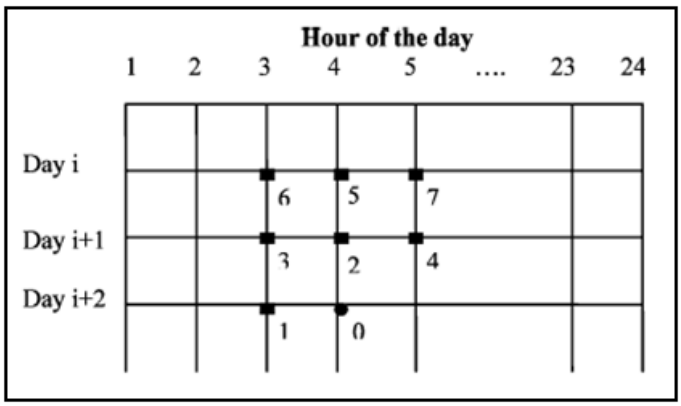

Figure 1.3. A weather window [2].

NE uses an ensemble approach consisting of a combination of three types of demand forecasts to predict hourly loads for the current and next two operating days [20], [19]:

1. A Similar Day ("Simday") approach develops an hourly load forecast for the next operating day by selecting up to five "similar days" from a database. Similar days are found based on the correlation of their weather with an average of three weather forecasts provided for the eight locations listed below, which are weighted according to the values indicated in parenthesis:

   (a) Logan Airport - Boston, MA (0.208)

   (b) Bradley Field in Windsor Locks, CT (0.277)

   (c) Bridgeport, CT (0.073)

   (d) Worcester, MA (0.212)

   (e) Providence, RI (0.049)

   (f) Concord, NH (0.057)

   (g) Burlington, VT (0.043)

   (h) Portland, ME (0.084)

Similar Days are screened and adjusted based on a comparison of the hourly loads in the hour preceding the forecast period, with that of the last hourly value of the current day's load forecast. Weather variables used in the summer Simday application include temperature-humidity index (THI), dew point, cloud cover, and precipitation. The user also has to specify a day of week (Sun, Mon, Tue-Wed-Thurs, Fri, Sat, and holiday).

2. An ANN model developed by the Electric Power Research Institute (EPRI) that generates hourly forecasts for the next seven operating days. The only weather input to the ANN model in summer is a temperature-humidity index (THI), and the model is re-trained annually.

3. A Next Day regression model developed by Metrix that uses a weighted THI to forecast load in summer. This model also utilizes THI as its only weather input, and is re-trained twice per month.

The aggregate model is updated in real-time if a prediction error of 400 MW or greater is experienced and expected to continue.

CHAPTER 2

REGIONAL LOAD AND WEATHER

2.1 New England Weather Data

A retrospective analysis (or 'reanalysis') data set provided by the National Aeronautics and Space Administration (NASA) called the Modern Era Retrospective-Analysis for Research and Applications (MERRA) reanalysis dataset will be used to provide New England weather variables. MERRA uses a new version of the Goddard Earth Observing System Data Assimilation System (GEOS-5) and incorporates the current suite of research satellite observations in the context of climate. Most variables of interest are available within MERRA's 2-D diagnostic fields, which offer native values time-averaged on an hourly basis [34].

A multitude of meteorological inputs are provided as initial conditions to a numerical weather prediction (NWP) model that is used to provide weather forecasts. However, as improvements are made to the NWP model, spurious shifts in the climate are introduced that obfuscate the interannual trends that are needed and sought by climatologists. Reanalysis was first proposed in 1988 as a solution to this problem by providing a consistent set of data to reveal these interannual trends [5]. Since climate reanalysis is performed with a fixed NWP model and data assimilation method, it produces a comprehensive, consistent, uniformly gridded, long-term dataset for the global climate system, including the atmosphere, oceans, and land surface. The atmospheric component often includes an extensive array of weather variables.

Although reanalysis data has myriad applications, it is most commonly used to track climatic trends (i.e., climate variability) [51]. In other applications, reanalysis

data are most useful when available data is scarce or derived from a variety of sources. Other applications have included the estimation of renewable energy resources, including the development of a wind resource map of Newfoundland [25] and a study of Danish wave energy [16].

The MERRA dataset is one of a few recent releases of reanalysis data. Another reanalysis dataset, National Centers for Environmental Prediction's (NCEP) North American Regional Reanalysis (NARR) project, was released in 2004, and has a 32 kilometer resolution and 3-hourly output [30]. The strength of these data is their representation of interannual variability on monthly to seasonal timescales [43]. The use of reanalysis data for wind resource assessment has been proposed [45], and, as already mentioned in Chapter 1.4, the use of MERRA data for long-term wind resource characterization has also been suggested.

Limitations of reanalysis data are often a result of imperfect global models. Given that the release of the MERRA dataset was relatively recent, very little has been published with respect to its specific strengths and deficiencies; however, a special collection of the first publications concerning MERRA have recently become available for early release (see [43], [44], [8], and [7]). As the MERRA dataset becomes more established in the climatological community, further validation of the dataset's quality will be conducted. And as improvements to the global models are made and computational power increases, the quality of data provided by reanalyses will also continue to improve commensurately [6].

A weather dataset produced by a reanalysis project was desired due to the homogeneity of data it offers over multiple decades [38] and region-wide. The MERRA dataset was selected due to the following features that are sought for this project:

1. Temporal granularity - MERRA data is available on an hourly basis for variables of interest, which matches the time resolution of load data provided by ISO New England.

2. Spatial granularity - the MERRA data of interest is available for each 1/2 degree latitude and 2/3 degree longitude. See Figure 2.1 to note the locations of MERRA data output for the New England region.

3. Record length - The MERRA dataset covers the time period from 1979 to 2010.

4. Performance drivers - The primary performance drivers for the global atmospheric model used for MERRA are desirable in the context of this project, and include temperature and moisture fields, wind fields for transport studies of the tropospheric chemistry communities, and climate-quality analyses to support studies of the hydrologic cycle [49].

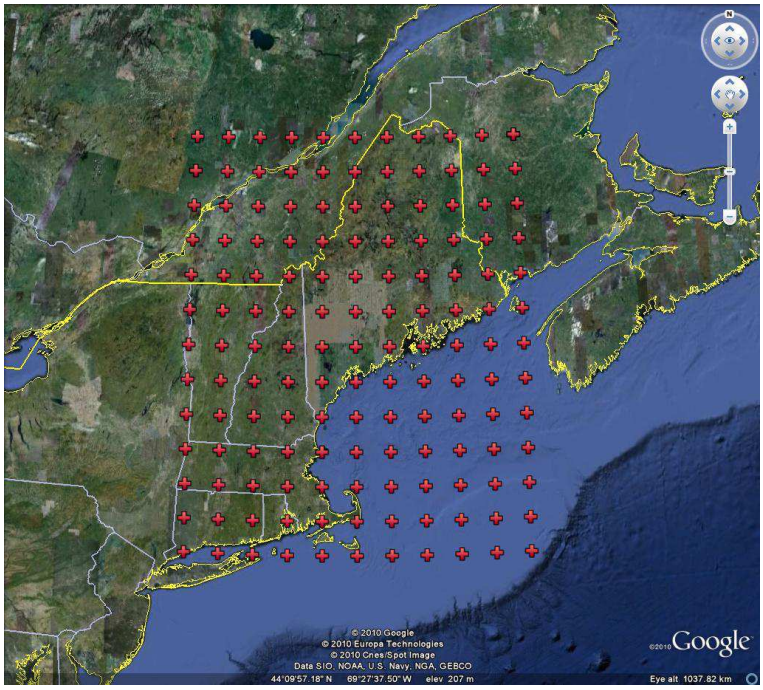

Figure 2.1. The 11x13 grid of MERRA data locations available for New England.

Figure 2.1 illustrates the grid of MERRA data that is available for the New England region. Selection of the final MERRA grid locations that are used for weather inputs to the hindcast model will be discussed in Section 3.3.

Three MERRA atmospheric fields are utilized to provide weather inputs to the hindcast model. Two of the fields are referenced with respect to a *displacement height*, which represents the height at which 'zero wind' occurs. Since the elevation of the displacement height is a function of the surface roughness, which is strongly influenced by vegetation and the built environment, it does not correspond exactly with the ground surface; however, atmospheric variables corresponding to two meters above displacement height are assumed to adequately represent weather conditions at the ground surface. The following is a description of the atmospheric fields used [34]:

1. T2M – The temperature (degrees Kelvin) two meters above the displacement height, available in the 2-D surface turbulent flux diagnostics data collection. This field provides temperature data used by the hindcast model.

2. Q2M – The specific humidity (kg-kg^{-1}) two meters above the displacement height, also available in the 2-D surface turbulent flux diagnostics data collection. This field provides humidity data used by the hindcast model.

3. SWGDN – The surface incident solar radiation (Watts-meter^{-2}) with wavelengths ranging from 0.175 to 3.85 microns, available in the 2-D surface and top of atmosphere (TOA) radiation fluxes data collection. This field provides solar insolation data used by the hindcast model.

Each selected atmospheric field is provided on an hourly, time-averaged basis; however, the timestamps of the averaged data outputs correspond to the middle of each hour (e.g., 4:30, 5:30, etc.). Given that they are time-averaged values, for the purposes of this research it is assumed that they will adequately represent weather conditions for the hour of the output, e.g., the MERRA temperature data output for 4:30 will sufficiently represent the temperature between 4:00 and 5:00.

2.2 New England Load

Hourly time series data of aggregate electricity load for all of New England dating back to 1980 are available on ISO New England's website [21]. These data exclude loads consumed by pumped-hydro facilities in New England, and therefore represent only the electricity consumed by end users. Each hourly value represents the total electrical energy in megawatt-hours (MWh) used over the hour.

Load for the region exhibits daily, weekly, and yearly cycles that are driven by human activity, weather conditions, calendar effects, and economic factors. At the regional level, demand for load is comprised of a variety of end users including residential, commercial, and industrial customers. Figure 2.2 depicts the regional hourly time series load data for the year 2006, which is a typical annual load shape for recent years. The maximum and minimum hourly loads, 28,130 MW and 9,171 MW, respectively, are labeled. The 2006 summer peak is the greatest on record for the region. As can be seen, New England's power system is a summer peaking system, with winter as a secondary peaking season that is much lower in magnitude. Also note the seasonal shift in the load shape, with the summer exhibiting the highest seasonal mean, winter the next highest, and the spring and fall (or 'shoulder' seasons) exhibiting the lowest. Figure 2.3 is a Fourier transform of the 2006 hourly load data that displays prominent semi-diurnal (12-hour cycle, or 0.083 cycles/hr), diurnal (0.0417 cycles/hr), and weekly (approx. 0.006 cycles/hr) frequencies. Magnification of the frequency axis close to the origin (not shown) uncovers a semi-annual cycle corresponding to the seasonal shift in load. This cycle would likely become more apparent if multiple years of data were used. Ranking these various seasonalities according to their respective spectral magnitude (greatest to smallest) yields: 1) diurnal; 2) seasonal; 3) semi-diurnal; and 4) weekly. The seasonalities exhibited by load likely classify it as a non-stationary system, which means that its first and second order statistical moments, mean and autocorrelation function, vary with time [4].

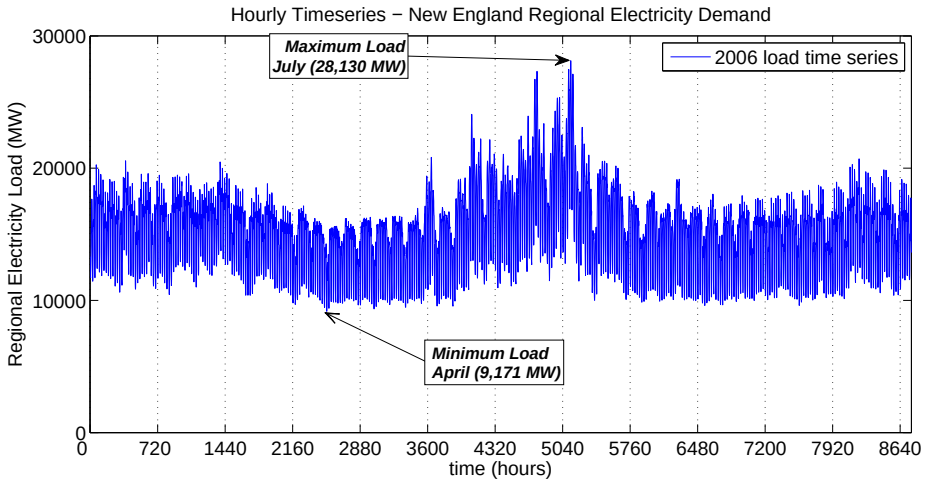

Figure 2.2. Hourly load time series for New England, 2006.

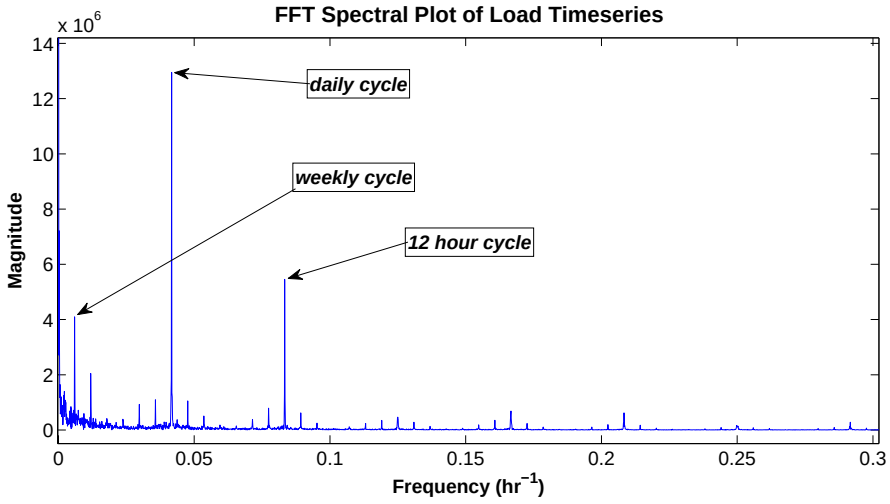

Figure 2.3. Spectral features of the 2006 New England load time series.

Figure 2.4 depicts the spatial distribution of the regional load intensity, with dark red indicating the highest intensity and dark green indicating the lowest [20]. Most of the load is concentrated in southern New England, in the states of Massachusetts and Connecticut. As would be expected, the load intensity is highly correlated with population density.

Subregional load data is also available for the years 2003 to present for each of the eight ISO-NE *load zones*, which are depicted in Figure 2.5. The eight ISO-NE load zones are as follows:

1. Connecticut (CT)
2. Rhode Island (RI)
3. West/Central Massachusetts (WCMA)
4. Southeast Massachusetts (SEMA)
5. Northeast Massachusetts, including Boston (NEMA)
6. Vermont (VT)
7. New Hampshire (NH)
8. Maine (ME)

Hourly subregional load data for 2006 is plotted in Figure 2.6, which illustrates variation in the spatial distribution of the load among the load zones that approaches five percent (see CT curve). The distribution of load intensity shown in Figure 2.4 will be used later to select MERRA gridpoints from those shown in Figure 2.1 that will represent the weather for each of the eight load zones.

Due to its weather-dependence, load is subject to similar weather-driven variability as wind, but in a less variable and more predictable manner. In addition,

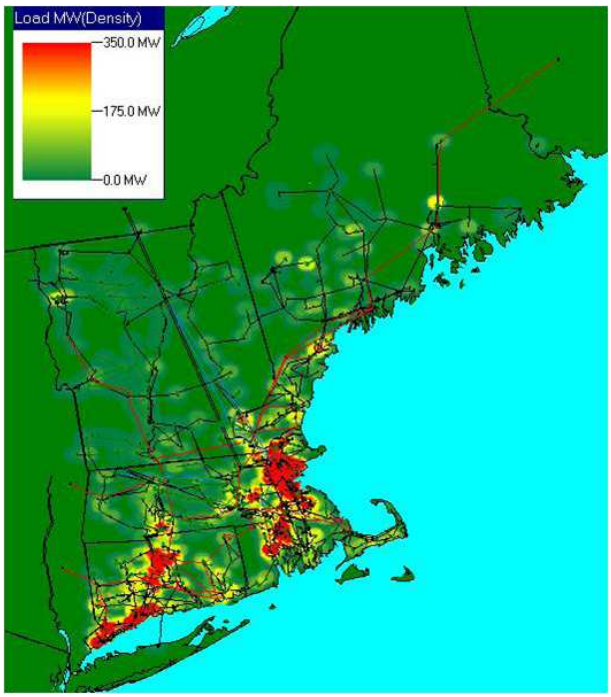

Figure 2.4. Spatial distribution of load intensity for New England.

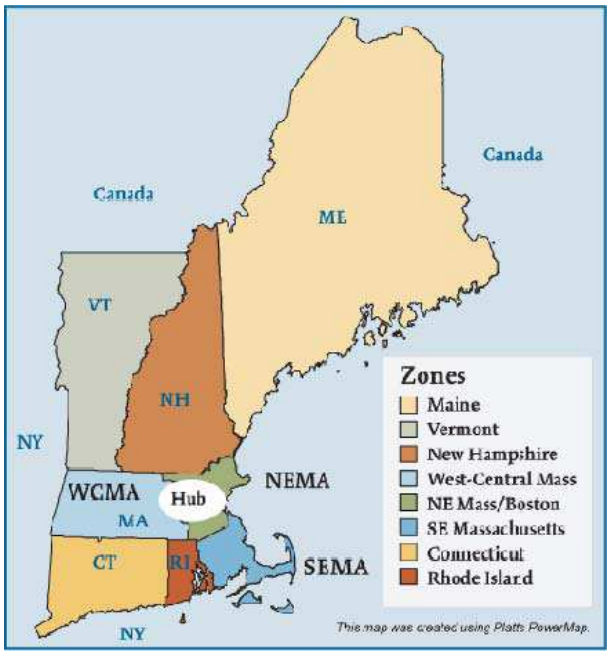

Figure 2.5. ISO New England load zones.

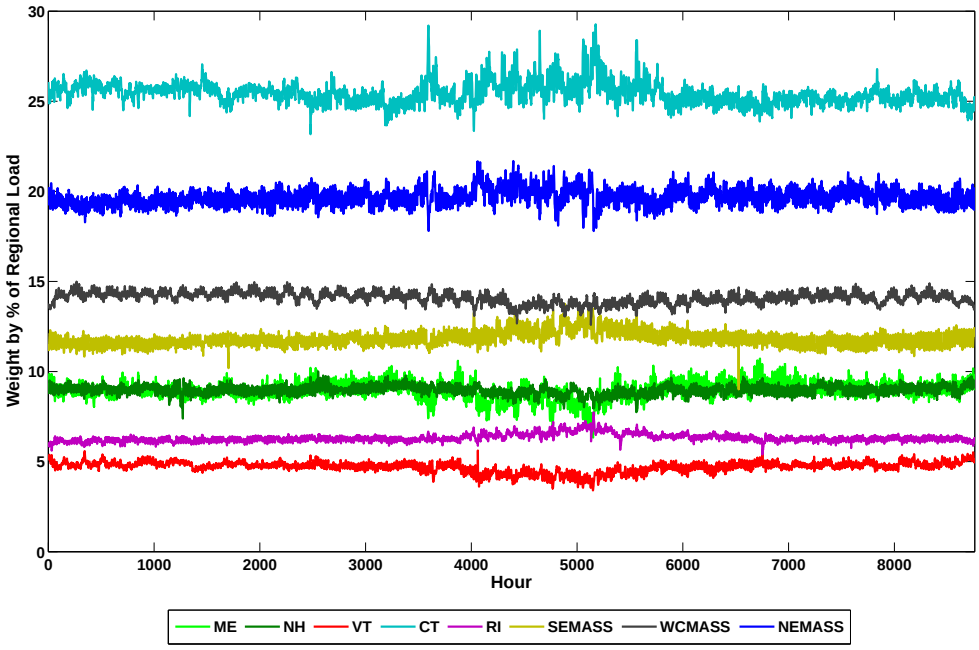

Figure 2.6. Hourly load zone weights - 2006.

usage patterns change over the course of decades due to changes in population, demographics, and consumer patterns, which are influenced by factors such as shifts in electricity-consuming technologies. For example, a dramatic shift in consumer patterns has resulted in the increased reliance on air-conditioning in New England over the last 30 years, which has led to much higher summer load peaks. A simple indicator of this change is a comparison of the annual summer-peak load factor for each year over the past three decades. The annual load factor is defined as the ratio of the average regional load over the course of the entire year to the summer peak demand. In general, a lower load factor is an indicator that a greater amount of dispatchable generation must be called upon to generate power infrequently, but nevertheless are needed to maintain system adequacy. As can be seen in Figure 2.8, these values have diminished significantly over the last 30 years, corresponding mostly to the increased reliance on air-conditioning in summer months. Figure 2.7 shows the load timeseries

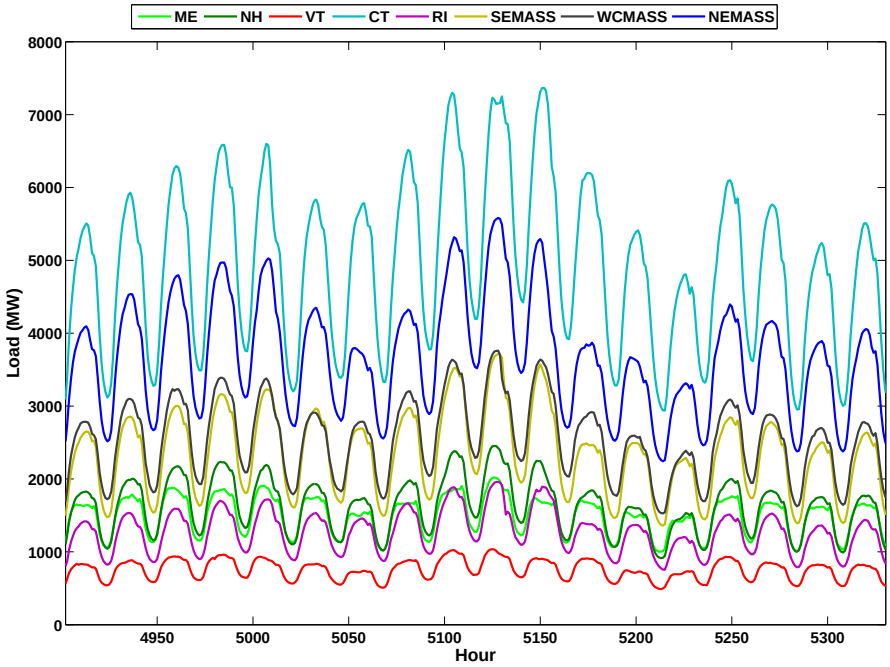

Figure 2.7. Hourly load zone MWs - Summer 2006.

for the eight load zones for approximately $2\frac{1}{2}$ weeks during the summer of 2006. Note that certain load zones have a greater peak load response, as represented by the dramatically steep profiles of the CT and NEMA load zones, whereas others such as VT and ME appear to have less of a peak response. This is certainly at least in part attributable to cooler weather in the northern part of New England than in the southern part of the region, but it may also be a function of differing load responses between subregions.

A comparison of the hourly time series for the years 1989 and 2006 was conducted in order to develop a sense of how the regional load curve has changed over the last couple decades. These years were selected since they are both non-leap years and start with the same day of the week, making them interchangeable in terms of the Gregorian calendar. This usually means that not only days of the week match, but also that holidays fall on the same days, resulting in easily comparable load data.

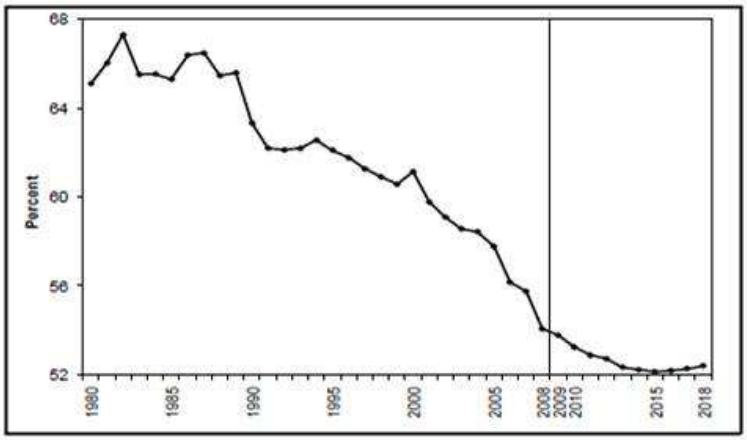

Figure 2.8. Annual summer peak load factor for New England, 1980 to 2018.

The time series of the regionally-averaged temperature was also compared for the two years in order to examine some of the weather-dependence of the differences in the load. Table 2.1 lists the maximum and minimum loads, mean load, and load factor for both years.

Table 2.1. Load characteristics - 1989 vs. 2006.

	Year	
Load Characteristic	1989	2006
Maximum Hour (MW)	19,722	28,130
Minimum Hour (MW)	7,011	9,171
Mean Hourly Value (MW)	12,785	15,079
Load Factor (percent)	64.8	53.6

Figure 2.9 is a plot of the coincident load (top) and temperature (below) for the years 1989 (blue) and 2006 (red) during one summer week (Monday, August 14 to Sunday, August 20). Since load in summer is highly temperature dependent, this week was selected because both years exhibit well-correlated temperatures, thus making the loads more comparable. Note the differences in the load shape between weekdays and weekend days — not only is the load magnitude different, but the diurnal shapes of the load curve are distinct, reflecting a different end use pattern. Some of this

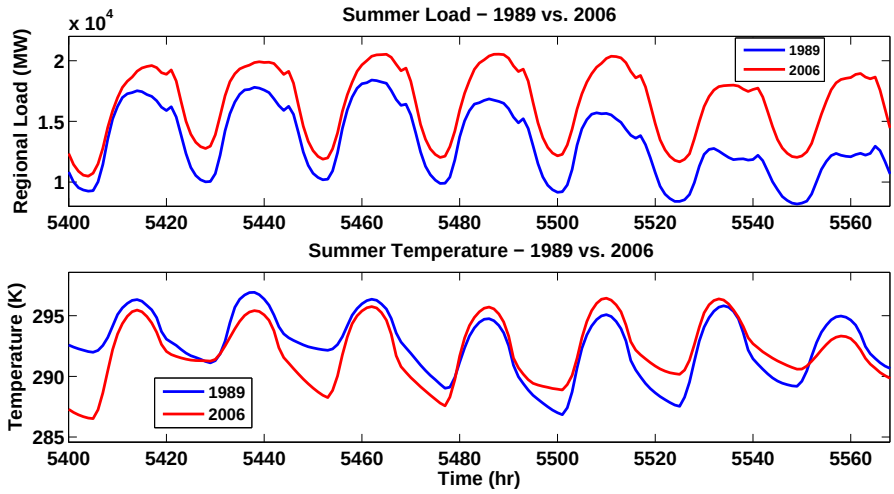

Figure 2.9. Load (top) and temperature (bottom) - August 14-20, 1989 (blue) and 2006 (red).

difference may also be attributable to other weather variables not shown, such as humidity.

By examining Figure 2.9 it is clear that the regional load has increased between 1989 and 2006. Plotting representative weeks for the other seasons (not shown) revealed similar trends. Overall, New England regional summer loads exhibit the most pronounced increase over this 17 year period, and winter load the least. It is likely that the disparity in seasonal trends can be attributed primarily to a greater reliance on air conditioning in summer, but perhaps also to the decrease in electric heating during the winter. Comparing the plots by days of the week, the weekend load has increased more than the weekday load. With respect to the diurnal load shape, no obvious changes from 1989 to 2006 are evident.

As illustrated in Figure 2.10, diurnal load characteristics vary significantly throughout the year. Since human activity is somewhat constant (save for holidays and vacations) the prime determinant of this variation is weather. Note the dramatic difference in both the timing and magnitude of the daily peak on a relatively high-load summer day (July 13, depicted in red) and the days representing the balance of the year.

Contrastingly, the load shape of April 19 (cyan line) exhibits the lowest load value for all hours. This suggests that there are times of year when the weather is having negligible influence on electricity demand. In other words, at near-optimal weather conditions (with respect to minimizing building energy demand), minimal energy is being used for the conditioning of building interior space. This phenomenon is observed in spring and fall. One could argue that with the exception of daylight effects (e.g. varying lighting needs due to varying length of daylight), the load during these times is exhibiting a fair degree of weather-independence. Exploiting the concept of weather-independence is a potential basis for determining the weather-sensitive load, which would be the load that remains. Given that the goal of this project is to use weather to predict load, the idea of separating weather-sensitive and weather-insensitive load could prove a useful method of effectively extracting the most salient features of the load-weather relationships.

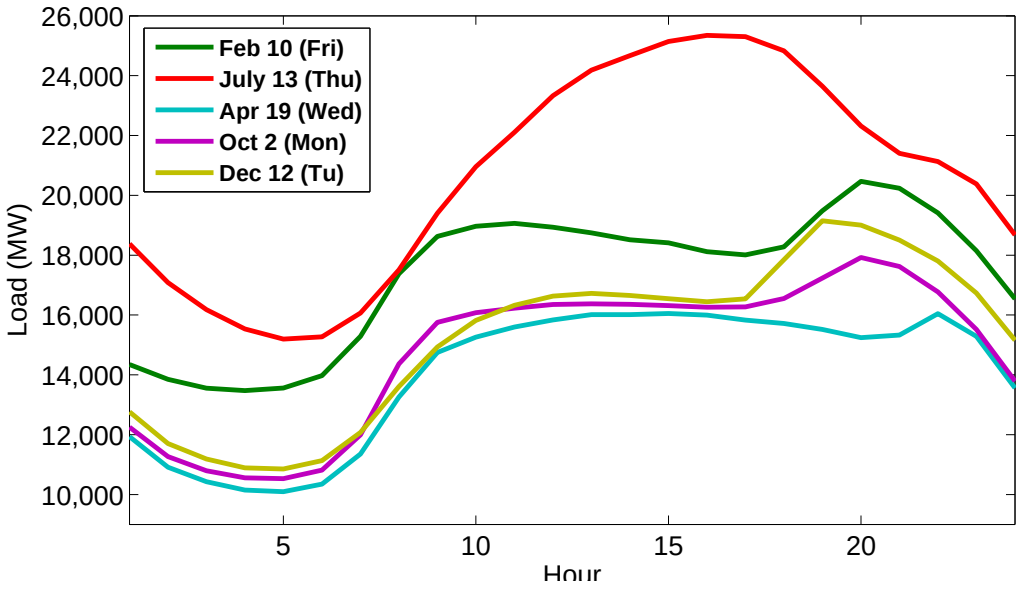

Figure 2.10. Diurnal load profiles for 5 weekdays interspersed throughout 2006.

2.3 Analysis of Load-Weather Relationships

As previously discussed, a significant portion of electricity demand is determined by the weather. This section describes analysis of load-weather relationships based on regional load and weather data from 2006. The weather-driven component of the load is highly nonlinear for many regions, and is a composite of the nonlinear and dynamic effects of weather variables like temperature and humidity, as well as deterministic effects of other variables such as cloud cover and wind speed [1]. It is hypothesized that solar insolation also factors into the load response due to the effects of solar heat gain on building energy consumption. The dynamic effects of temperature and dew point consist of the tendency of the current load to be a function of weather over previous time periods, and not simply the current weather. It is likely that this dynamic portion of the load is most prominent during summer months, due at least in part to patterns of air-conditioning usage that typically force the most nonlinear load response. In general, the relationship between load and weather is unique to each region and is a function of that region's climate.

Figure 2.11 is a scatter plot of regional temperature and load. Figure 2.12 is the same plot for only summer months – June, July, August, and September – which are designated by color. Figure 2.13 and Figure 2.14 are scatter plots of regional load versus humidity and solar insolation, respectively. A nonlinear relationship is evident for temperature and specific humidity, but the relationship between solar radiation and load is ambiguous. Significant (vertical) spread is observed in the load-temperature and load-humidity plots, but this is likely the result of varying relationships between load and weather predictors by season, day type (i.e. weekdays and weekend days), and by hour of the day (same day type). This was confirmed by the hourly load-temperature analysis conducted for weekdays described in Section 2.3.1.

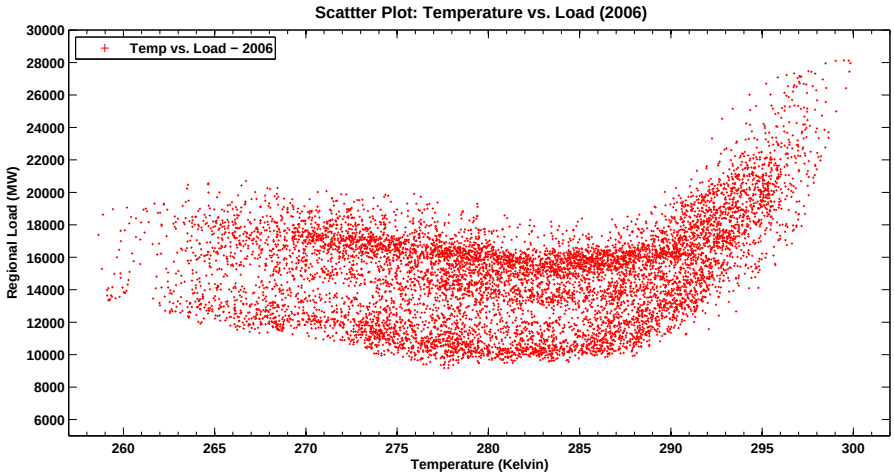

Figure 2.11. Scatter plot of temperature and load, 2006.

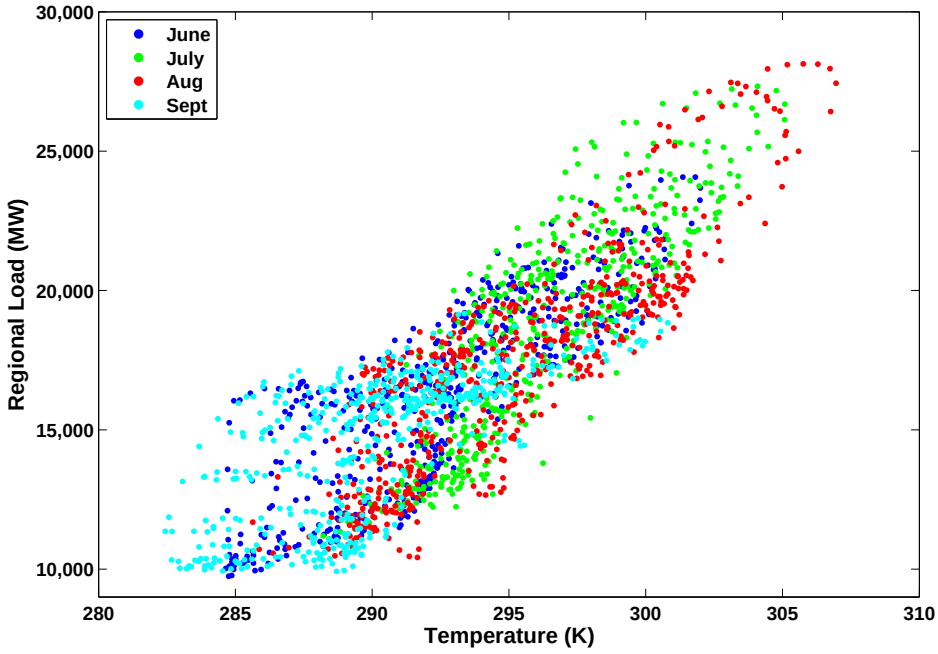

Figure 2.12. Scatter plot of temperature and load in summer, 2006.

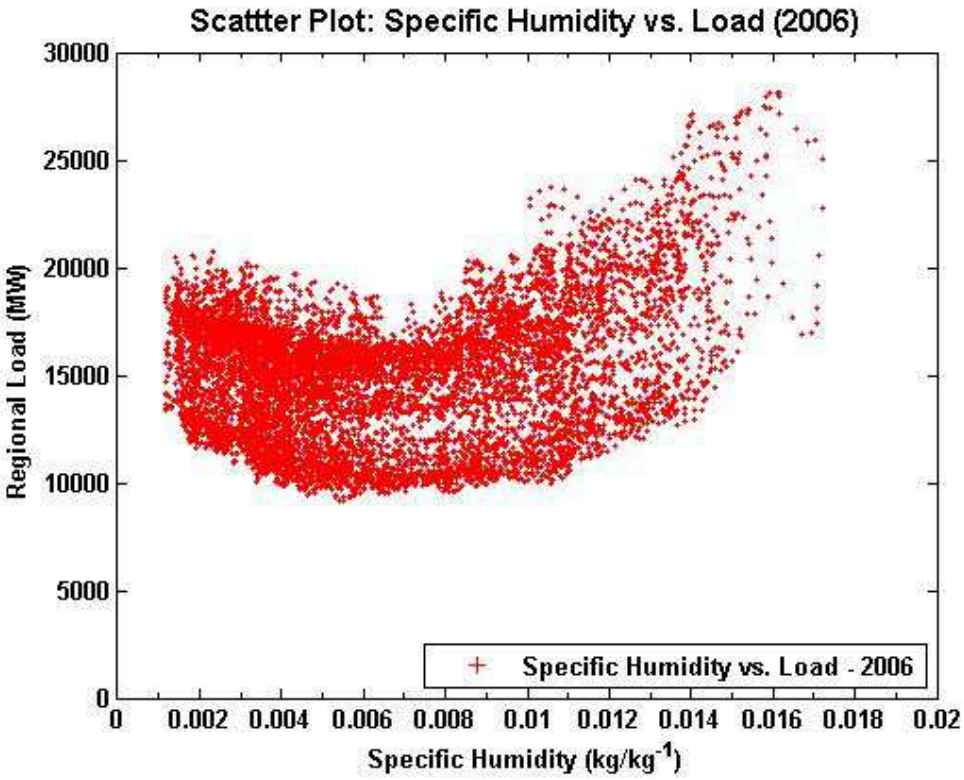

Figure 2.13. Scatter plot of specific humidity and load, 2006.

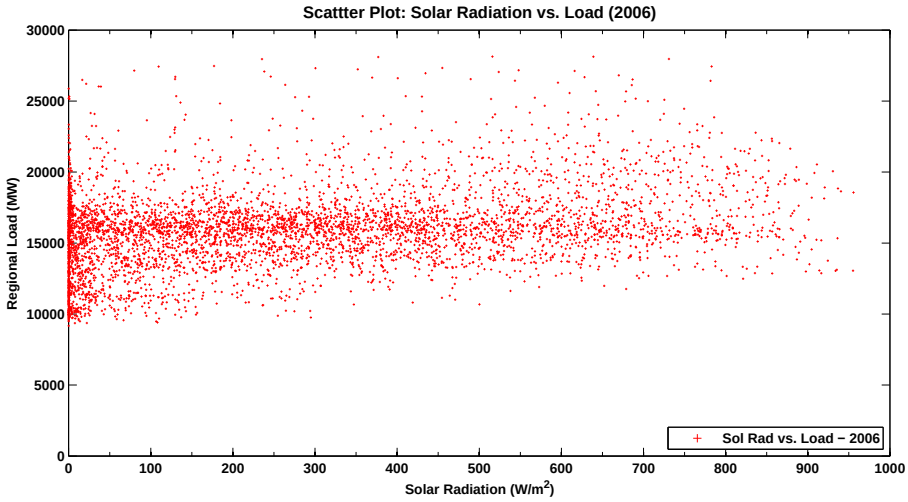

Figure 2.14. Scatter plot of solar radiation and load, 2006.

An initial analysis of the relationship between load and weather in New England in 2006 was conducted for three weather variables: dry bulb temperature, specific humidity, and solar insolation. Values used for each parameter consisted of the average of the hourly values for all the MERRA grid points shown in Figure 2.1 (11×13 points). Table 2.2 lists the cross-correlation values of the averaged weather variables and load. As these values indicate, the weather variables (i.e., the predictors) are highly correlated, especially temperature and specific humidity. This statistical phenomenon is referred to as multicollinearity, and could pose a problem when modeling since portions of the inputs are redundant in that they explain the same portion of the system response. It should also be cautioned that the values in the correlation coefficient matrix are simply the normalized measure of the strength of linear relationship between variables, and that correlation does not necessarily imply causation.

Table 2.2. Correlation matrix of load and weather variables, 2006.

	Temperature	Specific Humidity	Solar Radiation	Load
Temperature	-	0.922	0.373	0.267
Specific Humidity	0.922	-	0.165	0.299
Solar Radiation	0.373	0.165	-	0.383
Load	0.267	0.299	0.383	-

Figure 2.15 shows the 365 daily averages of load, temperature, specific humidity, and solar radiation for 2006. It is likely based on observation of the averages of the load and corresponding weather variables that some weather variables actively contribute to the seasonal periodicity exhibited by the load.

As stated above, several of the candidate weather variables are extremely collinear, e.g., temperature and humidity. One of the assumptions of most multiple-input models is that the inputs to the system are independent. Methods of addressing multicollinearity in predictors are specific to either the time domain or frequency domain. Since a time domain solution is desired, the method of principal component analysis (PCA) was reviewed; however, the PCA method would not be useful since it obscures

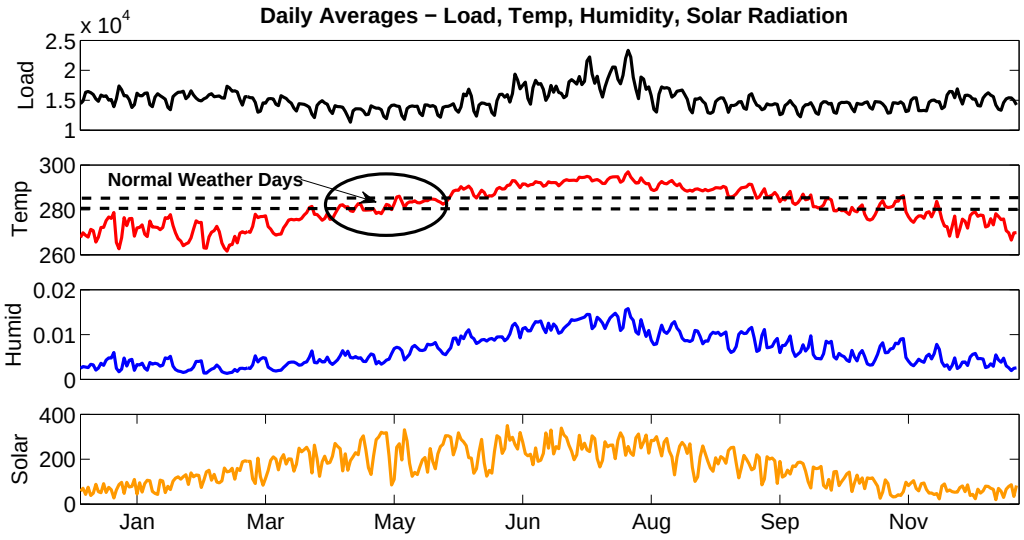

Figure 2.15. Plots of daily averages for load, temperature, specific humidity, and solar radiation, 2006.

the physical meaning of the predictors, which is not desired for hindcasting modeling. An example of a frequency domain solution is the conditioned spectral density function [4]. Given that time-domain analyses and modeling are sought for this project, the effects of collinear weather variables are not easily addressed. The researcher must consider this issue during model development and validation.

2.3.1 Hourly Load-Temperature Analysis

Since there is suspected variation in load-weather relationships by season, day type (weekday vs. weekend day), and perhaps even by hour of day, a more granular analysis is warranted. To evaluate this, a time of day analysis of load and temperature was conducted for weekdays only. Figures A.1 and A.2 in Appendix A are scatter plots of the results. The data points are color-coded by season as indicated in the accompanying legend. Please note that each hourly plot represents an equivalent area of Cartesian space (i.e., the product of delta-x and delta-y depicted for each hour is the same) which enables direct comparison of the slopes of each seasonal component

for all hours. These figures illustrate a crisper relationship between temperature and load than Figure 2.11, which exhibits significant vertical spread. (Outliers present in the figures are likely the result of holidays, which display dramatically different load shapes than typical weekdays.) This not only confirms that the suspected source of the vertical spread in Figure2.11 is mostly attributable to differing levels of human activity throughout the day, but it also suggests that in developing a weather-based hindcasting model, a time-of-day based approach will likely offer greater fidelity. For example, as Figures A.1 and A.2 illustrate, the load-temperature relationship varies slightly over the course of the day.

In order to get a better grasp of how load-weather relationships change throughout the course of a weekday, Figure 2.16 depicts quadratic fit curves that were developed based on the load-temperature scatter plots for hours 1, 4, 7, 10, 12, 15, 19, and 23. As is clearly illustrated, these relationships are different from hour to hour. The results of this analysis support a time-of-day (by hour) modeling approach to capture the unique hourly weather-load relationships.

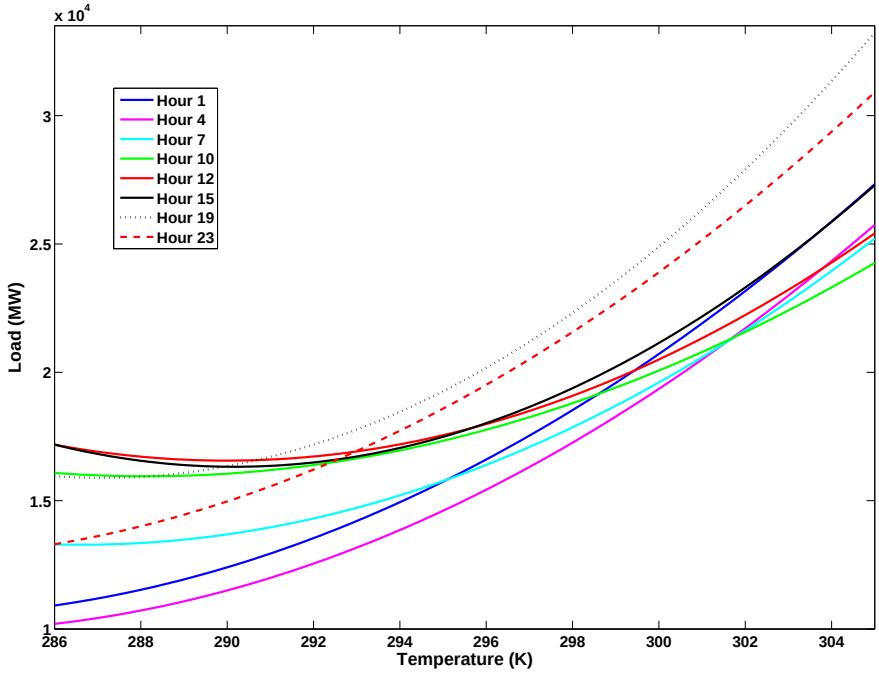

Figure 2.16. Quadratic fits of temperature-load relationships by hour - 2006.

CHAPTER 3

MODELING APPROACH AND PROBLEM FORMULATION

Based on the analysis of weather and load relationships discussed in previous sections, the load response due to weather is a function of time-of-day, and is likely a nonlinear, non-stationary system. Furthermore, it is possible that the load response differs between the subregions of New England. What follows is a brief description of the load response system, the by-hour and by-subregion modeling approaches selected, and some of the distinguishing features and potential applications of the hindcast model.

3.1 System Description

The system to be modeled is New England's regional load response during the summer period, which is defined as the months of June, July, August, and September. While the system is very likely non-stationary in nature, it exhibits somewhat homogeneous behavior over time, i.e., although its output does not have a constant level, it is cyclical with consistent periodicities (e.g., daily, weekly, and seasonal). The load response system contains both deterministic and stochastic components, including yearly, seasonal, weekly, and intra-day cycles. Additionally, the load response exhibits relationships with current weather and associations with each period of day. There is also a relationship between the current electricity demand and its past values, which are the result of the dynamic effects of weather occuring during the preceding period. Thus, variables used to estimate the load response may include weather

from the previous hour, day or week, representing the dynamic aspect of the load response. Figure 3.1 is a simple model illustrating the multiple-input/single-output (MISO) system to be modeled.

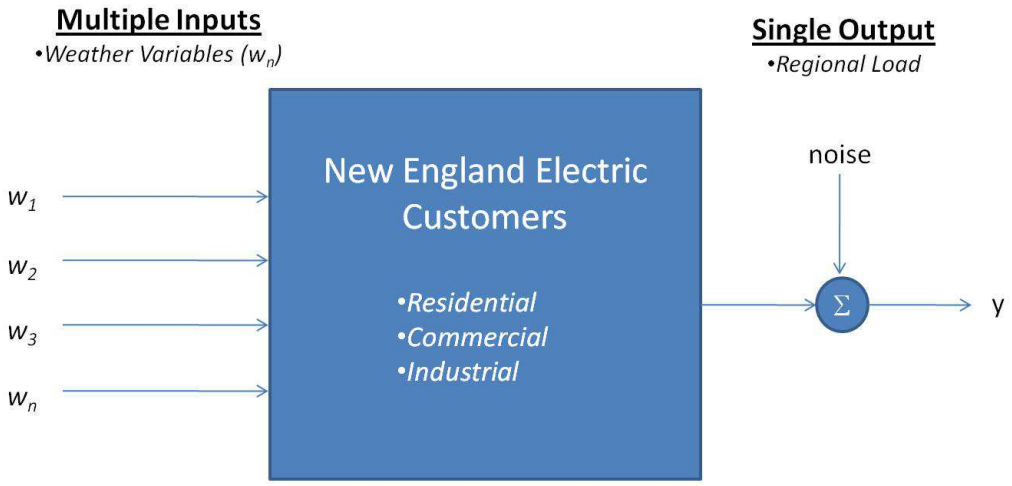

Figure 3.1. Basic system model - multiple input, single output (MISO) system.

3.2 By-Hour Approach

Since load is the direct result of human activity, it exhibits distinct diurnal patterns that are unique for weekdays, weekend days, and holidays. Non-holiday weekdays, the focus of the modeling for this research, follow a general pattern, with occasional exceptions during Monday mornings and Friday afternoons, as well as weeks that are heavily vacationed. As demonstrated in Chapter 2, the result of the link between load and human activity is that the relationship between load and weather changes over the course of a weekday, making it a nonlinear, non-stationary system. However, further analysis of the load-weather relationship reveals a distinct and consistent diurnal pattern for each individual hour (refer to Figure 2.16). This pattern will be exploited by taking a by-hour approach to hindcasting load, whereby each

weekday hour will be modeled separately using hourly bins of weather and load data. It is assumed that modeling the load response for each hour separately will effectively enable the weather's influence to become more apparent.

3.3 By-Subregion Approach

Given that high-quality, subregional load data is publicly-available, regional load will be represented as the sum of load responses that are modeled independently for each ISO-NE load zone. This approach has been chosen for three reasons: (1) Different load zones likely have a different mix of residential, commercial, and industrial electricity consumers that result in a different load profile, and (2) load-weather relationships may differ between sub-regions (for reasons other than differing mixes of customer classes, which are discussed below), or (3) some combination of the two aforementioned reasons. Therefore, it's possible that each subregion may have a unique load response model.

Due to the variety of weather conditions experienced throughout New England, each sub-region's built environment (where most end use of electricity occurs) may embody different characteristics. For example, since Connecticut is typically more hot and humid than Maine in the summer, the buildings in each state may be designed and equipped differently so that they deliver the cooling demand typical of each. If this is so, it could translate into a different load response when Maine has an unusually hot and humid period as opposed to Connecticut's load response when it experiences the same weather. Modeling each load zone separately thus lends the model-building process a sensitivity to these potentially different load responses. If unique models result for each zone, it may be assumed that the aggregate regional load response will exhibit greater adaptability to these subregional differences than if a static weather-weighting scheme was implemented.

Another possible justification for the by-subregion approach is that differing state policies (e.g., building codes, energy efficiency goals/incentives, and renewable portfolio standards) may affect the design and/or operation, and therefore performance, of building construction.

Figure 3.2 shows the MERRA data locations selected to represent the weather for each load zone. Locations were hand-selected based on the distribution of load intensity throughout the region illustrated in Figure 2.4. Below is a list of the locations listed by load zone:

1. SEMA – 4 MERRA gridpoints, indicated in orange
2. NEMA – 2 MERRA gridpoints, indicated in blue
3. WCMA – 1 MERRA gridpoint, indicated in yellow
4. CT – 2 MERRA gridpoints, indicated in red
5. RI – 1 MERRA gridpoint, indicated in green
6. ME – 3 MERRA gridpoints, indicated in white
7. NH – 2 MERRA gridpoints, indicated in teal
8. VT – 1 MERRA gridpoint, indicated in pink

All weather inputs are averaged for each timestep over all the MERRA gridpoints for each load zone. Note that the locations selected for each subregion are open to interpretation, and that a different combination of gridpoints may in fact be more suitable.

The sum effect of the by-hour, by-subregion modeling approach is illustrated in Figure 3.3, which depicts the NEMA subregion's hourly load (top plot), temperature (second plot), specific humidity (third plot), and solar insolation (bottom plot) during

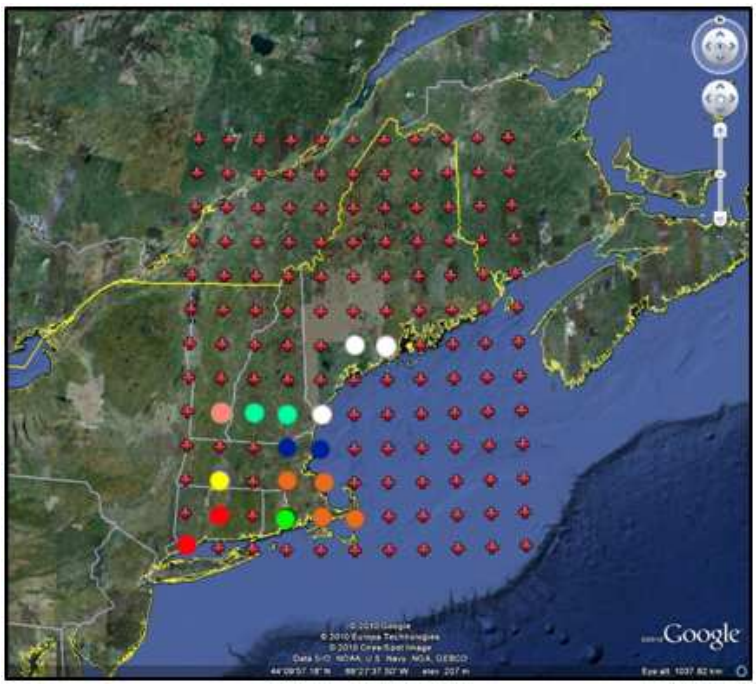

Figure 3.2. The MERRA data locations selected for the eight load zones.

Hour 13 of the summer dataset. By binning weather and load data by hour of the the day for summer weekdays, these are the type of weather inputs that the hindcast model will be given.

3.4 Distinguishing Features of the Hindcast Model

Although load hindcasting is related to load forecasting, the following key differences exist between them:

1. The hindcast model will be used to predict load based solely on weather inputs and the load-weather relationships derived during model training. This likely requires more weather inputs than those typically used for load forecasting. For example, a number of lagged weather variables were created for building of the hindcast model for this project, and will be described in Section 4.2. Other

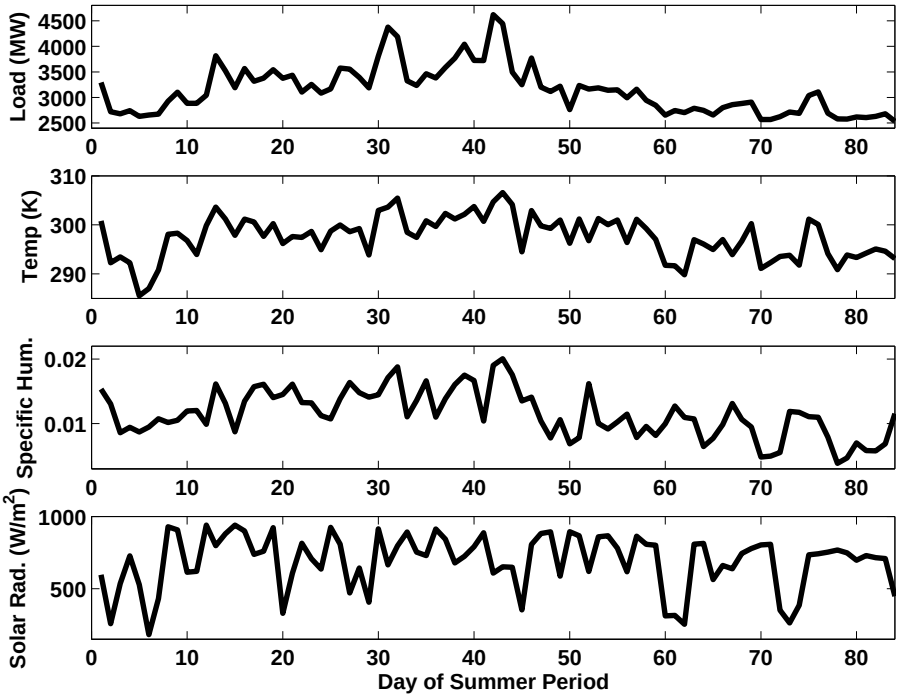

Figure 3.3. Load, temperature, specific humidity, and solar data for NEMA during Hour 13 of 84 days of Summer 2006.

non-weather related variables are commonly utilized in load forecasting, such as economic factors.

2. No online adjustments are available to the hindcast model since hindcast loads cannot be realistically compared to historic loads (see comparison of 1989 and 2006 loads above), and also cannot be reliably compared to itself. This means that extra care must be exercised to ensure that hindcast loads are not trending in an unrealistic fashion, since reliable feedback will not be available to the model, save for during training.

3. Weather inputs will be provided by the MERRA reanalysis dataset, rather than a weather forecast. The MERRA reanalysis offers homogeneity over multiple decades and a far more extensive list of possible weather predictors than weather forecasts provide to load forecasters.

4. Modeling of the regional load will consist of eight independent subregional models that will enable a sensitivity to unique load responses to weather in different parts of New England. Load forecasting typically involves the use of a relatively static weighting scheme to represent the different weather throughout the region, such as the weather weights used by ISO-NE that were discussed in Chapter 1.5.1. Weighting schemes such as these assume an identical weather-load relationship across a region.

3.5 Potential Applications of the Load Hindcasting Methodology

The load hindcasting methodology under development has potential applications to the characterization of weather-driven power generation resources including wind, solar, or hydro power. Specifically, the general approach of using a reanalysis dataset

such as MERRA affords data consistency and relative spatial granularity over large geographical regions.

Another possible application is in regional load forecasting. An interesting application to load forecasting would be to leverage the weather-load relationships of the hindcast model to explore the effects of distributed weather-driven power generation that is embedded in the distribution system. Since distributed generation is a 'behind-the-meter' generation source it is not under the control of a regional system operator like ISO-NE, and is only 'visible' to the system operator as a load modifier. Given that weather predictors such as wind speeds and solar insolation are available to the model, the hindcast methodology may lend itself to modeling this generation. Overall, the hindcast model may prove useful to power system operators, transmission or distribution system owners and operators, and/or academics conducting related research.

CHAPTER 4

MODEL-BUILDING METHODS

In the absence of a true functional relationship between weather inputs and the load response, an empirical model that uses data to approximate a function is needed. Multiple linear regression (MLR) is a common empirical approach and will be used to develop the load response model. This chapter describes the MLR modeling methods used, the creation of candidate regressors, variable selection methods, and parameterization (i.e., training) of the MLR models. All modeling was performed using the MATLAB computing environment, created by The MathWorks, Inc. [50].

4.1 Multiple Linear Regression Model

MLR is a method of developing relationships between predictor variables and system response by fitting empirical data to a linear model via the method of least-squares, sometimes referred to as ordinary least-squares (OLS). The general form of the MLR model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_k x_k + \epsilon, \tag{4.1}$$

where β_0 is the y-intercept, the β_k's are the linear coefficients or parameters for each predictor variable, x_k, and ϵ is the error term in the output of the system. Since there is more than one regressor (as opposed to simple linear regression where there is only one), each parameter β_k represents the change in response y due to a change in each each regressor x_k, when all other regressors are held constant. Therefore, each β_k is often referred to as a partial regression coefficient.

MLR offers the flexibility of including both higher-ordered single terms and interaction between terms. For instance, in the following generic expression:

$$y = \beta_0 + \beta_1 {x_1}^2 + \beta_2 x_1 x_2 + \epsilon \tag{4.2}$$

a second order term, $\beta_1 {x_1}^2$, and an interaction term, $\beta_2 x_1 x_2$, are a part of the model. This enables the handling of nonlinear relationships between predictors and response while maintaining a model that remains linear in each of its parameters.

All MLR modeling was conducted using MATLAB's *regress* function, which is available in the Statistics Toolbox.

4.1.1 Parameter Estimation Using Ordinary Least-Squares

As already mentioned, the method of ordinary least-squares (OLS) is used to estimate the regression coefficients of the MLR model. OLS is frequently used to approximate a solution for problems that are overdetermined, i.e., when there are more equations than unknowns [32]. Rather than seeking an exact solution, the goal in OLS is to minimize the sum of the squares of the *residuals*, which are defined as the deviations between the responses estimated by the model and the observed responses. In the application of OLS for load hindcasting, the residuals are the difference between the predicted and observed hourly loads.

Inherent to the use of MLR are probabilistic assumptions about the underlying error distribution of the regression model, which will be discussed further in Chapter 5. When these assumptions are valid, OLS produces what is known as the maximum-likelihood estimate of the parameters. Even when these assumptions are not valid, it has been demonstrated that OLS still produces useful results [33].

The mathematical background of OLS that follows in matrix notation has been adapted from Montgomery, Peck, and Vining [33]. Matrix notation allows for a convenient and compact expression of the MLR model. Assuming that there are n

observations and k regressors, such that $n > k$, the general MLR model in matrix notation is as follows:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{4.3}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In the above matrix notation, $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times k$ matrix of regressors, β is a $k \times 1$ vector of coefficients, and ϵ is a $n \times 1$ vector of errors.

Using OLS, we are seeking a vector of least-squares estimators, $\hat{\beta}$, that minimizes the following expression for the least-squares function:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = (\boldsymbol{\epsilon'\epsilon}) = \boldsymbol{(y - X\beta)'(y - X\beta)}$$

In the above expression, $S(\beta)$ can be written as follows:

$$S(\boldsymbol{\beta}) = \boldsymbol{y'y - B'X'y - y'X\beta + \beta'X'X\beta = y'y - 2\beta'X'y + \beta'X'X\beta}$$

This is because $\beta'\mathbf{X}'\mathbf{y}$ reduces to a scalar, and therefore, its transpose $(\beta'\mathbf{X}'\mathbf{y})'$ $= \mathbf{y}'\mathbf{X}\beta$ is the same scalar. Therefore, the least-squares estimators needs to satisfy the following expression:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\boldsymbol{X'X}\hat{\boldsymbol{\beta}}$$

which simplifies to:

$$\boldsymbol{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{4.4}$$

Equation 4.4 is referred to as the set of *least-squares normal equations*, which are solved by multiplying both sides of the matrix by the inverse of $\mathbf{X}'\mathbf{X}$, which makes the least-squares estimator of β become

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{4.5}$$

If all regressors are linearly independent, no column of $\mathbf{X}$ will be a linear combination of any other columns, and the inverse matrix $(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$ exists, which means that Equation 4.5 is solvable.

MATLAB's *regress* function uses an orthogonalization algorithm known as *QR factorization* to compute the least squares solution in order to avoid the potential for the normal equations to become singular even when the columns of $\mathbf{X}$ are independent [32].

4.2 Potential Design Variables

As was already discussed in Section 1.5, the load response to weather is a dynamic process, which means that in addition to the current weather, previous weather inputs (i.e., lagged weather) affect the current load response. This dynamic aspect of the load response is attributable to the physics of building energy and heat transfer. The overall heat transfer coefficient of a building includes conduction, infiltration, ventilation and solar gain effects [3]. In warm and humid climates, the cooling load has both a sensible (the cooling of dry air) and latent (the cooling and ultimate removal

of water vapor from humid air) component. As would be expected, the latent cooling load increases as a function of ambient specific humidity, as does the portion of the total cooling load that is latent [40]. To reduce the amount of latent cooling load that must be provided to a building, cooling systems are typically designed to minimize the amount of fresh air intake while also maintaining sufficient air quality within the occupied environment. Additionally, heat flow rates through building envelopes vary continuously and unpredictably due to changes in the temperature gradients and thermal storage effects, which give rise to a phenomenon known as thermal lag [39]. Below is a summary of some of the weather-related factors that can influence the amount of cooling load required for an individual building:

1. The influence of ambient air temperature on the conduction/convection of heat through the building's thermal envelope, which can have lagged effects due to thermal storage

2. The influence of ambient air humidity on the exchange of moisture through the building envelope

3. The amount of solar insolation passing through windows (solar heat gain), and solar insolation's heating effects on wall and roof surface temperatures, which in turn affect the conductance of heat through the building's thermal envelope

4. The perception of acceptable comfort, which may be influenced by weather factors such as the persistence of severe weather

While load forecasters do not explicitly attempt to model building energy dynamics directly, the summer load response is predominantly driven by air conditioning load. Therefore, modeling how these dynamic aspects of building energy translate into the electric load response may be considered the most challenging aspect of electric load estimation. Furthermore, the lagged, interrelated influence of different

weather variables on summertime building energy consumption suggests that a realistic model of the electric load response likely consists of a highly complex transfer function. Unfortunately, load forecasting literature offers definitive guidance on neither which lagged weather variables to include in a model nor their most appropriate form(s), i.e., whether they should be singular hourly values, time averages over a number of hours, or deviations from previous timesteps.

Of the array of weather variables available in the MERRA dataset, the design variables under consideration for inclusion in the MLR hindcast model include dry bulb temperature (T), specific humidity (H), and solar insolation (S), which are assumed to be the primary weather influences during the summer period. (Although other weather factors are likely also an influence, e.g., wind speed, these are assumed to be secondary influences that should be incorporated (if needed) after the base model is developed.) Specific humidity is defined as the ratio of water vapor (m_v) to the mass of dry air (m_a) in a particular parcel of air. Since the specific humidity is in units of mass per unit mass, it is sometimes referred to as the mass mixing ratio.

The following expression is an example of the use of lagged weather variables in a MLR summer weekday model that is taken from Soliman et al [46]:

$$\begin{aligned} Y(t) =& a_0(t) + a_1(t)T(t) + a_2(t)T^2(t) + a_3(t)T^3(t) \\ &+ a_4(t)T(t-1) + a_5(t)T(t-2) + a_6(t-3) + a_7(t)H(t) \\ &+ a_8(t)H(t-1) + a_9(t)H(t-2) \end{aligned} \tag{4.6}$$

Where:

$Y(t)$ =load at time t;

$T(t)$ =temperature deviation at time t (explained below);

$H(t)$ =the humidity factor (degrees Celsius)

$a_0(t)$ =base load at time t;

$a_1, a_1(t), ..., a_9(t)$are the regression parameters to be estimated at time t

The temperature deviation is calculated as the difference between the dry bulb temperature at time t and the average dry bulb temperature of the previous 20 weekdays, and the humidity factor is essentially a temperature-humidity index (THI), and is calculated as follows:

$$H(t) = 0.55T_d(t) + 0.2T_p(t) + 5.05 \tag{4.7}$$

Where T_d is the dew point temperature. Note that the humidity factor is set to zero if the dry bulb temperature is less than 25^o C, which is assumed to be room temperature.

The expression above may be viewed in compact form as:

$$Y(t) = f^T(t)X(t) \tag{4.8}$$

Where:

$$f(t) = \left\{ \begin{array}{c} 1 \\ T(t) \\ T^2(t) \\ T^3(t) \\ T(t-1) \\ T(t-2) \\ T(t-3) \\ H(t) \\ H(t-1) \\ H(t-2) \end{array} \right\} \tag{4.9}$$

And $X(t)$, the estimated parameter vector is of the form:

$$X(t) = \begin{Bmatrix} a_0(t) \\ a_1(t) \\ \vdots \\ a_8(t) \\ a_9(t) \end{Bmatrix} \tag{4.10}$$

Preliminary use of the Soliman model as a potential hindcasting model revealed that its performance exhibited large residuals, especially during peak load hours, and was therefore considered unacceptable.

Given the lack of understanding regarding what forms of lagged weather variables most clearly represent the load response, one objective of this research was to explore how a number of different types of lagged weather variables contribute to a model's predictive performance. To achieve this, a set of different form(s) of lagged weather was created for potential inclusion in the model. In addition to the current weather variables, the following different forms of lagged weather variables were made available to the model:

1. Values corresponding to one, three, and 24 hours ago – For example, 'Sm3' is the solar radiation three hours prior (i.e., 'Sm3' stands for S minus 3 hours) and 'Tm24' is the dry bulb temperature 24 hours ago.

2. Average values over the past three hours, 24 hours, 48 hours, 72 hours, and 168 hours. For example, 'T3' is the average temperature over the past 24 hours, and 'H72' is the average specific humidity over the past 72 hours.

3. The change in value from 24 hours ago. For example, 'delTm24' is the change in dry bulb temperature from 24 hours ago (i.e., 'delTm24' stands for the delta between 'T0' and 'Tm24').

Table 4.1 lists a total of 84 potential variables that were considered for the load hindcasting model by category.

Table 4.1. Potential variables

Variable Type	**Potential Variables**
Order 1	T0, Tm1, Tm3, Tm24, T3, T24, H0, Hm1, Hm3, Hm24, H3, H24, S0, Sm1, Sm3, S3, S24, T48, T72, T168, H48, H72, H168, S48, S72, S168, delTm24, delHm24
Order 2, 1 Term	($T0^2$), ($H0^2$), ($S0^2$), ($T3^2$), ($H3^2$), ($S3^2$), ($T24^2$), ($H24^2$), ($S24^2$)
Order 2, 2 Terms	(T0)(H0), (T0)(S0), (H0)(S0), (T3)(H3), (T3)(S3), (H3)(S3), (T24)(H24), (T24)(S24), (H24)(S24), (T0)(H3), (T0)(H24), (T0)(S3), (T0)(S24), (T3)(H0), (T3)(S0), (T3)(H24), (T3)(S24), (T24)(H0), (T24)(H3), (T24)(S0), (T24)(S3), (H0)(S3), (H0)(S24), (H3)(S0), (H3)(S24), (H24)(S0), (H24)(S3), (Tm1)(Hm3), (Tm1)(Hm24), (Tm1)(Sm3), (Tm3)(Hm1), (Tm3)(Sm1), (Tm3)(Hm24), (Tm24)(Hm1), (Tm24)(Hm3), (Tm24)(Sm1), (Tm24)(Sm3), (Hm1)(Sm3), (Hm3)(Sm1), (Hm24)(Sm1), (Hm24)(Sm3), (delTm24)(delHm24)
Order 3, 1 Term	($T0^3$), ($H0^3$), ($S0^3$)
Order 3, 3 Terms	(T24)(H24)(S24)
Order 4, 2 Terms	($T24^2$)($H24^2$)

Sm24 was not included in the model since solar insolation in a given hour can be a highly discrete phenomenon in that it can exhibit poor correlation with solar insolation in other timesteps. In contrast, average solar insolation over the previous 24 hours ('S24') is a potential indicator of the heat absorbed by the thermal mass of the occupied built environment, thereby playing a role in aggregate consumer cooling demand.

4.3 Variable Selection

Model building, also known as the variable selection problem, involves two conflicting objectives: (1) the need to include a large number of regressors to ensure that their information content adequately represents the response, and (2) a desire to include the fewest regressors in order to minimize the variance of the response estimation, which increases with the number of regressors [33]. The challenge of building a viable model is further exacerbated when the basic form of the model is unknown, the set of candidate regressors is large, and the correct functional form of the regressors is also unknown.

Ideally, all possible combinations of candidate regressors would be tested to find the most optimal subset, a model selection method called *all possible regressions*; however, the total possible combinations of N number of candidate regressors is equal to 2^N. Therefore, with 84 candidate regressors, the entire search space covers 1.9×10^{25} potential solutions, a value that far exceeds the capabilities of even the most advanced parallel computational algorithms. In light of this computational limitation, a method of selecting the variables to include in the MLR model is needed. In addition to the all possible regressions method already noted and ruled out, many variable selection methods have been proposed, including: best subsets regression, stepwise regression, ridge regression, principal components regression, latent root regression, and stagewise regression [10]. These are typically considered the standard variable selection methods. In addition to these, some nonstandard variable selection methods have also been proposed, including genetic algorithms and simulated annealing [22]. These nonstandard methods are actually global optimization techniques that have been repurposed from the field of optimization.

All of the aforementioned methods of selecting the best regression equation from a set of candidate regressors are limited and are not guaranteed to yield the identical, most optimal solution. This is especially true as the number of candidate regressors increases. For this reason, research aimed at developing new or improved selection methods remains active. In consideration of this, the following model building approach was developed:

1. Choose two variable selection methods, one standard and one nonstandard.
2. Separately develop an hour-based model for all eight load zones using each method.
3. Compare the performance of both methods via residual analysis and performance metrics.

This enables the comparison of two separate models, which may allow a deeper exploration of the explanatory potential of candidate regressors as well as a couple of different approaches to model building. The two variable selection methods that will be compared are stepwise regression and a genetic algorithm-based selection, which are described further in Sections 4.3.2 and 4.3.3.

Recall that due to the by-hour, by-subregion approach that has been selected, the variable selection task entails searching for the best subset of regressors to predict the load response during one particular hour within an individual load zone. This will result in a total of 24 regression equations for each load zone, and a total of 192 regression equations for the region for each variable selection method. (Additionally, these equations were parameterized using three different summer datasets, resulting in 1,152 unique regression equations!)

4.3.1 Variable Scaling

In order to ensure that equation 4.3 contains a well-conditioned $\mathbf{X}$ matrix, all weather variables are scaled between zero and one. Giving candidate regressors a consistent range of values lends them all the same basis for determining variance, which is key to model parameterization. Additionally, a convenient by-product of this scaling is that the final parameters estimated by OLS are easily comparable and not influenced by units of measurement.

4.3.2 Stepwise Regression Variable Selection

Stepwise regression-based variable selection for MLR uses measures of statistical significance to sequentially add or remove candidate regressors to determine the 'most' statistically significant subset of regressors. This means that stepwise regression is guided by the in-sample fit of the regressor matrix and response variable. The common tests for statistical significance in MLR are p-values and partial F-tests, which are both part of the analysis of variance (ANOVA) suite of statistical measures that are

used in regression. Thus, p-values and partial F-tests are measures of a variable's contribution to the overall variance of the response variable, which is often simply referred to as a variable's *contribution*.

A p-value is a statistical inference test that is a measure of the strength of evidence against the null hypothesis, H_0. The null hypothesis is the assertion that the significance shown by means of a variable's p-value is the result of random chance alone. The smaller the p-value, the stronger the evidence against the null hypothesis, and the more 'significant' a variable is in terms of explanatory power [48]. The threshold for significance, α, can be selected by the user, but it is typically $\alpha < 0.05$; however, it is not certain in actuality that a p-value lower than this threshold necessarily implies significance, but rather that there is a high *probability* that it does based on the data considered. For this reason, the use of the p-value has been criticized as a somewhat arbitrary tool. This valid criticism becomes somewhat less applicable as the values for α become smaller, since the probability that the null hypothesis is wrongly rejected is also reduced.

A partial F-test (or just 'F-test' if there is only one variable in the regression) is a measure of the contribution of a newly added variable given that all other regressors selected previously are in the model. Stepwise regression uses the partial F-test iteratively as each variable is added or removed from the regression, and therefore, it is actually used to measure the effect of *sets* of regressors.

It should be noted that the partial F-test possesses maximum fidelity in its variable selection guidance when the data corresponding to each variable are orthogonal to that associated with other variables in the $\mathbf{X}$ matrix, which indicates that there is linear independence between all of the columns of this matrix [33]; however, true independence rarely occurs in actual data. Consequently, when a near-collinear relationship exists between variables, which is the case with the weather variables avail-

able for selection, this test's abilities degrade. This should be noted by the user of stepwise model building.

Although there are inherent limitations to using p-values and partial F-tests to guide variable selection, experience has shown that this approach is still valid even when the aforementioned conditions that warrant caution are present. As such, stepwise regression has remained in widespread use for decades.

The procedural logic that typifies stepwise regression is described as follows [10]:

1. Determine the first variable to enter into the regression – Calculate the p-values of all predictor variables with the response. The variable with the lowest p-value is entered into the regression.

2. Determine if the regression equation is significant – Perform an OLS fit with the single regressor and perform an F-test to determine if the regression equation is significant. If it is not, terminate the stepwise regression and use a constant response model where the response is the mean of the observed responses. Otherwise, proceed to the next step.

3. Find the next variable to add – Calculate the p-values of all the currently excluded predictor variables with respect to the response that remains unexplained by the included regressors, and enter the variable with the lowest p-value into the regression model.

4. Determine if the new regression equation is significant and test for the least useful predictor currently in the equation – Perform OLS fit with both regressors and calculate a partial F-test value for both regressors now included in the model. Test whether removing the regressor with the lowest F-value would improve model significance. If so, remove that regressor and proceed to the next step; if not, proceed to the next step.

5. Repeat steps 3 and 4 until no further variables should be added or removed. The 'best' subset of regressors are those included in the regression model when the stepwise regression is terminated.

Both entry and exit p-value tolerances are chosen by the user to guide the stepwise procedure. Again, while the stepwise regression approach to model building is standard to most regression analyses, it is not guaranteed to yield the optimal subset of regressors.

4.3.2.1 MATLAB Stepwise Regression

Stepwise variable selection was performed using MATLAB's *stepwisefit* function. The procedural logic that this function uses to perform the stepwise regression is consistent with what was described above. MATLAB's default p-value thresholds used for variable entry ($\alpha < 0.05$) and exit ($\alpha < 0.1$) were used. These are considered conservative values, in that they tend to allow variables whose contributions have slightly weakened after the addition of a new variable to be retained in the model.

Stepwisefit allows the user to specify an initial set of variables to begin the stepwise regression. Using this feature, different initial sets of variables will sometimes yield different results, although there is no guarantee that this will produce a better fit [50]. Although this feature was not incorporated into this research, it would be of value to test its effect on stepwise variable selection, and is a worthwhile topic for future work. Regardless, this algorithm's sensitivity to different initial sets of variables further validates the limitations of the stepwise method, while also offering a potential approach to further maximizing its effectiveness by identifying the optimal subset of variables.

4.3.3 Genetic Algorithm-Based Variable Selection

Given the limitations of the stepwise regression method, a second variable selection method based on a genetic algorithm (GA) was developed. GAs were originally

invented by John Holland in the 1960s as a method to apply the mechanisms of natural selection to computer systems [31]. Although the field of evolutionary computation already existed, Holland's major contribution was his introduction of a population-based algorithm that included genetic-inspired operators such as "crossover" and "mutation". The use of GAs as an global optimization tool for solving engineering problems has since become widespread. Further, the use of GAs as a nonstandard method of variable selection in regression models has already been proposed [22] [15]. Therefore, it was germane to determine if this method is in fact well-suited to the combinatorial problem posed by the selection of regressor variables from a large set of candidates. While the use of GAs is typically more relevant to the optimization of continuous-valued functions, the aforementioned research has shown significant value in applying GAs to the highly discrete-valued search space of the variable selection problem, especially as the number of candidate regressors increases.

The basic elements of a GA consist of an initial population that is randomly generated and the selection of evolving populations of new offspring based on the concept of *fitness*, which uses one of several available methods of crossover among 'fit' individuals, and random mutation. Below is a summary of the basic components of a GA [31], including (where necessary) a brief description of how they were adapted to the needs of the variable selection problem:

Genes – The building blocks of chromosomes; within the context of variable selection a single gene represents one of the candidate regressors.

Chromosomes – Strings of bits (or genes) that represent the genetic code of one individual. As each individual represents a potential solution to the variable selection problem, it is an 84-character binary string representing all the candidate regressors, or genes, in which a '1' signifies an active regressor, and a '0' an inactive one.

Initial Population – A randomly generated population of chromosomes. In the context of variable selection, each individual represents a possible combinatorial solution to the selection problem.

Fitness Function – A function that assigns a score to each individual chromosome in the current population that is a measure of how well they solve the problem at hand. In the context of variable selection, a fitness function was designed to be a measure of an individual's out-of-sample predictive performance (described further below).

Parents – Individuals that have the best fitness that are selected to breed.

Offspring – The new individuals that are created via crossover and mutation (described below). These are the new combinatorial solutions that are created by crossover/mutation amongst the most fit variable subsets found within the current generation.

Crossover – The varieties of breeding mechanics that are used by GAs, which include both uniform crossover and one-point crossover. One-point crossover is the swapping of genetic material between parents at a single point such that all genes either before or after that gene location (first parent before, second parent after) are swapped. The crossover location is randomly generated. Uniform crossover exchanges half the genes between parents to form offspring between multiple crossover points. The locations of the crossover points are randomly generated.

Mutation – The random flipping of a specific gene's operator, the occurrence of which is determined by a specified mutation rate. In the context of variable selection, mutation implies a variable either being turned off or on, depending on its initial state.

Maximum Fitness – This is simply the maximum fitness that the GA has found.

Generation – Each successive population of offspring that results from both crossover and mutation. A maximum number of generations is specified by the GA user, at which point the GA algorithm terminates. (When a GA is utilized for the global optimization of a continuous-valued function, the GA algorithm can also be instructed to terminate when a specified number of successive generations do not increase the maximum fitness found.)

As was mentioned above, the out-of-sample performance of an individual combinatorial solution was selected as the measure of fitness to implement into the GA's fitness function. To do this, the regression variables for each chromosome were used to build a parameterized regression model with one summer dataset (the in-sample data), and the model was used to predict the load response using the weather of a second summer dataset (the out-of-sample data). Performance was measured using the mean absolute percent error (MAPE), which is expressed as follows:

$$MAPE = \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{4.11}$$

where y_i is the load response estimated by the model and $\hat{y}_i$ is the observed value, and $y_i - \hat{y}_i$ is the expression for the residual. The MAPE over all hours of prediction was selected as the out-of-sample performance metric.

Preliminary tests of this form of the fitness function (using a single summer dataset for training, and a single summer for performance assessment) revealed a tendency to overfit the in-sample data used for parameterization. Overfitting occurs when a statistical model begins to describe random error associated with the data, rather

than the underlying relationships between regressors and response. This tendency was discovered by assessing the parameterized model's predictive ability on a second out-of-sample summer dataset, which was almost invariably much poorer than the performance during the first out-of-sample prediction (because MAPE during this period was defined as the fitness to be minimized). By reversing the roles of the datasets, it was further discovered that different hourly models could result depending on the training/prediction years chosen, i.e., training with 2005 data and predicting using 2007 weather data would yield a different set of regressors and parameters than if training with 2006 data and predicting using 2005 weather data.

Since poor out-of-sample performance is a strong indicator of a suboptimal model, a second (and final), more robust fitness function was designed. The final form of the fitness function will be referred to as *triple cross-validation*, and is defined as a method of selecting a fixed set of variables that minimizes MAPE over six out-of-sample prediction datasets, using a total of three summers weather and load data. Using triple cross-validation, the fitness of the subset of regressors represented by each individual combinatorial solution produced by the GA is calculated using the following procedure:

1. Using MATLAB's *regress* function, create a parameterized regression equation using 2005 summer weekday data for training, then use the parameterized model to predict loads for summer weekday weather data from both 2006 and 2007, and record the MAPE for both prediction years.

2. Repeat using 2006 data for training and predict loads for the years 2005 and 2007.

3. Repeat using 2007 data for training and predict loads for the years 2005 and 2006.

4. The average of the six MAPE values for all six training-prediction year combinations is the *fitness* of the combinatorial solution represented by the individual chromosome.

Thus, triple cross-validation is performed on every individual of each generation to develop an optimal regression equation for every hour for each load zone. As implemented, the GA's only optimization goal is performance, i.e., minimization of the MAPE for the given hour. As such, it could have a tendency to overfit the training data. This often results in a model that is overly complex. An effective means of avoiding overfitting is to test a model's predictive performance on a new set of data.

The rationale behind using a fixed set of regressor variables for triple cross-validation is that while different weather may occur from one summer to another, the overall "system" should not be all that different in its response to weather. This means that for each hour, a consistent set of weather regressors should adequately represent the load response no matter what the weather used for training. Therefore, forcing the variable selection engine to minimize the MAPE for all training/prediction year combinations should yield a set of regressors that more closely represents the load response for all years, rather than weather phenomena specific to one particular summer. Furthermore, consistent results after reestimating the model form (i.e., using new data to estimate new parameters for the same model regressors) strongly support that the model is viable under broader circumstances that those contained in the original data [33]. Thus, fixing the regressors and training separately using three sets of data will enable further validation of the model via inspection of the coefficients. This will be discussed further in Chapter 6.

4.3.3.1 MATLAB GA code

The GA MATLAB routine developed for variable selection is included in Appendix H. The base GA code is a simple genetic algorithm (GA) created by Burjorjee [9],

which is a vectorized implementation of a simple genetic algorithm in MATLAB that is based on specifications described by Mitchell [31]. This code was augmented to suit its application for variable selection.

Based on preliminary runs of the fully-developed and adapted GA algorithm, the following values were selected for the final GA variable selection runs:

Initial Population – 120

Crossover Method – Uniform crossover

Crossover Rate – 0.3

Mutation Rate – 0.03

This combination of initialization parameters was found to yield an algorithm that exhibited good overall performance, exhibiting consistent increases in fitness as evolution of generations proceeded and a high degree of maximum fitness (i.e., low average MAPE across all training/prediction combinations). However, it should be noted that there may be more suitable combinations of these GA parameters.

4.4 Model Training – Estimation of Parameters

An important part of the model building process involves using data to estimate the parameters (via OLS) of each regressor to develop a full parameterized regression model. This parameterization is referred to as model training, since it in effect 'trains' the model on a particular set of data. Model training is used in the variable selection process as described in Sections 4.3.2 and 4.3.3 above. Once a subset of regressors has been selected, the full parameterized model is needed so that its predictive abilities can be tested on out-of-sample data (i.e., a dataset used for prediction).

Initial testing of the model building engines (i.e., stepwise and GA-based) was conducted by splitting an individual summer dataset into a training set and a prediction set. This revealed that using less than an entire summer's worth of data to train

the model did not yield desired performance, and that in general, training with more data yields more favorable results. The direct relation between additional data and increased predictive performance is attributable to the increased information content in larger training dataset, i.e., it embodies a more diverse set of weather phenomena than less data. This indicates that selecting an "information rich" set of training data is of great import to the model building process.

For the purposes of this research, the length of the training period is chosen to be one summer. As already indicated above, one-half of a summer of data does not appear to be enough to adequately train a model. A variety of weather typically occurs over the course of a summer, so that a full summer of data should provide relatively adequate training. Commensurately, the length of the out-of-sample dataset is also chosen to be one summer. This will enable the assessment of the predictive performance of a parameterized model over an entire summer period containing a wide range of weather, and a fair number of week long strings of consecutive days to observe the resulting time series.

In consideration of the the in-sample and out-of-sample data lengths, MERRA weather data and ISO-NE load zone data from the summer period of years 2005, 2006, and 2007 were compiled in order to provide a suite of training/prediction combinations. Consistent with the regimen of training/prediction dataset combinations employed for the triple cross-validation fitness function that was described in Section 4.3.3, model training will consist of the following six training/prediction combinations:

1. Train with 2005 summer data, predict using 2006 summer data
2. Train with 2005 summer data, predict using 2007 summer data
3. Train with 2006 summer data, predict using 2005 summer data
4. Train with 2006 summer data, predict using 2007 summer data
5. Train with 2007 summer data, predict using 2005 summer data

6. Train with 2007 summer data, predict using 2006 summer data

To ensure a high quality of load data, weekdays that were holidays or were adjacent to holidays and may have been part of a holiday weekend or week were removed from the datasets since load responses during such weekdays are generally atypical. For example, July 3-5 and September 4 were removed from the summer 2006 dataset to account for the 4^{th} of July and Labor Day weekends. Additionally, days that featured hours when active demand response resources were dispatched by ISO-NE were also removed. Active demand response resources are load-side entities that are obligated under contract to reduce their energy consumption when called upon to do so. Since these resources are typically dispatched during times when there is a shortage of generation capacity, which predominantly coincides with the greatest peak summer loads, some of the peak summer load days were removed to avoid exposing the models to artificially lower loads during training. The resulting datasets for each summer contain approximately 80 days ($n = 80$), plus or minus a couple of days depending on the year.

CHAPTER 5

MODEL VALIDATION

Model validation is a part of the model-building process that consists of testing the stability and reasonableness of the regression parameters and the overall usability and adequacy of the regression model [36]. Several 'in-sample' (i.e., using the training data only) adequacy testing techniques are available, e.g., the commonly used *adjusted coefficient of multiple determination*, R_a^2, which is a measure of how much of the variance in the response is explained by the variance in the training data. However, since ample MERRA and New England load data exist with which to test the adequacy of the model, these techniques will be forgone in favor of 'out-of-sample' methods. Out-of-sample methods test the performance of the parameterized model on new data, which is the best way to validate a model.

In order to check the adequacy of a model, it is important to understand the assumptions that are made when using MLR and OLS parameterization to build a model. In addition to the assumption that there is at least an approximate linear relationship between the regressors and response, the following assumptions are made regarding the residuals produced by a model [33]:

1. The residuals have a mean of zero.

2. The residuals are normally distributed.

3. The residuals are uncorrelated.

4. The residuals have constant variance (a condition called "homoskedasticity").

Given that the assumed qualities of its residuals are the foundation of a model's construction using MLR, model validation and adequacy checking will rely heavily on residual analysis. Typically, more than one iteration of variable selection and model adequacy checking are needed to decide upon a viable model. However, given the breadth of the modeling approach – modeling eight load zones separately, while allowing each model to freely select from the candidate regressors an hourly regression equation that may be different from that of other hours – only the first iteration will be completed as part of this research.

5.1 Analysis of Residuals

By analyzing the properties of the residuals generated by a model one can determine whether it is an adequate representation of the actual system [11]. As it represents a model's realized error, the residual is a measure of the variability of the response variable that is not explained by the model [33], and therefore, analysis of it is an effective means of measuring model adequacy. The ultimate goal of model development is a residual that exhibits characteristics of white noise – zero mean, constant variance, and normal distribution. The following three forms of residual analysis will be used to test whether a model is a viable representation of the load response:

1. Normal distribution plots to check the assumed condition of normality

2. Scatter plots of the estimated response versus residual value to check the assumed condition of constant variance, and also to check for outliers in the response estimations

3. The use of two performance metrics to assess the suite of out-of-sample predictions made be each model

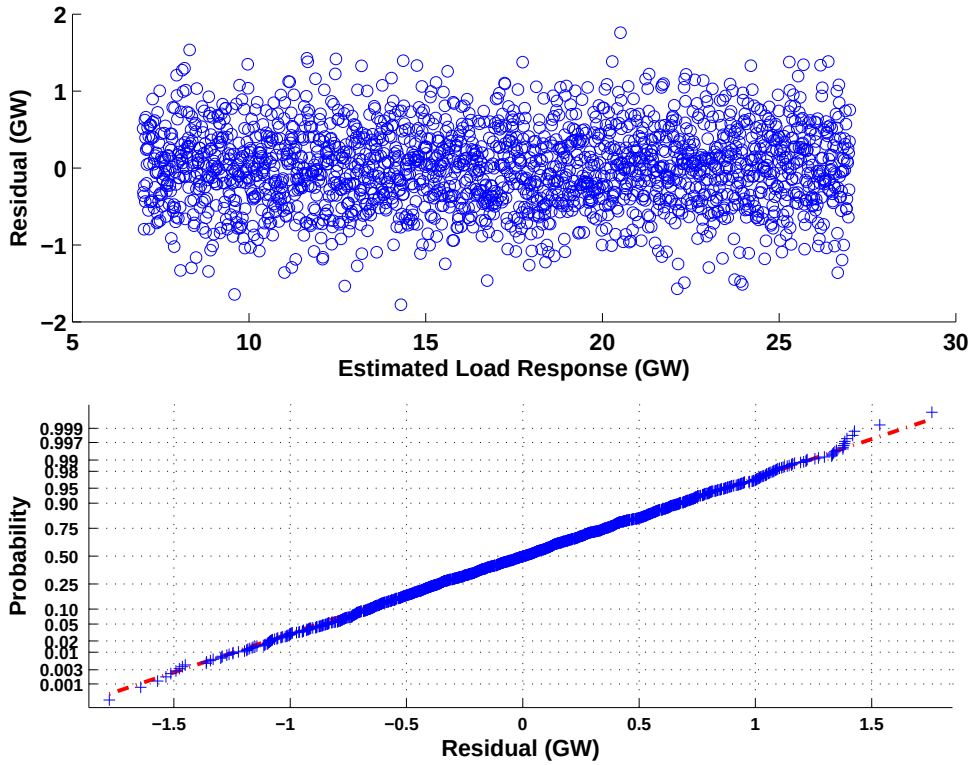

Figure 5.1. Ideal residual versus estimated load response plot (top) and normal probability plot (bottom)

(a) Mean absolute percent error (MAPE), for a measure of overall model performance

(b) MAPE during top load hours to assess model performance during peak loads. MAPE5 is defined as the MAPE during the hours with the highest 5 observed loads, and MAPE10 is defined as the MAPE during the hours with the highest 10 observed loads. MAPE5 is used to measure an individual hourly performance within the same hour throughout a summer, whereas MAPE10 is used to assess regional performance across all hours throughout a summer.

Figure 5.1 illustrates both a scatter plot of estimated load responses versus residuals (above) and a normal probability plot (below) for an ideal hypothetical model

output. These plots demonstrate that the model output satisfies the assumptions regarding the residuals outlined above, thus validating the model. The scatter plot demonstrates that the residual exhibits a consistent variance no matter the value of the estimated load response (i.e., it is homoskedastic) and is centered about a mean of zero. While there are myriad validation techniques that are available from the field of statistical regression (e.g., Mallows C_p criterion [36]), the combination of these two plots are dependable, high-level indicators of model adequacy, and will suffice for the preliminary modeling conducted as part of this research.

Performance metrics have been evaluated and tabulated for all subregional and regional load prediction results, and will be discussed in Chapter 6; however, given that the ultimate goal of this research is to model the regional load response, which is an aggregate of all eight load zone predictions, the use of the residual scatter and normal probability plots will focus mainly on the overall regional results. Emphasis on these validation techniques for the aggregate model will provide a sense of whether the overall by-hour, by-subregion modeling approach, and the use of MERRA weather data to model regional load, are reasonable. That said, since each load zone is modeled independently, these validation techniques are highly relevant to the subregional load estimation results. Applying these and other validation techniques to each subregional model's further development will be an important next step for subsequent research.

CHAPTER 6

DISCUSSION OF RESULTS

6.1 Regression Equations

Twenty-four separate hourly subsets of candidate regressors were generated by each variable selection method for each load zone, yielding a combined total of 384 subsets for both methods. Each of these regressor subsets were then parameterized using three different summer datasets, resulting in 1,152 unique regression equations! Due to this high volume, only the regression equations for the NEMA load zone have been included in this report. These equations generated by the stepwise and GA-based variable selection methods are included in Appendices B and C, respectively.

The stepwise hourly models selected using the same training year have fixed variables and parameters. In contrast, while there is significant overlap, different variables are consistently present in the same hourly models selected using different training years. It is likely that multiple subsets of regressors are virtually equivalent in their overall explanatory power, so it is difficult to glean an indication of poor model specification from these regression equations alone.

Due to the nature of the triple cross-validation utilized as part of the GA's fitness function the same variables are present in each GA-based hourly model regardless of the training year used. This enables a comparison of the parameter estimations resulting from the different training years. (Consistent parameter values across all training years would be a strong argument for a high degree of model adequacy.) Comparing the parameters reveals that some change significantly in value, and others even change signs (albeit rarely). Some of this variation is attributable to different

weather in each of the summer training periods; however, hours with instances of significant parameter deviations likely indicates poor variable selection in those models. Given that different regressors resulted for the stepwise selection for each training year, this parameter comparison cannot be performed for this method.

The presence of negative regression coefficients may also be an indication that there are problems with the regression model, due to either the presence of multicollinearity in the regressor data or because important regressors were excluded from the model [33]. It is also possible, however, that negative coefficients are warranted in the regression models due to misspecified functional forms of weather variables. If variables are misspecified (e.g., $T0^2$ should actually be $T0^{\frac{1}{2}}$), the removal of some portion of an interactive variable or a higher ordered variable would be required to create a strong fit to the data. Given what has already been discussed in Sections 4.2 and 1.5 regarding the complexity of the dynamic interaction between weather and the load response, it is difficult to determine which of these explanations is most valid, especially during this initial phase of model building.

A high-level comparison of the hourly equations resulting from the stepwise and GA methods reveals that overall, stepwise regression offers more parsimony and more consistency in variable selection from hour-to-hour than the GA. The average number of variables selected by the stepwise method for all three training years is 6.3 variables per hourly regression equation, while the GA averaged 10.3 variables per equation. A typical rule of thumb is that the number of regressor variables included in a model should not exceed the number of observations by 5-10 times [10]. This suggests that a total of 8-16 variables may be appropriate when using one summer of data (approximately 80 observations) to estimate the model. Furthermore, the hourly models selected by both variable selection methods vary in composition from hour-to-hour within each subregional model, and also between subregions. This observation

seems to validate the rationale for employing both the by-hour and the by-subregion modeling approaches.

To show an example of the hourly models resulting using each variable selection method, the following expression is the load response during Hour 15 within the NEMA load zone using the stepwise method:

$$\begin{aligned} y_{15} =& 3053.7 + 1622.6(T24)(H24)(S24) - 402.6(H0)(S24) \\ & - 94.4(T24^2)(H24^2) + 102.5(T0) + 1651.5(T0)(H0) \\ & - 37.2(H24^2) - 162.3(H0) + 214.0(H168) \end{aligned} \tag{6.1}$$

While the same hourly equation using GA-based variable selection is:

$$\begin{aligned} y_{15} =& 2673.5 + 151.1(Tm3) - 393.3(T24) - 1043.3(Tm1) \\ & + 916.5(T0^2) + 1653.6(T3)(H24) + 620.9(T48) \\ & + 268.6(Hm24)(Sm3) - 236.1(Hm3) - 17.8(H0)(S0) \\ & + 467.9(Hm24) + 230.4(Tm1)(Sm3) + 826.6(delHm24) \\ & - 1.9(delTm24)(delHm24) \end{aligned} \tag{6.2}$$

Comparing Equations 6.1 and 6.2 clearly shows more parsimony in the stepwise selection (8 regressors in contrast to 13).

In theory, the y-intercept term (parameter β_0) should represent the amount of load present when all weather variables are at their summer minimum values (i.e., when their scaled values approach zero) and as such may be viewed as an approximation of the amount of load that is a not sensitive to weather. This value should change from hour to hour to reveal a diurnal pattern of approximated weather-insensitive

load. To test this theory, all β_0 values were plotted over all hours for each variable selection method and training year. Figure 6.1 is the resulting plot for the CT load zone, which also includes the diurnal load shapes corresponding to the 2005 summer days exhibiting the maximum (June 14) and minimum (September 30) average load. Similar curves for the seven other load zones are included in Appendix G. Assuming that the minimum load day is an indication of the weather insensitive load (at least in terms of the summer data provided to the model), inspection of these plots reveals what could be described as a reasonable approximation of the weather-insensitive load. The values of β_0 trace the minimum load day well for most hours, although there are some deviations. Overall, it appears that the regression did an adequate job of separating out the weather-sensitive portion of the load response. (The plots for the other load zones exhibited similar trends). This gives an initial sense of the weather insensitive load profile, which could be useful for the further development of the model if other types of regression are attempted.

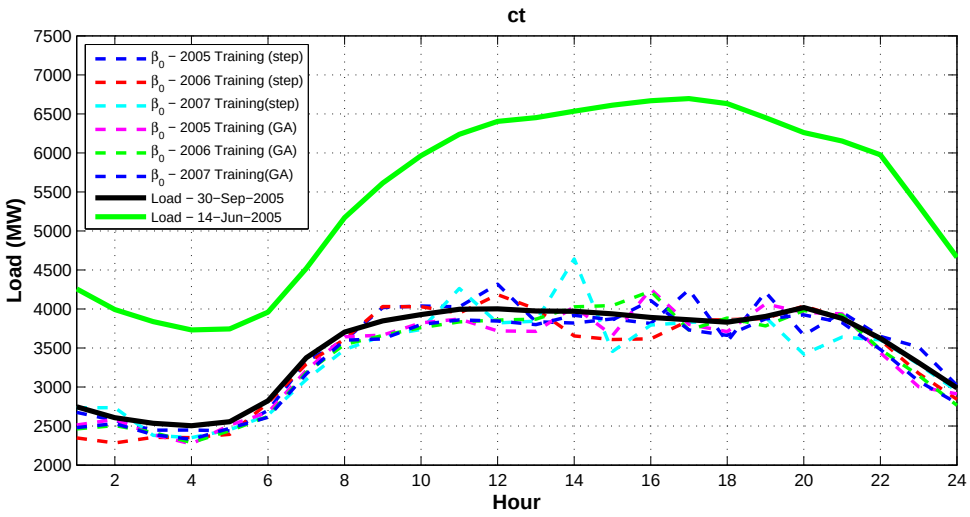

Figure 6.1. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the CT load zone.

6.2 Residual Analysis

The residual scatter plot and normal probability plot for all training/prediction combinations for the stepwise model are illustrated in Figures 6.2 and 6.3. These same plots for the GA-based model are illustrated in Figures 6.4 and 6.5. All of the scatter plots for both methods depict inconsistent variance over the range of estimated load responses, as well as a tendency for straying from zero mean. Overall, the normal probability plots depict a light-tailed distribution, which indicates a slight departure from normal distribution. When using actual data, departure from the ideal depicted in Figure 5.1 in Section 5.1 is expected, although the extent of departure that is acceptable is somewhat unclear. Based on these residual plots, it appears that the models selected by both methods are mediocre in their adherence to their presumed characteristics.

6.2.1 Model Performance

6.2.1.1 Regional Performance

Table 6.1 lists the regional MAPE and MAPE10 averaged over the full suite of training/prediction combinations. Figure 6.6 is a bar graph illustrating the same results. Detailed hourly tabulations of the performance results are included in Appendix D. Regional load prediction performance for both the stepwise and GA-based variable selection methods resulted in a MAPE of less than 3%, which indicates good overall performance. Slightly weaker predictive performance is noted during the top 10 load hours of the summer period, with average MAPE10 over all training/prediction combinations ranging from 1.54%–6.03% for the the stepwise model, and 0.83%–4.05% for the GA model. Overall, the GA-based model performed somewhat better over all training/prediction year combinations, with an average MAPE of 2.19%, as compared to the stepwise model's average MAPE of 2.83%.

Table 6.1. Mean absolute percent error (MAPE) and mean absolute percent error over top 10 load hours (MAPE10) – Stepwise (Step) and genetic algorithm (GA) variable selection Methods – All years

	Step - MAE	**Step - MAE10**	**GA - MAE**	**GA - MAE10**
Train - 05, Pred - 06	2.62	4.73	2.05	2.36
Train - 05, Pred - 07	2.93	1.54	2.15	0.83
Train - 05, Pred - 05	2.98	6.03	2.17	4.05
Train - 06, Pred - 07	2.70	1.93	2.11	2.02
Train - 07, Pred - 05	2.93	2.67	2.44	2.83
Train - 05, Pred - 06	2.80	2.15	2.22	2.35
Averages	**2.83**	**3.18**	**2.19**	**2.41**

Figure 6.7 is the average hourly MAPE and MAPE5 for the entire region, which again, is the aggregate of the values for all the load zones. Values for MAPE are shown to be consistent across all hours, with hourly MAPEs between 2% and 4% for most hours over the diurnal cycle. MAPE5 values are generally greater, especially between the hours of one and seven. As can be seen, the GA model outperformed the stepwise method for almost all hours in terms of overall MAPE and also MAPE5.

Figures 6.8 through 6.13 are time series plots of the regional predicted and actual loads corresponding to a summer peak load week. Figure 6.8 and 6.9 are timeseries plots of actual versus predicted load for the week of July 18-23, 2005, using training years of 2006 and 2007, respectively. Figures 6.10 and 6.11 are timeseries plots of actual versus predicted load for the week of July 17-22, 2006, using training years of 2005 and 2007, respectively. Figures 6.12 and 6.13 are timeseries plots of actual versus predicted load for the week of July 30 to August 4, 2007, using training years of 2005 and 2006, respectively.

The time series plots show a fairly accurate load response estimation for both methods. In general, the prediction traces follow the actual values well, but less so during the overnight hours and during the peaks of a couple days shown.

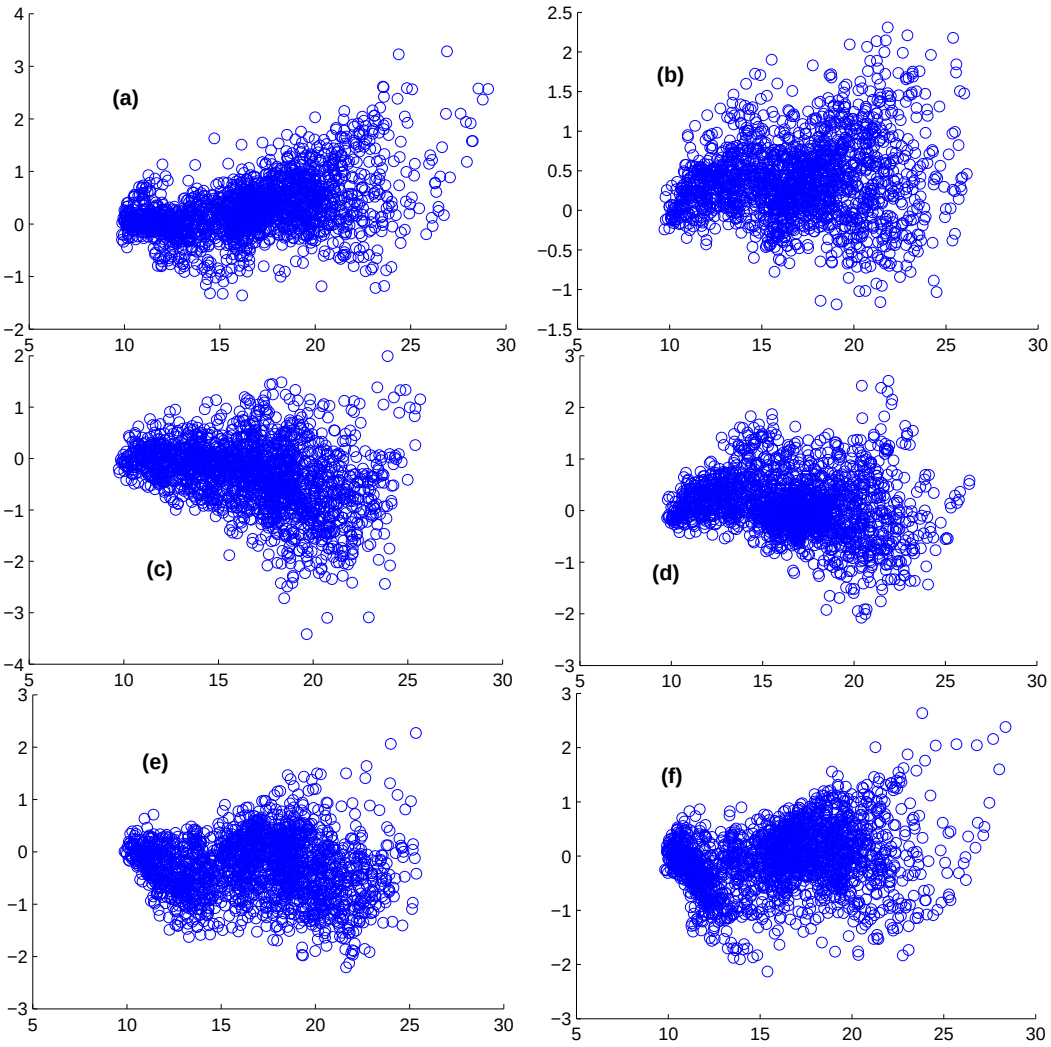

Figure 6.2. Stepwise Variable Selection - Residual plots for the following training/prediction pairs: (a) 2005/2006 (b) 2005/2007 (c) 2006/2005 (d) 2006/2007 (e) 2007/2005 (f) 2007/2006

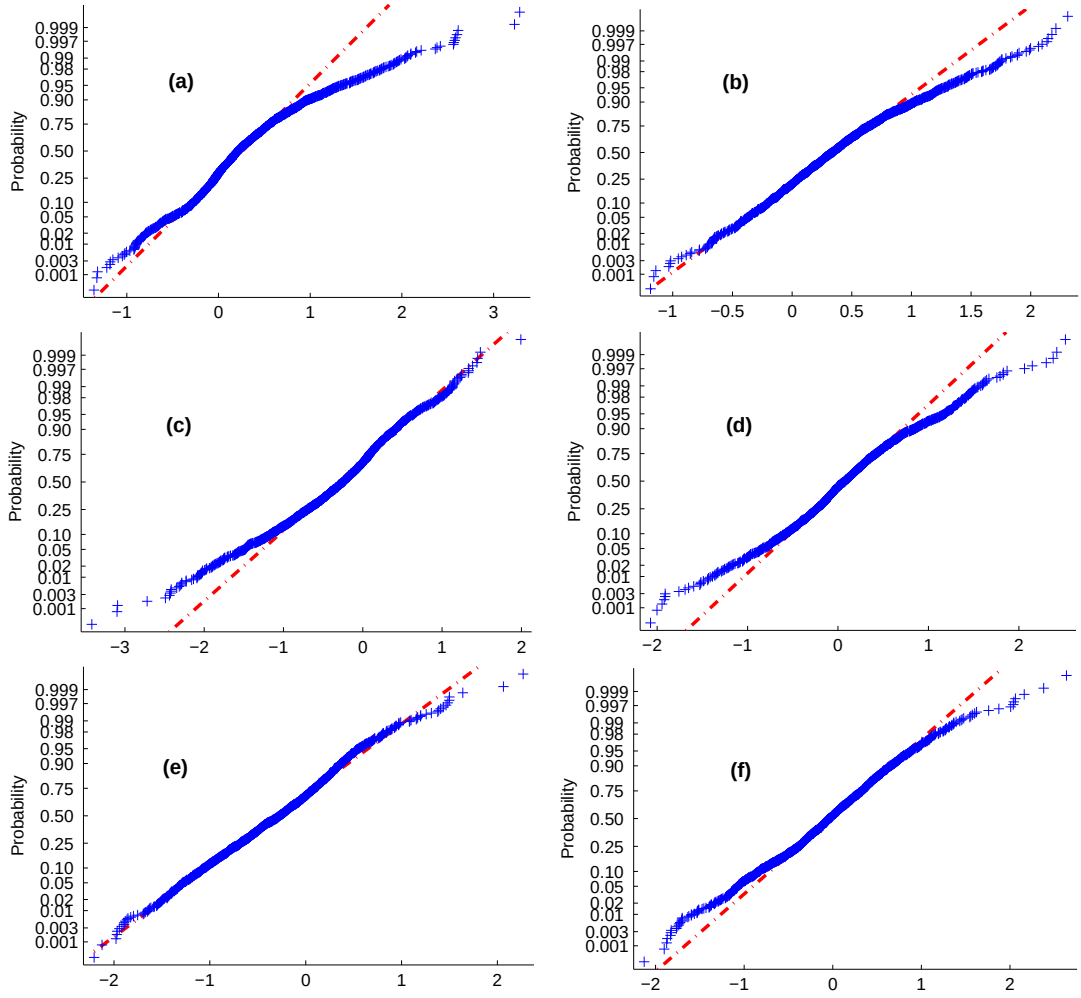

Figure 6.3. Stepwise Variable Selection - Normal distribution plots for the following training/prediction pairs: (a) 2005/2006 (b) 2005/2007 (c) 2006/2005 (d) 2006/2007 (e) 2007/2005 (f) 2007/2006

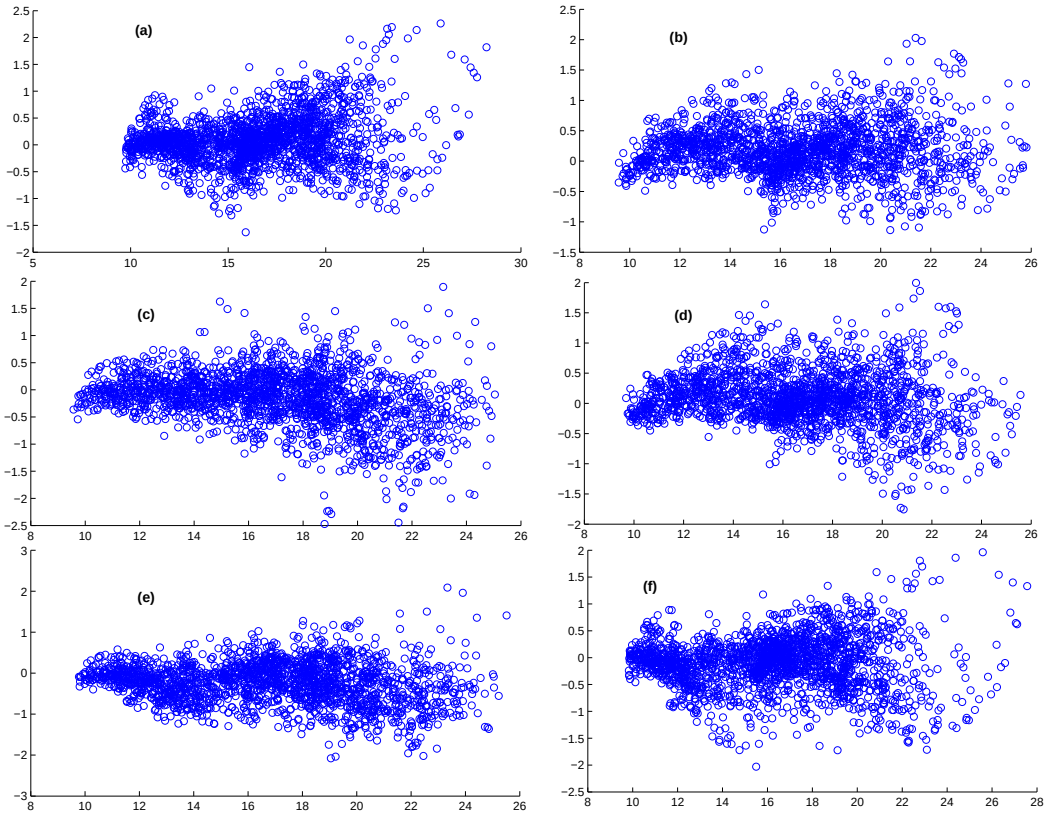

Figure 6.4. GA Variable Selection - Residual plots for the following training/prediction pairs: (a) 2005/2006 (b) 2005/2007 (c) 2006/2005 (d) 2006/2007 (e) 2007/2005 (f) 2007/2006

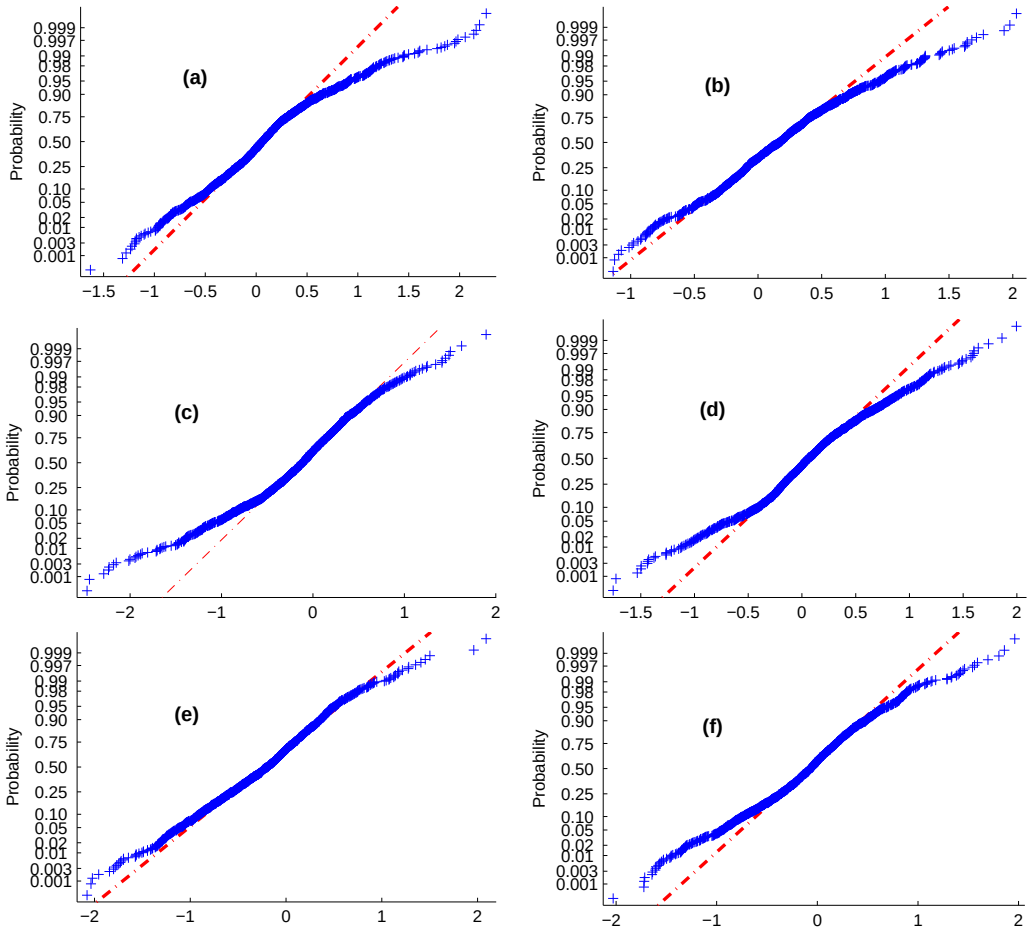

Figure 6.5. GA Variable Selection - Normal distribution plots for the following training/prediction pairs: (a) 2005/2006 (b) 2005/2007 (c) 2006/2005 (d) 2006/2007 (e) 2007/2005 (f) 2007/2006

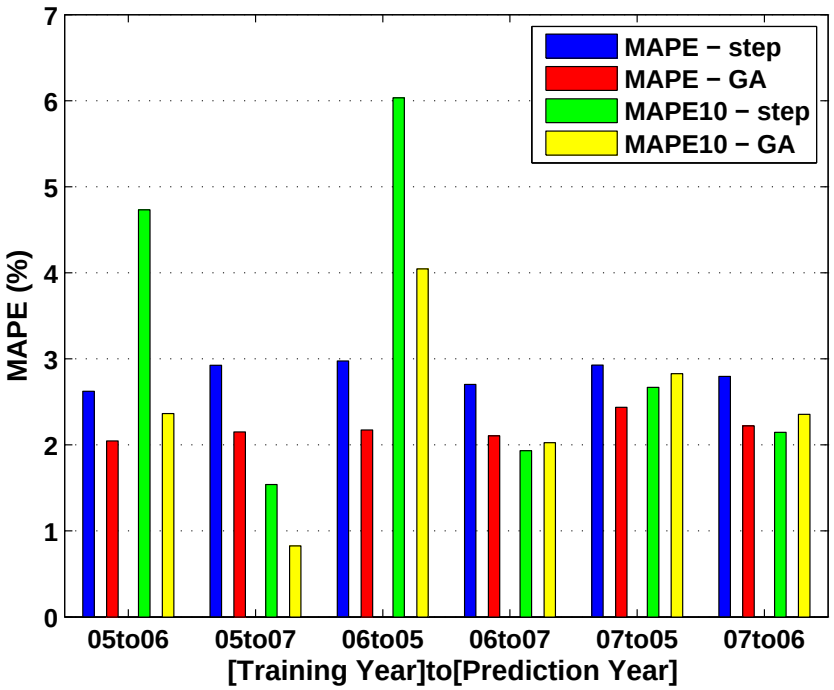

Figure 6.6. Regional MAPE and MAPE10 for all training/prediction combinations

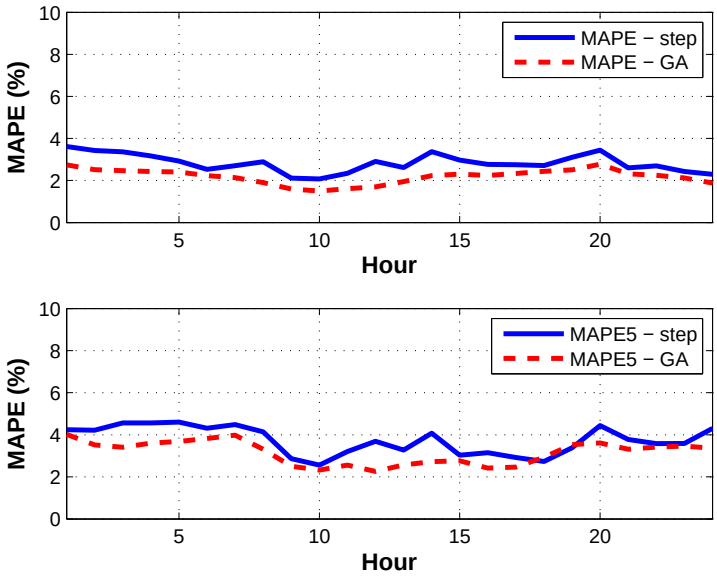

Figure 6.7. Regional - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

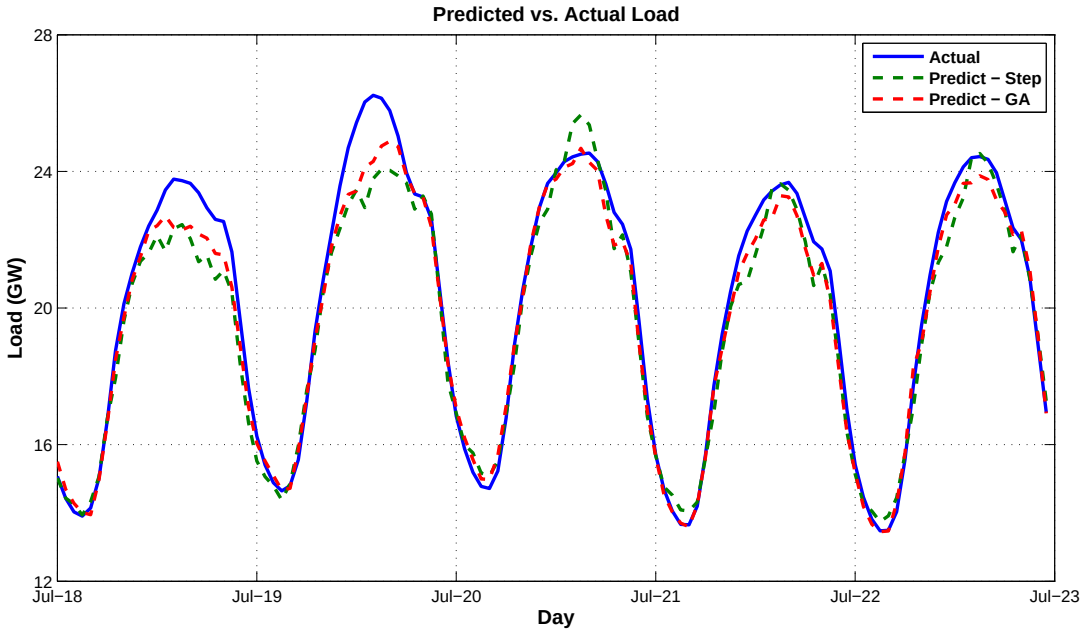

Figure 6.8. Regional actual vs. predicted loads - July 2005; training year - 2006

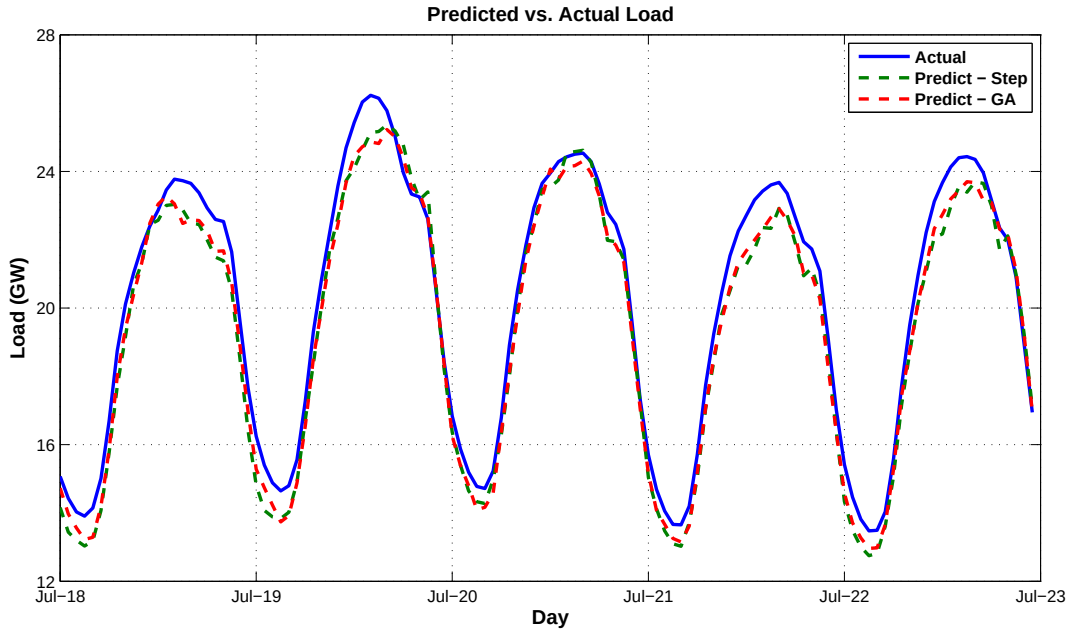

Figure 6.9. Regional actual vs. predicted loads - July 2005; training year - 2007

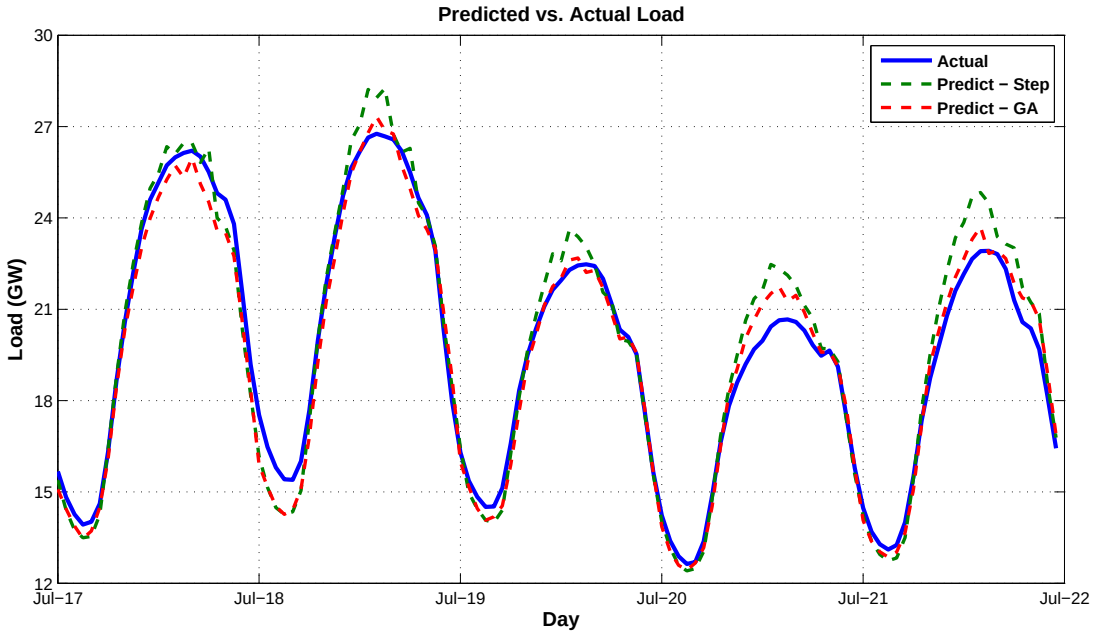

Figure 6.10. Regional actual vs. predicted loads - July 2006; training year - 2005

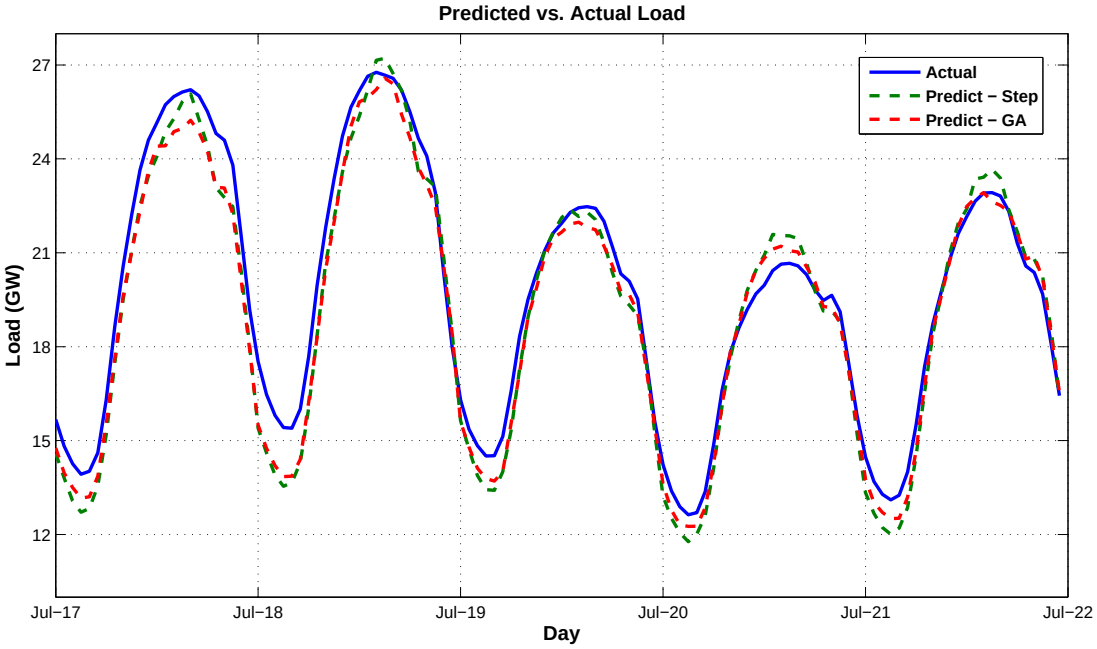

Figure 6.11. Regional actual vs. predicted loads - July 2006; training year - 2007

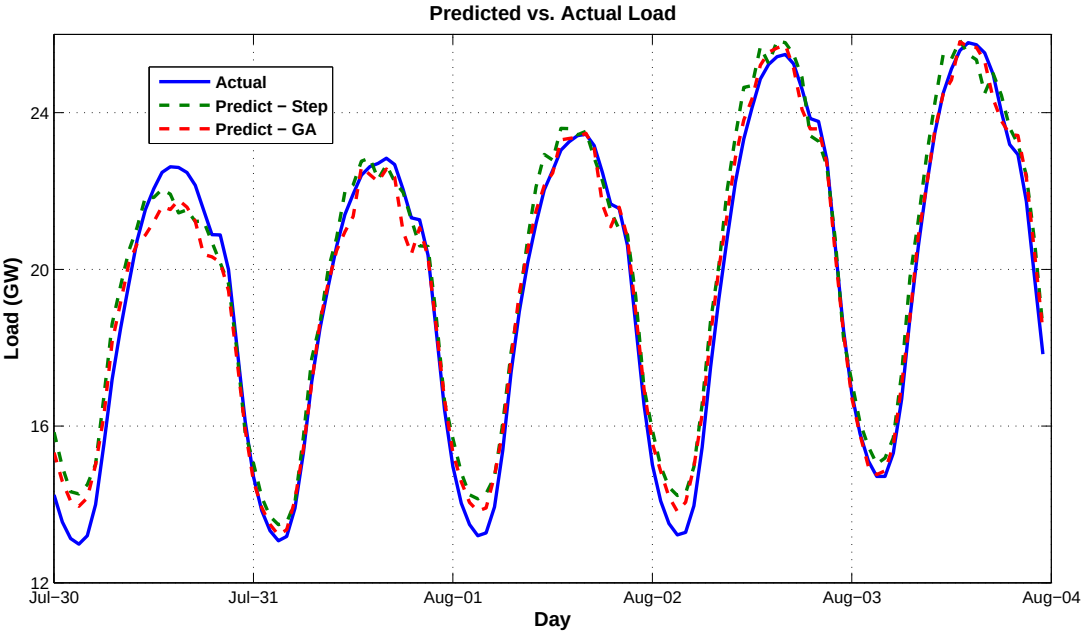

Figure 6.12. Regional actual vs. predicted loads - August 2007; training year - 2005

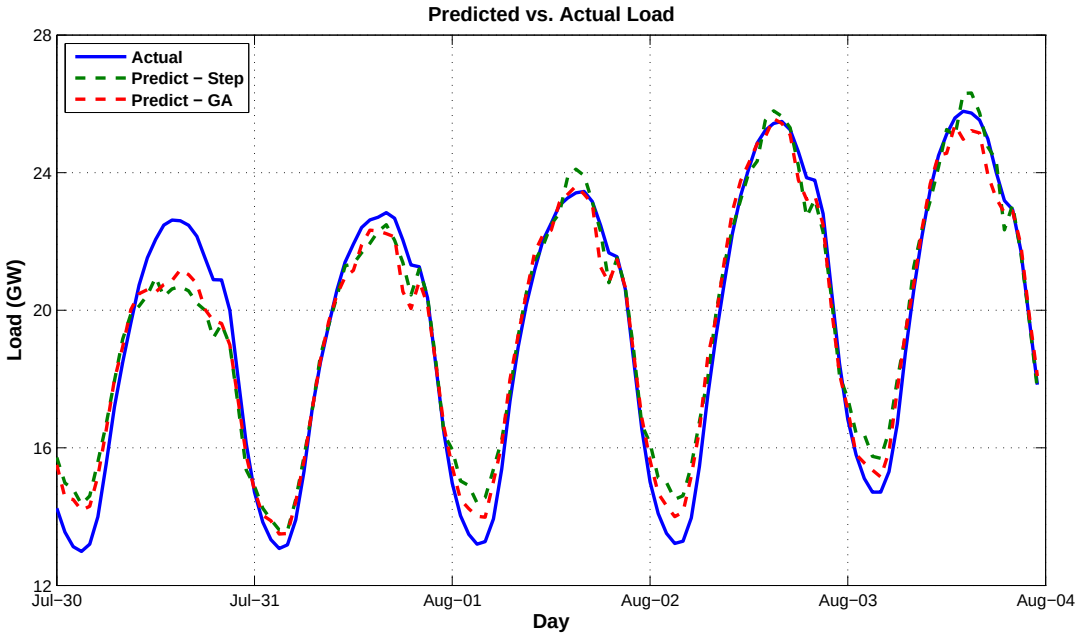

Figure 6.13. Regional actual vs. predicted loads - August 2007; training year - 2006

6.2.1.2 Subregional Performance

Figure 6.14 is a bar chart illustrating the MAPE and MAPE5 corresponding to each model for all eight ISO-NE load zones. The models for the CT load zone, comprising the largest portion of New England's regional load of all load zones (approximately 25%), exhibit the poorest overall performance for both variable selection methods. The source of this poor performance is unknown, but could be attributed to missing regressors or a poor selection of MERRA gridpoint locations to represent the CT load zone.

Figures 6.15 through 6.22 are plots of the hourly MAPE and MAPE5 for all of the New England load zones. These values are the averages over the full suite of training/prediction combinations. For a detailed tabulation of the MAPE and MAPE5 values associated with each hourly model of every training/prediction pair, refer to Appendices E and F.

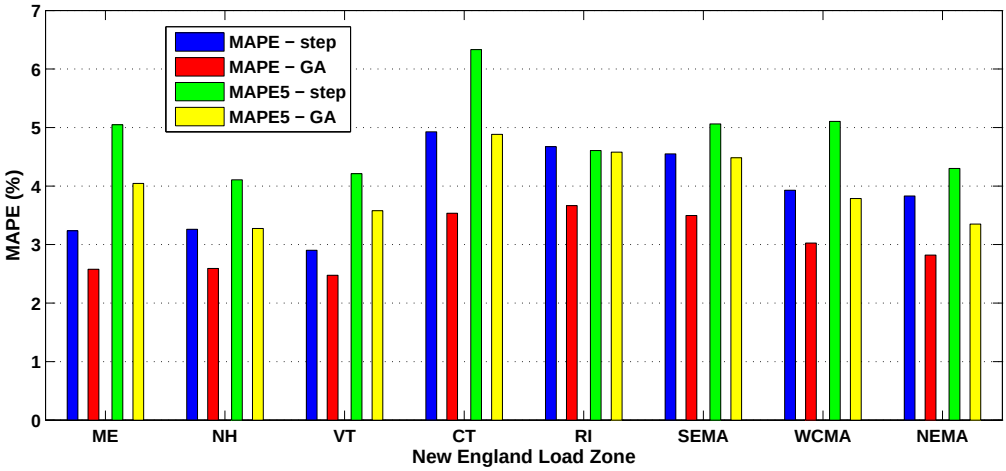

Figure 6.14. MAPE and MAPE5 for each load zone - average values over all training/prediction combinations

In general, each model's performance varied (sometimes significantly) depending on the summer dataset used for training, suggesting that careful consideration must be given to the set of data used to parameterize a MLR model.

Overall, the subregional predictive performance is poorer, with average MAPE values across all training/prediction year combinations ranging from approximately 2.5%–6% among the eight load zones. This suggests that increased regional predictive performance likely resulted from uncorrelated biases in the subregional model outputs. Another possible explanation for the higher regional performance would be if there were "leakage" in the load zone load data, i.e., if load from one load zone was counted within more than one different load zone at different times. This leakage could result in complementary biases in adjacent load zones; however, since the load zone data are the result of ISO-NE's market settlements, which are heavily vetted, the probability of the leakage described is highly unlikely.

The poor prediction performance noted for most subregional models during the early morning hours (hours 1-6) may be a result of the variable scaling method em-

ployed. Weather variables (e.g., temperature) are typically at their lowest values during these hours, meaning that their scaled values approach or are equal to zero. The presence of regressors within the $\mathbf{X}$ matrix in Equation 4.3 that approach zero may be an indication that potential problems (e.g., rank deficiencies) may arise while MATLAB's QR factorization algorithm is attempting to estimate parameters using OLS. Many rank deficiencies were noted while the GA algorithm was calling MATLAB's *regress* function for these hours, and are therefore believed to be a result of the scaling method. When this occurs, the $\mathbf{X}$ matrix has more columns than rows, the square matrix $\mathbf{X'X}$ is singular, and $(\mathbf{X'X})^{-\mathbf{1}}$ does not exist, which means that Equation 4.5 completely degenerates. The GA-code was written so that when serious rank deficiencies that resulted in a non-number (NaN) were output from the fitness function, combinatorial solution was ascribed a very large MAPE value so that the code could continue to operate. Although this 'fix' worked, it also means that many combinatorial solutions tested during these hours were discarded even though they were as valid as those that generated numerical solutions, which significantly reduced the performance of the GA-based variable selection.

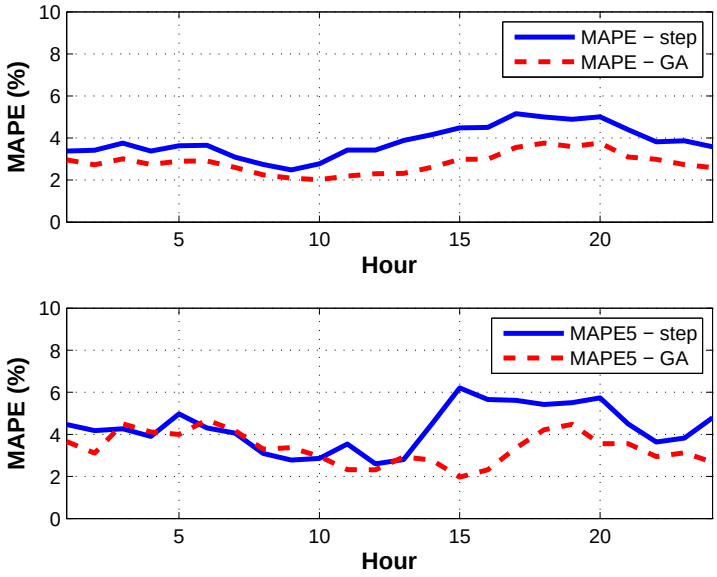

Figure 6.15. NEMA Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

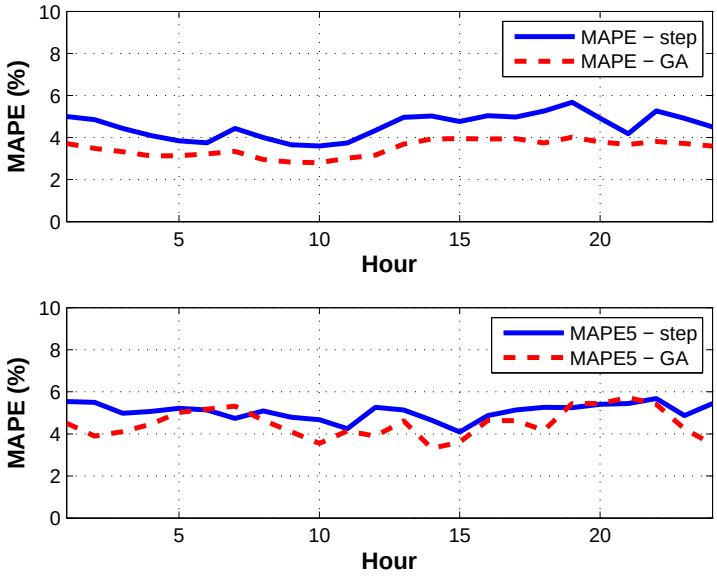

Figure 6.16. SEMA Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

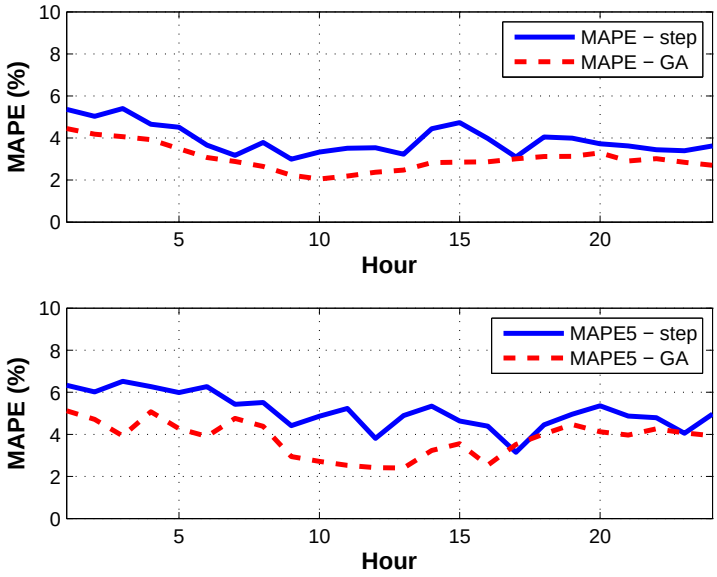

Figure 6.17. WCMA Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

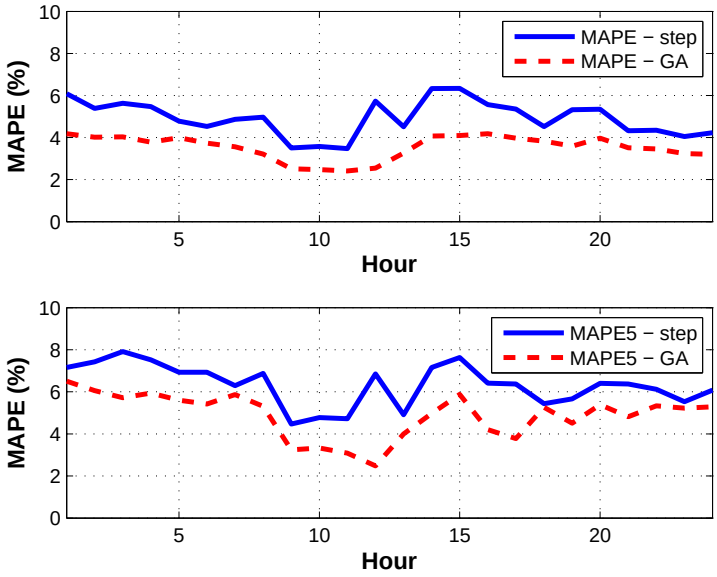

Figure 6.18. CT Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

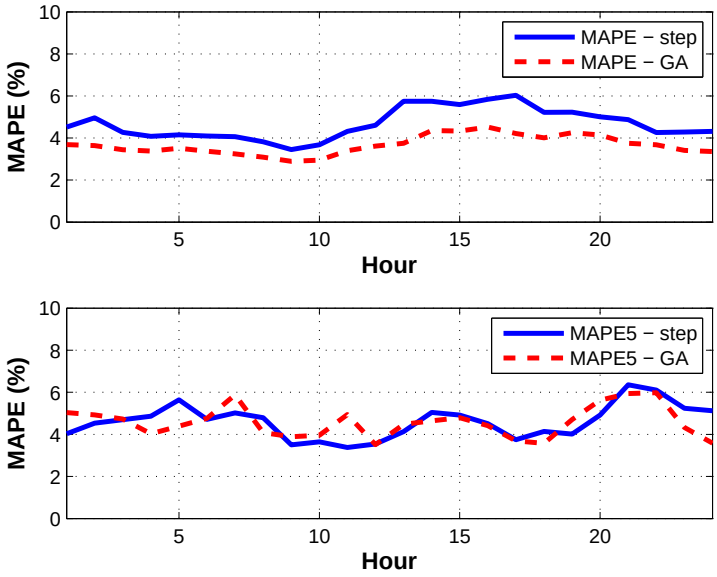

Figure 6.19. RI Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

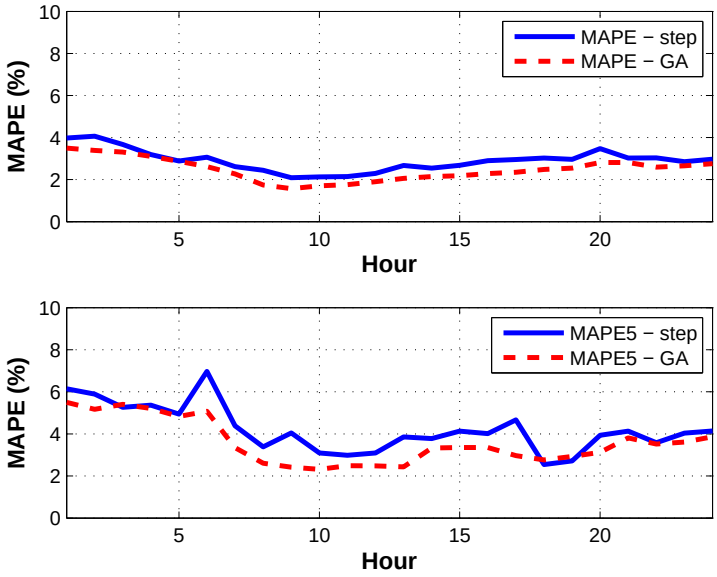

Figure 6.20. VT Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

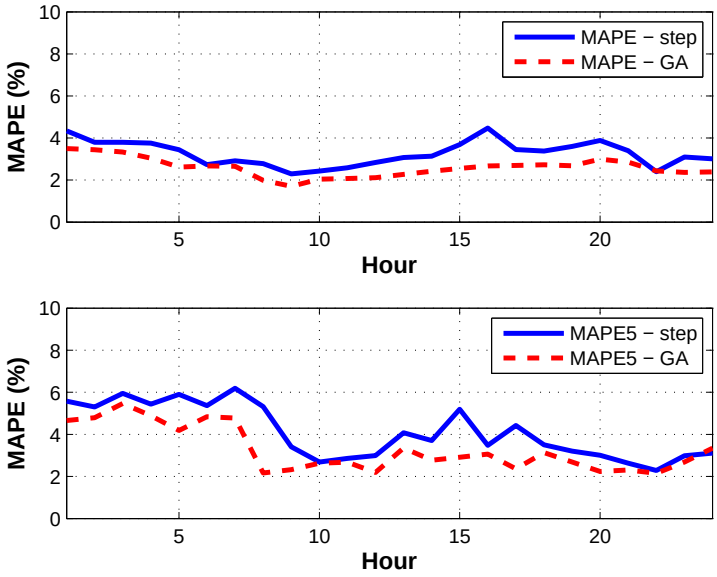

Figure 6.21. NH Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

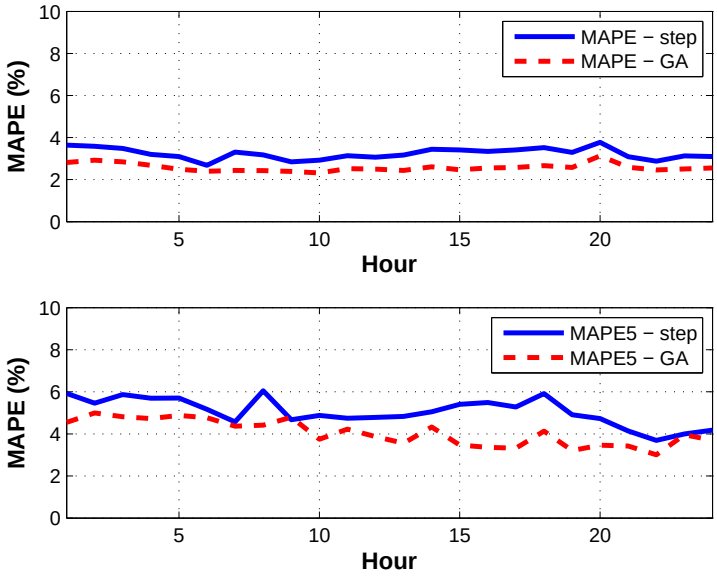

Figure 6.22. ME Load Zone - Hourly MAPE (top plot) and MAPE5 (bottom plot) for Stepwise and GA variable selection methods. Values are averages for all training/prediction combinations

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

Based on the results of this research, the use of weather data provided by NASA's MERRA reanalysis dataset to hindcast regional load has demonstrated merit. Thus far, three atmospheric fields from the MERRA have been integrated into the model building engines; however, many more potential weather regressors are available. Including other fields containing weather variables (e.g., wind speeds and precipitation) may provide the summer load response model additional fidelity. Furthermore, there may be other combinations of MERRA gridpoints that more suitably represent the weather influencing the load response in each subregion. Other than the set of MERRA gridpoints used and discussed in this report, no others have been tested. Increased model performance may result from improved gridpoint selection. It would be useful to validate the MERRA weather data used for this research, perhaps with data collected and maintained by the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC). Due to the inherent limitations of the operational model(s) used to generate the MERRA dataset, there may be data quality issues associated with the atmospheric fields used in the hindcasting model. Also, because this is a relatively new dataset (released in 2008), the first body of journal publications concerning MERRA is just beginning to emerge. Reviewing this literature as it becomes published could prove helpful in further validation of the MERRA dataset with respect to load hindcasting.

The hourly models selected by both the stepwise regression and GA-based variable selection vary in composition from hour-to-hour within each subregional model, and

also between subregions. This observation seems to validate the rationale for employing both the by-hour and by-subregion modeling approach. However, this does not entirely rule out a potential need for introducing other candidate regressors that are either based on a different functional form of weather already used (e.g., $T0^{1.5}$), new weather variables (e.g., wind speed), and/or the integration of both. Similarly, some of the candidate lagged weather regressors explored so far may be redundant. When included together in a regression model, weather predictors such as 'Sm3' and 'S3' may not embody much predictive capability that is substantially unique from one another. If this is true, one of these regressors should be eliminated, thus simplifying the variable selection problem and increasing the likelihood of finding a more optimal combinatorial solution.

With respect to using the hindcasting tool for the determination of wind's capacity value, further research is need to perform long-term wind resource characterization using the MERRA dataset, as has already been proposed [17]. As the hindcasting tool is further developed and validated, complementary long-term wind power and load datasets could be synthesized, and capacity value calculations could be made.

The aggregate regional load model was found to perform well, especially considering that this research constitutes the first attempt at modeling. These performance results seem to indicate that the overall approach (by-hour, by-subregion) holds promise for the hindcasting methodology. Preliminary regional load prediction performance for both the stepwise and GA-based variable selection methods resulted in a MAPE of less than 3%, which indicates strong overall performance. Slightly weaker predictive performance is noted during the top 10 load hours of the summer period, with average MAPE10 over all training/prediction combinations ranging from 1.54%–6.03% for the the stepwise model, and 0.83%–4.05% for the GA model. Since higher performance during times of peak load are critical to accurate (CV) calculations, additional work is required to improve estimations of the peak load response.

The subregional predictive performance is poorer, with average MAPE values across all training/prediction year combinations ranging from approximately 2.5%–6% among the eight load zones. This suggests that increased regional predictive performance likely resulted from uncorrelated biases in the subregional model outputs. Another possible explanation for the higher regional performance would be if there were "leakage" in the load zone load data, i.e., if load from one load zone was counted within more than one different load zone at different times. This leakage could result in complementary biases in adjacent load zones; however, since the load zone data are the result of ISO-NE's market settlements, which are heavily vetted, the leakage described is highly unlikely.

Poor prediction performance was noted for the subregional models during early morning hours (hours 1-6) and is believed to be a result of the variable scaling method employed. The nature of the problem involved rank deficiencies in the $\mathbf{X}$ regressor matrix, which were observed frequently when the GA algorithm was calling MATLAB's *regress* function for these hours. Although a temporary fix was added to the GA-code to make it function in the presence of rank deficiencies, the result was that a number of potentially viable solutions were discarded. To rectify this problem, another variable scaling method should be employed. Two potential scaling methods are unit normal scaling and unit length scaling [33].

In general, each model's performance varied (sometimes significantly) depending on the summer dataset used for training, suggesting that careful consideration must be given to the set of data used to parameterize a MLR model. The GA-based model performed somewhat better over all training/prediction year combinations, with an average MAPE of 2.19%, as compared to the stepwise model's average MAPE of 2.83%. As previously stated, the most effective method of validating a regression model is testing its predictive performance on new data. Although all validation employed out-of-sample methods, the GA model's out-of-sample predictive performance

was based on three summers that were part of the fitness function used to develop the model. Thus, although these predictions are technically out-of-sample because model parameterization was performed with different data than load prediction, the resulting out-of-sample performance may be unique to the three years involved in the triple cross-validation. Therefore, further validation of this model via additional out-of-sample performance evaluations are needed to verify that the GA models are in fact better.

Additionally, it is worth considering using more than one summer dataset to train the models to ensure a rich variety of weather phenomena are included. Perhaps a better option would be an amalgam of hand-selected data from multiple summers, so that there is more control over the weather that is introduced to the model during parameterization. Whatever the method, a more strategically-selected training dataset would only serve to improve the model building and its overall performance.

Although the two variable selection engines seemed to perform reasonably in their maiden runs, there is much to improve. With respect to the GA-based method, other means of mitigating the its tendency to overfit could be developed, e.g. introducing parsimony pressure to balance performance with model simplicity and efficiency. With respect to the stepwise method, specifying different initial models to MATLAB's *stepwisefit* function could be tested to see if better fits will emerge. An interesting topic of future work would be to integrate the two variable selection methods, e.g., feeding the GA output into the stepwise fit function as an initial model. Also, it is possible that a better variable selection method exists. As such, an algorithm could be developed to implement a third method that may be better-suited to the highly discrete nature of the combinatorial solution sought. A potential candidate is simulated annealing, which has already been proposed for variable selection [22].

Some of the validation techniques used in this research were done so exclusively on the regional load response predictions and not the subregional predictions, namely the

residual scatter plot and normal probability plot. Since each load zone is modeled independently and their predictions are aggregated to arrive at the regional predictions, application of these techniques are highly relevant to the subregional load estimation results. Applying these and other validation techniques to the further development of each subregional model will be an important next step for regional hindcasting performance.

Overall, a vast amount of information was generated during the model building performed as part of this research – a total of 1,152 regression equations, along with performance results (including time series outputs) corresponding to each of them for all training/prediction year combinations – and time did not permit the culling of all of it for this report. Improving each subregional model may be possible simply by integrating these results more than time allowed thus far. Additionally, a solid foundation of MATLAB routines has been developed to perform everything from downloads of MERRA data, organization of a cohesive body of load and weather data, variable selection and model building, and the tabulation and plotting of results to enable the user to assess the performance of the model's predictive capability and overall adequacy. Further development of the hindcast model is warranted and additional research opportunities are available at every phase of the model building procedure.

Lastly, after the summer weekday model has been fully developed and validated, the modeling techniques covered in this report could be applied to modeling the load response during weekend days and other seasons. Then there would be a cohesive model that could perform retrospective load predictions for an entire year.

APPENDIX A

HOURLY WEATHER-LOAD PLOTS

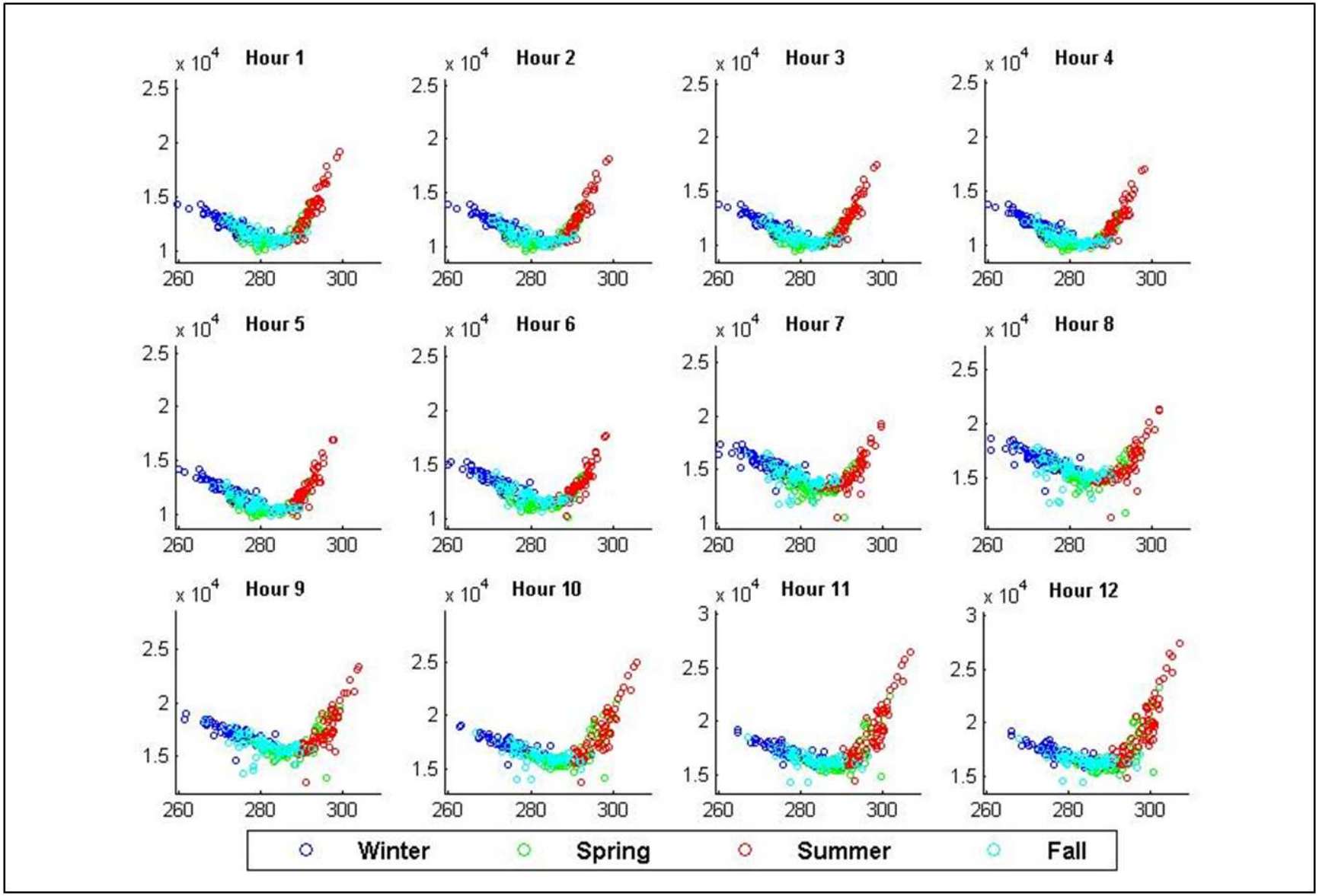

Figure A.1. Scatterplots of temperature vs. load, binned by hour - hours 1-12, summer 2006, seasons by color.

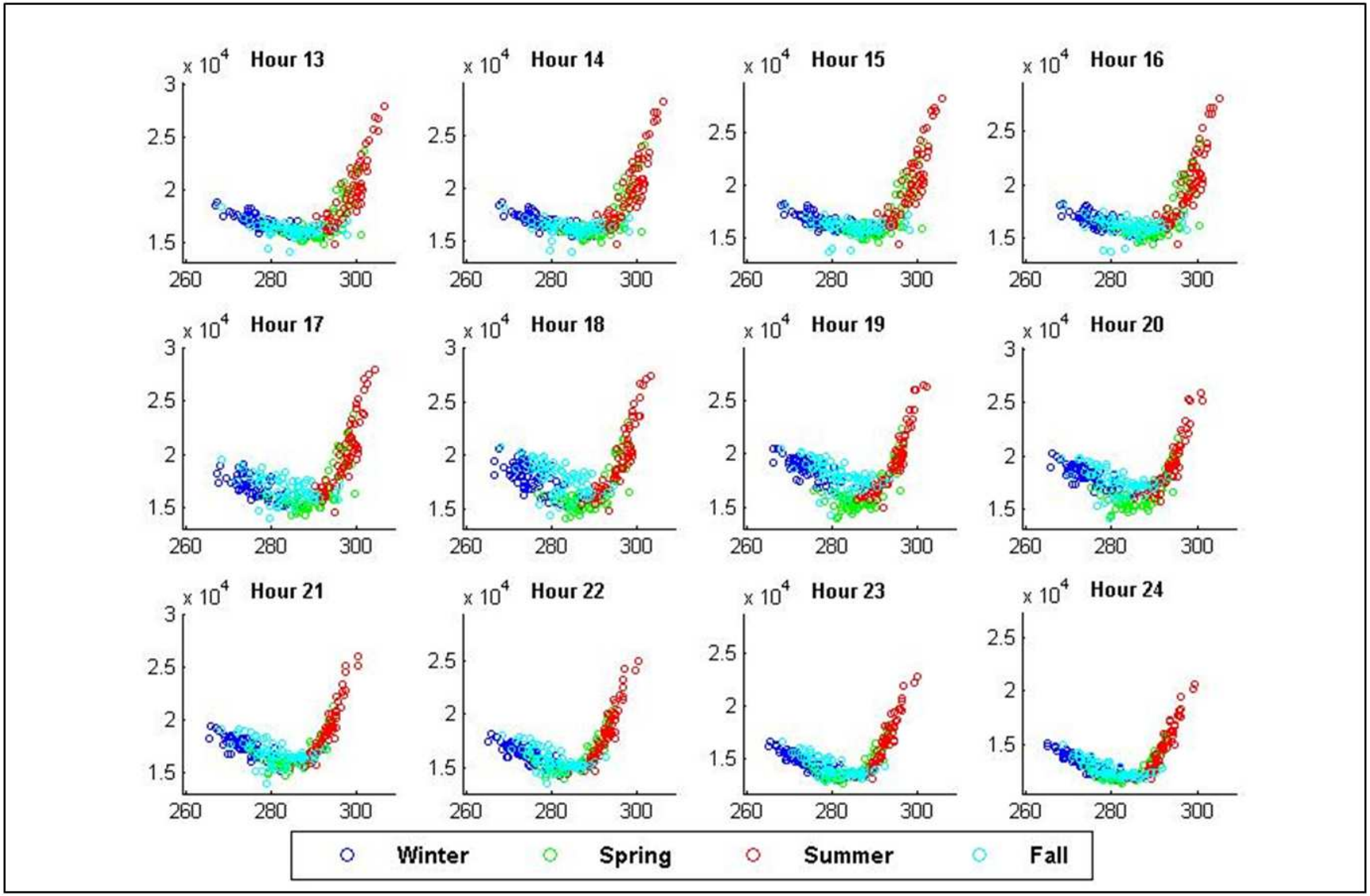

Figure A.2. Scatterplots of temperature vs. load, binned by hour - hours 13-24, summer 2006, seasons by color.

APPENDIX B

REGRESSION EQUATIONS FOR STEPWISE VARIABLE SELECTION METHOD - NEMA LOAD ZONE

Table B.1. Hourly regression equations; Load zone - NEMA; Variable selection method - step; Training Year - 2005; Prediction Years - 2006 and 2007

Hour	Regression Equation
1	2097.8 +621.7(T24)(H24)(S24) +366.3($H0^3$) +183.4(H72) +704.4(T24)(H3) -18.3(H3) +305.2(Tm1)(Hm24) -254.2(Hm24) +85.0(S72) +32.2(S24)
2	1956.5 +515.3(T24)(H24)(S24) +390.6($H0^3$) +120.9(H72) +883.1(T24)(H3) -154.6(H3) +29.4(Tm24) +66.4(S72) +57.4(S24)
3	1790.5 +367.3(T24)(H24)(S24) +277.6($T24^2$) +317.4($H0^2$) +173.2($T0^3$) +330.5(T0)(H24) +230.4(H72) +497.3(T24)(H3) -601.6(H3) +112.2(S72) +194.3(delHm24)
4	1979.3 +1312.6(T24)(H24)(S24) -615.3(H24)(S24) +483.1($H0^3$) +224.6(H3)(S24) +63.1($T24^2$)($H24^2$) +68.3(H72) +364.8(Tm24)(Hm3)
5	2011.7 +1094.9(T24)(H24)(S24) -507.0(H24)(S24) +395.7($H0^3$) +227.8(H3)(S24) +122.7($T24^2$)($H24^2$) +497.1(Tm24)(Hm3)
6	2064.7 +423.0(T24)(H24)(S24) +308.9($T24^2$) +384.1(Tm1) +903.0($H3^2$) +130.5(H72) -614.1(Hm3) -41.4($S24^2$)
7	2523.1 +822.1($H0^2$) +752.8($T24^2$)($H24^2$) -244.5($H24^2$)
8	2854.0 +868.0($H0^2$) +851.1($T24^2$)($H24^2$) -401.6($H24^2$) +268.7(Hm24)(Sm3) -54.2($S3^2$)
9	2997.8 +872.6(T24)(H24)(S24) -640.2(H24)(S24) +634.0($H0^3$) +636.7(H3)(S24) +471.0($T24^2$)($H24^2$) +53.6(Hm24)(Sm1)
10	3032.1 +151.2(T24)(H24)(S24) +373.0(H24)(S24) +225.1($H0^3$) +191.0($T24^2$)($H24^2$) +662.0(T0)(H3) +392.0(T24)(H3)
11	3085.7 +295.3(H3)(S24) +488.8($T24^2$)($H24^2$) +744.3($H3^2$) +281.6(T48) -67.1(Hm3) +990.9(T0)(S24) -417.0(S24)
12	3043.7 +471.5(H0)(S24) +548.5($T24^2$)($H24^2$) +405.0(T72) +1434.0(Tm1)(Hm3) -283.4(Hm3)
13	3115.6 +434.2(T24)(H24)(S24) +333.7($T24^2$) +183.5($H3^2$) +342.7(H48) +1838.3(T0)(H3) -495.3(H3)
14	2970.6 +515.3(H0)(S3) +481.6($T24^2$) +824.6($T0^3$) +695.3($H3^2$) +352.6(H48)
15	2580.0 +452.9($H0^3$) +1023.4(H3)(S24) +971.7(T48) +1139.3(Tm3)(Hm24) -391.3(Hm24) -144.0($S24^2$) +491.1(delTm24)
16	3132.1 +1161.8($T3^2$) +932.5($H0^3$) +1107.5(H0)(S24) +364.6(T72) -186.5($S24^2$) -482.4(delHm24)
17	2760.8 +1655.6(T24)(H24)(S24) -622.5(H24)(S24) +358.0($H0^3$) -142.5($T24^2$) +649.9(Tm1) +426.5(T48) +1004.8(Tm1)(Hm3)
18	3184.8 +3555.1(T3)(H0) +594.5(H48) -1517.9(H0)
19	3173.4 -554.5($T24^2$)($H24^2$) +3589.0(T3)(H0) +1069.5($H24^2$) -1659.3(H0)
20	3040.6 +1282.4($T0^3$) +240.8($T24^2$)($H24^2$) +128.5($H3^2$) +438.6($H24^2$)
21	3000.7 +1029.3($T0^3$) +205.3($T24^2$)($H24^2$) -312.5($H24^2$) +1046.0(Tm24)(Hm3)
22	2727.0 +271.8(Tm3) +319.3($T24^2$)($H24^2$) +1768.8(T3)(H0) +230.9(T48) -556.8(H0) +363.3(Hm24)(Sm3)
23	2522.8 +536.5(H0)(S3) +389.5($T24^2$) +494.9($T0^3$) +781.2(T24)(H0) +156.8(H48) -190.6(H0)
24	2196.0 +380.8(T24)(H24)(S24) +344.0(H24)(S24) -81.4($H0^3$) +1400.0(Tm3)(Hm1) +330.3(H48) -413.7(Hm1) +164.3(delHm24)

Table B.2. Hourly regression equations; Load zone - NEMA; Variable selection method - step; Training Year - 2006; Prediction Years - 2005 and 2007

Hour	Regression Equation
1	2045.5 +591.0(T24)(H24)(S24) +430.9(H24)(S24) +561.8($T0^3$) +516.4($H3^2$) +27.0(T24)(H3) +148.2(H168) +49.6(delTm24)(delHm24)
2	1870.1 +746.2(T24)(H24)(S24) +86.9(H24)(S24) +767.1(T0)(H24) +379.4($H3^2$) -167.4(Hm3) +172.2(H168) +115.9(S72) +161.8(delTm24)
3	1950.3 +1102.6(T24)(H24)(S24) -16.4(H24)(S24) +946.7($H3^2$) -154.7(Hm3) +153.6(H168)
4	1822.9 +1101.9(T24)(H24)(S24) -215.8(H24)(S24) +496.1($H3^2$) +374.2($H24^2$) -102.6(Hm3) +213.0(H168) +91.2(S72) +174.0(delTm24)
5	1968.6 +714.1(T24)(H24)(S24) -100.5(H24)(S24) -9.7($H0^3$) +107.7(H0)(S24) +594.3($T0^3$) +159.3($T24^2$)($H24^2$) -204.2(H3) +413.9(Hm3) +122.1(H168) +447.7(H24)(S0) -180.2($S0^2$) -57.9(delTm24)(delHm24)
6	2120.9 +854.9(T24)(H24)(S24) +606.2($H0^3$) -132.9($T24^2$)($H24^2$) +345.8($H24^2$) +128.4(H168) +21.5(delTm24)(delHm24)
7	2445.9 +1041.8(T24)(H24)(S24) +163.5(H24)(S24) +696.4($H0^3$) -306.6(H0)(S24) +218.3(Tm1)
8	2733.1 +336.5(T24)(H24)(S24) +516.3(H24)(S24) +559.2($H0^3$) -79.0(T24) +416.9($T24^2$)($H24^2$) +184.6(H3)
9	2857.9 +447.3($H0^3$) +396.9(H0)(S24) +122.6($T24^2$)($H24^2$) +752.8(T0)(H24) -206.0($H3^2$) -7.4($H24^2$) +376.5(T24)(H24) +109.7(H168)
10	2992.3 +213.1(H0)(S3) +576.5(T24)(H24)(S24) +124.1($T3^2$) -8.9(H24)(S24) +403.7($H0^3$) +848.3(T0)(H24) +13.3(H24) +84.6(H168)
11	3019.5 +916.3($H0^3$) -235.6($H0^2$) +1081.6(H24)(S3) +254.3(H72) +835.8(T0)(S3) +193.1(S48) -263.7(S3) -418.9($S3^2$)
12	3063.3 +502.5(H0)(S3) +303.6($H0^3$) +790.5($T0^3$) +124.0($T24^2$)($H24^2$) +623.4($H24^2$) +162.4(H168)
13	3191.0 +301.3(H0)(S3) +716.3($H0^2$) +583.0($T0^3$) +597.9($H24^2$) -344.4(H24) -148.7(Hm1) +156.8(Hm24)(Sm3) +171.5(H168) +816.4(T24)(S3) -328.6(S3)
14	3062.4 +382.1($T24^2$)($H24^2$) +1917.4(T0)(H0) -112.7($H24^2$) -342.0(H0) +496.5(Hm24)(Sm3) +253.6(H168) +228.9(S72) -140.1(S3)
15	3053.7 +1622.6(T24)(H24)(S24) -402.6(H0)(S24) -94.4($T24^2$)($H24^2$) +102.5(T0) +1651.5(T0)(H0) -37.2($H24^2$) -162.3(H0) +214.0(H168)
16	3012.1 +1757.0(T24)(H24)(S24) -617.4(H3)(S24) +26.8($T24^2$)($H24^2$) +1089.1($H3^2$) -220.9(T24)(H24) +240.0(H168) +1157.4(T0)(S24) -329.0(S24) +44.5(delTm24)
17	3243.2 +1691.7(T24)(H24)(S24) -97.6($T24^2$)($H24^2$) +851.7($H3^2$) -376.8(H24) +314.5(H168) +1252.0(T0)(S24) -744.2(S24) -53.5(delTm24)
18	3188.1 +782.0($T3^2$) +1362.7(H3)(S24) +433.5($T24^2$)($H24^2$) -329.7(H24) +364.8(H168) +197.9(Hm24)(Sm1) +534.7(T24)(S24) -780.5(S24)
19	2968.3 +978.7(H3)(S24) -396.2(H0)(S24) +935.1($T0^3$) +514.9($T24^2$)($H24^2$) +251.5(H24)
20	3302.4 -433.3($T3^2$) -320.9($H0^3$) +937.6($T0^3$) +2118.4(T3)(H3) -710.0(H3) -268.1(delTm24)
21	3056.1 +886.2(T24)(H24)(S24) +935.5($T0^3$) -132.0($T24^2$)($H24^2$) +58.4(T3)(H0) +244.7($H3^2$) +201.3($H24^2$) -207.0(Hm1)
22	2798.1 +421.4(H0)(S24) +806.5($T0^3$) -41.0($T24^2$)($H24^2$) +793.7(T24)(H24)
23	2641.9 +927.6(H0)(S3) +414.1($H0^3$) +93.9($T24^2$)($H24^2$) -1163.9($H24^2$) +2328.3(T24)(H24) +719.3(Tm1)(Hm3) -1079.4(Tm24)(Hm3) -303.7($S3^2$) -257.7(delTm24)(delHm24)
24	2378.1 +961.3(H0)(S24) +164.2($T24^2$)($H24^2$) +1880.4(T3)(H24) +208.2($H24^2$) -971.3(T24)(H24) +210.7(T72) -485.8(Hm3) -202.3(S24)

Table B.3. Hourly regression equations; Load zone - NEMA; Variable selection method - step; Training Year - 2007; Prediction Years - 2005 and 2006

Hour	Regression Equation
1	2173.8 +383.3(T3^2) +328.6(H0)(S24) +272.4(T0^3) +762.5(T24^2)(H24^2) +105.5(H24^2)
2	2078.5 +378.3(T3^2) +301.7(H0)(S24) +197.1(T0^3) +783.2(T24^2)(H24^2) +53.3(H24^2)
3	2029.2 +364.5(T3^2) +291.9(H0)(S24) +146.7(T0^3) +755.5(T24^2)(H24^2) +38.7(H24^2)
4	2044.1 +484.6(T3^2) +176.5(H0^2) -11.3(T0^3) +813.7(T24^2)(H24^2) -60.0(H24^2) -188.3(S24) +276.5(S24^2)
5	2079.7 +206.3(H0^2) +307.9(T0^3) +961.7(T24^2)(H24^2) -337.7(H24^2) +242.7(H48) +82.0(H0)(S0)
6	2237.8 +595.6(T3^2) +61.8(H0^3) +648.2(T24^2)(H24^2) -113.7(H24^2)
7	2545.6 +566.0(T3^2) +138.6(H0^3) +573.4(T24^2)(H24^2) +213.1(H72) -153.6(H24^2) -177.9(H168)
8	2836.3 +578.4(H0^3) +589.6(T24^2) +126.6(Hm24)(Sm3) -8.6(delHm24)
9	2919.2 +627.5(T24^2)(H24^2) +354.4(T0^2) +278.6(T0)(H0) -132.8(H24^2) +235.4(H48) +316.5(H0)
10	3185.1 +457.4(Tm3) +614.4(T24^2)(H24^2) -335.1(H24^2) +954.3(Tm3)(Hm1) -19.3(delTm24) -216.5(delHm24) +98.0(S0^3)
11	3246.0 -506.7(H0^3) +341.5(H0)(S24) +929.0(Tm1) +1503.6(H3^2) -392.2(Hm3) -130.2(delTm24) -314.7(delHm24)
12	3403.7 +1271.2(T3^2) +795.0(H3^2) -323.8(delHm24)
13	3130.8 +1508.4(Tm3) +899.9(H3^2) -388.5(delHm24)
14	2766.7 +1487.3(Tm3) +733.5(H3^2) +362.2(H48)
15	3279.8 +504.2(H72) +3687.5(T3)(H3) -1263.9(H3)
16	3106.3 +1451.4(H0)(S3) -1256.4(Hm1)(Sm3) +689.2(T24)(H24)(S24) +859.3(T3^2) -152.0(H24)(S24) +456.4(T24^2) +1081.7(H3^2)
17	3116.7 +908.6(T24^2) -421.9(H0^2) +2721.6(T0)(H0) -375.4(T24)(H0) +244.1(H48) -341.0(H0)
18	2929.0 -155.3(Tm3) +1184.5(T24^2) +1644.1(T3)(H0) +237.9(H72) -393.9(H0) +413.2(delTm24)
19	2989.6 +774.4(T3^2) +850.4(H0^3) +788.6(T24^2)
20	3302.7 +456.2(T3^2) +667.2(H0^3) -150.7(H0)(S24) +1103.7(T24^2) +445.1(Tm3)(Hm24) -582.9(Hm24) -333.6(S72) +159.7(S0^2)
21	3352.7 +213.5(H0^3) +779.9(T24^2) -21.3(T0^3) +1824.4(Tm3)(Hm1) -385.0(H0) -356.1(Hm1) +203.4(Hm3)(Sm1) -321.9(Hm24)(Sm3) -509.6(S24) +334.5(S24^2) -66.8(delTm24)(delHm24)
22	3027.7 +334.1(H24)(S3) +1499.4(T24^2)(H24^2) -961.8(H24^2) +1677.9(T3)(H3) -124.5(delTm24)
23	2649.4 +712.1(T24^2) -46.8(T0^2) +1195.5(T3)(H3) -152.1(H3) -60.7(delTm24)(delHm24)
24	2385.7 +642.9(T24^2) -15.3(T0^2) +983.5(T3)(H3) -90.7(H3) -84.6(delTm24)(delHm24)

APPENDIX C

REGRESSION EQUATIONS FOR GENETIC ALGORITHM VARIABLE SELECTION METHOD - NEMA LOAD ZONE

Table C.1. Hourly regression equations; Load zone - NEMA; Variable selection method - GA; Training Year - 2005; Prediction Years - 2006 and 2007

Hour	Regression Equation
1	2124.0 +461.6(T24)(H24)(S24) +510.5(T0)(H0) +473.6(T3)(H0) +277.7($H24^2$) -89.4(Tm3)(Hm24) -26.4(S168) +236.5(T24)(S24) +88.1(T0)(S24) -103.7(delTm24) -8.8(delHm24)
2	2053.2 -692.2(T24) +237.2(Tm1) +1811.4(T24)(H0) +558.3($H24^2$) +285.1(T48) -927.6(Tm24)(Hm1) +17.1(H24) -30.1(S72) +433.8(T24)(S24) -171.4(delTm24)(delHm24)
3	1889.1 +715.1(T24)(H24)(S24) +161.6(H72) +870.7(Tm3)(Hm1) +19.0(H24) -11.4(H168) -25.9(S168) -49.4(T24)(S24) +85.1(S48)
4	1880.2 +763.4(T24)(H24)(S24) +526.5(Tm3) -468.7(T3) +225.9($H0^3$) +394.0(T0)(H0) +254.2($H24^2$) +172.0(T72) -46.4(Tm3)(Hm24) -41.8(S168) +26.3(S48)
5	1896.7 +440.3(H0)(S24) +362.5($H24^2$) +586.2(T24)(H3) +33.7(T168) +149.9(Tm1)(Hm24) +43.4(S72) -36.7($S24^2$)
6	2051.1 +198.7(H24)(S24) +242.5(H0)(S24) +697.1(T0)(H24) +297.4(Tm3)(Hm1) +95.3(T168) +162.8(Hm24)(Sm1) +68.4(S48) -97.3(S3) -144.1(T3)(S0) -27.0(delHm24)
7	2436.2 -62.4(H0)(S3) +285.8(Tm3) +114.1($T24^2$) +193.8($H0^2$) +349.2(H24)(S3) -339.9(T0) -32.3($H3^2$) +157.4(H72) +885.0(T24)(H0) -141.2(H168) -17.6(T168) -140.1(Hm24)(Sm1) +26.6(Tm3)(Hm24)
8	2751.5 +653.1(T24)(H24)(S24) +594.4(T3) +217.3($H0^3$) -474.9(Tm1) +663.8(T0)(H0) +213.3($H24^2$) -27.0(H48) +28.2(Tm24)(Hm1) -295.8(T0)(H3) -231.3(Hm24)(Sm1) +157.8(Tm3)(Hm24) -0.4(Tm24) +77.7(T0)(S3)
9	2889.9 +95.6(Hm1)(Sm3) +502.0(T24)(H24)(S24) +313.5($H0^2$) +70.0(H72) +61.4(H0) +735.2(T3)(H3) -289.3(Tm1)(Hm3) +203.6(Tm24)(Hm3) +83.7(H24)(S0) +19.9(Tm24)(Sm3)
10	2786.9 -85.9(H0)(S3) +481.8(T24)(H24)(S24) +376.8($H0^3$) +107.3($T24^2$) +91.0($T0^3$) -165.0(T0)(H0) +523.2(T3)(H24) +108.7(H72) +239.3(T48) -25.4(H48) +770.4(T0)(H3) -649.7(Tm1)(Hm3) +170.2(Hm24) -107.5(Tm24) +287.2(T0)(S3) -74.7(S3) +324.7(delHm24)
11	2910.5 +228.9(Hm1)(Sm3) +524.4($H0^2$) +107.9($T24^2$)($H24^2$) +503.5($T0^2$) +491.3(T3)(H24) +340.6(T48) +66.4(H24) +69.3(H168) -59.1(T168) +33.5(S72) -9.7(T24)(S3) +16.0(delTm24) +21.0($S0^3$)
12	2946.4 +316.0(H3)(S3) +322.2($T0^3$) +462.0(T3)(H0) +759.7(T3)(H24) +339.2(T48) -83.9(Hm3)(Sm1) +144.8(Hm3) +179.8(Tm1)(Hm24) -16.9(Tm24) +90.0(delHm24)
13	2574.5 -229.1(H24)(S3) +794.3($T0^3$) +148.2($T24^2$)($H24^2$) +539.6(H48) +331.0(Tm24)(Hm1) +149.1(H24)(S0) +1353.3(Tm3)(Hm24) -532.2(Tm1)(Hm24) +255.2(Tm1)(Sm3) -840.5(T3)(S3) +540.7(T24)(S3) +235.2(S3) +642.8(delHm24)
14	2977.6 -292.2(H3)(S3) +320.8($H0^3$) +1702.2(T3)(H24) +356.6(T48) +547.8(H3)(S0) -356.5(Hm24)(Sm1) +452.0(T24)(S3) +46.5($S3^2$) -117.7(S0) -189.8(delHm24) +467.1(delTm24)(delHm24)
15	2498.6 +883.3(Tm3) -324.1(T24) -778.7(Tm1) +764.8($T0^2$) +1099.4(T3)(H24) +463.3(T48) -15.0(Hm24)(Sm3) +104.9(Hm3) +218.8(H0)(S0) +543.4(Hm24) +4.4(Tm1)(Sm3) +507.0(delHm24) +24.3(delTm24)(delHm24)
16	2764.3 -220.3(T24) +905.0($T0^3$) +1789.4(T24)(H24) +550.6(S24) -295.4($S3^2$) +356.6(delHm24)
17	2800.0 +1994.8($T3^2$) +432.1(H24)(S24) -286.5($T24^2$)($H24^2$) +1081.4($H24^2$) +581.9(Hm3)(Sm1) +37.4(H168) +250.4(Tm24)(Sm3) -1412.1(T3)(S3) +373.5($S24^2$) +24.0(T3)(S0) +118.7(delHm24)
18	2833.4 +196.2(Tm1) +528.5(H24)(S3) +664.3($T0^3$) +147.1(T0) +1113.3(T24)(H24) +15.7(H24) +152.0($S24^2$) +77.4(Tm24)(Sm1) -526.1(Sm1) +370.8(delHm24) +234.8($S0^3$)
19	2892.5 +1507.3(Tm3) -39.2($H0^2$) -1812.0(Tm1) +1345.3($T0^2$) +560.1(T24)(H0) +320.6(H48) +521.1(Tm24)(Hm1) +21.9(T72) -86.8(T168)
20	2918.6 +178.2(T3) +940.6($T0^3$) -692.1(T0)(H0) +1287.8(T24)(H0) +95.3(H48) +193.6(Hm1) +163.6(Tm24) -184.1(S168)
21	2965.5 -279.9(T3) +901.4($T0^2$) +980.3(T24)(H0) +92.5(T48) -69.7(S168) -253.4(T3)(S24) +541.3(T24)(S24) -20.3(S48)
22	2735.6 +710.9($T0^3$) +360.3($T24^2$)($H24^2$) +192.4(T0) +765.5(Tm24)(Hm1) +246.4($S24^2$)
23	2471.6 +5.6(Tm3) +478.9($T24^2$)($H24^2$) +345.2($T0^2$) +130.5(H72) +754.4(T24)(H0) +172.1(S72) +342.9(T0)(S24) -82.0(S48)
24	2266.6 -113.3($H0^3$) +571.1(H0)(S24) +194.1(Tm1) +498.5($T0^3$) +458.8($T24^2$)($H24^2$) -2.8(T0) +151.6(Tm24)(Hm1) +282.4(Tm24)(Hm3)

Table C.2. Hourly regression equations; Load zone - NEMA; Variable selection method - GA; Training Year - 2006; Prediction Years - 2005 and 2007

Hour	Regression Equation
1	2134.4 +1023.4(T24)(H24)(S24) +279.9(T0)(H0) +699.7(T3)(H0) +352.1($H24^2$) -301.0(Tm3)(Hm24) +8.9(S168) -335.0(T24)(S24) +531.5(T0)(S24) -91.4(delTm24) -13.8(delHm24)
2	2084.5 -1364.4(T24) +771.5(Tm1) +1638.8(T24)(H0) +1019.1($H24^2$) +283.7(T48) -1006.5(Tm24)(Hm1) -150.9(H24) +48.9(S72) +683.4(T24)(S24) -251.6(delTm24)(delHm24)
3	1887.5 +1172.1(T24)(H24)(S24) -21.3(H72) +856.5(Tm3)(Hm1) +66.6(H24) +158.4(H168) +13.8(S168) -296.0(T24)(S24) +117.3(S48)
4	1873.5 +773.7(T24)(H24)(S24) +2173.0(Tm3) -2392.3(T3) -201.5($H0^3$) +1098.9(T0)(H0) +429.0($H24^2$) +115.5(T72) -104.3(Tm3)(Hm24) -1.7(S168) +59.7(S48)
5	1865.5 +306.5(H0)(S24) +629.3($H24^2$) +633.3(T24)(H3) +111.4(T168) -9.8(Tm1)(Hm24) +16.7(S72) +58.8($S24^2$)
6	2030.3 +243.6(H24)(S24) +137.7(H0)(S24) +952.3(T0)(H24) +229.9(Tm3)(Hm1) +78.8(T168) +192.8(Hm24)(Sm1) +52.5(S48) -123.9(S3) -57.3(T3)(S0) +17.5(delHm24)
7	2475.2 -215.1(H0)(S3) -128.4(Tm3) +33.3($T24^2$) +239.7($H0^2$) +925.3(H24)(S3) -246.5(T0) -316.3($H3^2$) -41.5(H72) +1315.4(T24)(H0) +158.8(H168) -28.0(T168) -599.0(Hm24)(Sm1) +373.9(Tm3)(Hm24)
8	2777.9 +847.2(T24)(H24)(S24) +299.0(T3) +85.3($H0^3$) -496.0(Tm1) +1223.4(T0)(H0) +343.5($H24^2$) +44.6(H48) +212.0(Tm24)(Hm1) -674.0(T0)(H3) -104.4(Hm24)(Sm1) +83.0(Tm3)(Hm24) -119.4(Tm24) +127.5(T0)(S3)
9	2873.7 -7.2(Hm1)(Sm3) +643.8(T24)(H24)(S24) +847.3($H0^2$) +88.0(H72) -155.7(H0) -798.8(T3)(H3) +872.2(Tm1)(Hm3) +236.2(Tm24)(Hm3) +235.7(H24)(S0) +95.2(Tm24)(Sm3)
10	2741.7 +116.7(H0)(S3) +532.1(T24)(H24)(S24) +469.8($H0^3$) -12.0($T24^2$) +380.6($T0^3$) -738.9(T0)(H0) +999.7(T3)(H24) +63.9(H72) +182.9(T48) -16.8(H48) +1924.8(T0)(H3) -1905.2(Tm1)(Hm3) +205.9(Hm24) +20.0(Tm24) +18.1(T0)(S3) +96.9(S3) +301.6(delHm24)
11	3028.0 +40.7(Hm1)(Sm3) +536.9($H0^2$) +114.6($T24^2$)($H24^2$) +448.6($T0^2$) +779.8(T3)(H24) -0.8(T48) -16.4(H24) +151.5(H168) -91.4(T168) +30.4(S72) +324.0(T24)(S3) -89.3(delTm24) +8.9($S0^3$)
12	2954.3 +364.2(H3)(S3) +146.8($T0^3$) +583.6(T3)(H0) +885.4(T3)(H24) +249.3(T48) -0.8(Hm3)(Sm1) -128.1(Hm3) +353.9(Tm1)(Hm24) -17.9(Tm24) +180.7(delHm24)
13	2650.6 +308.9(H24)(S3) +761.1($T0^3$) -251.0($T24^2$)($H24^2$) +238.7(H48) +249.3(Tm24)(Hm1) +127.7(H24)(S0) +2417.1(Tm3)(Hm24) -1063.3(Tm1)(Hm24) +782.9(Tm1)(Sm3) -1651.0(T3)(S3) +621.7(T24)(S3) +154.1(S3) +639.4(delHm24)
14	2999.2 +100.2(H3)(S3) +564.7($H0^3$) +1161.5(T3)(H24) +334.1(T48) +278.3(H3)(S0) -242.5(Hm24)(Sm1) +570.5(T24)(S3) +68.8($S3^2$) -130.1(S0) -145.7(delHm24) +344.9(delTm24)(delHm24)
15	2673.5 +151.1(Tm3) -393.3(T24) -1043.3(Tm1) +916.5($T0^2$) +1653.6(T3)(H24) +620.9(T48) +268.6(Hm24)(Sm3) -236.1(Hm3) -17.8(H0)(S0) +467.9(Hm24) +230.4(Tm1)(Sm3) +826.6(delHm24) -1.9(delTm24)(delHm24)
16	2825.7 -627.0(T24) +763.5($T0^3$) +2243.3(T24)(H24) +399.9(S24) -16.4($S3^2$) +357.5(delHm24)
17	2800.1 +1564.1($T3^2$) +409.6(H24)(S24) -37.1($T24^2$)($H24^2$) +561.6($H24^2$) +458.5(Hm3)(Sm1) +264.9(H168) +156.4(Tm24)(Sm3) -1127.6(T3)(S3) +307.3($S24^2$) +153.7(T3)(S0) +130.8(delHm24)
18	3025.2 -121.4(Tm1) +498.6(H24)(S3) +1007.4($T0^3$) -393.2(T0) +1180.6(T24)(H24) +89.3(H24) +284.6($S24^2$) +232.9(Tm24)(Sm1) -447.8(Sm1) +300.5(delHm24) +101.9($S0^3$)
19	3023.7 +1456.0(Tm3) +159.2($H0^2$) -2212.4(Tm1) +1773.0($T0^2$) -105.9(T24)(H0) +238.1(H48) +726.7(Tm24)(Hm1) +51.1(T72) -42.6(T168)
20	3111.3 -147.1(T3) +1170.6($T0^3$) -427.7(T0)(H0) +1170.1(T24)(H0) -112.3(H48) -75.1(Hm1) +240.4(Tm24) -154.2(S168)
21	3019.3 -295.8(T3) +1395.4($T0^2$) +464.8(T24)(H0) -170.7(T48) -197.1(S168) -1116.8(T3)(S24) +1510.9(T24)(S24) +193.6(S48)
22	2805.0 +969.6($T0^3$) +243.8($T24^2$)($H24^2$) +10.1(T0) +694.3(Tm24)(Hm1) +193.1($S24^2$)
23	2442.5 -90.6(Tm3) +368.3($T24^2$)($H24^2$) +763.9($T0^2$) +275.4(H72) +377.8(T24)(H0) +10.9(S72) +144.5(T0)(S24) +134.2(S48)
24	2313.6 -79.6($H0^3$) +395.1(H0)(S24) +712.9(Tm1) +687.5($T0^3$) +195.3($T24^2$)($H24^2$) -702.2(T0) +350.1(Tm24)(Hm1) +283.0(Tm24)(Hm3)

Table C.3. Hourly regression equations; Load zone - NEMA; Variable selection method - GA; Training Year - 2007; Prediction Years - 2005 and 2006

Hour	Regression Equation
1	2109.5 +637.5(T24)(H24)(S24) -640.6(T0)(H0) +1337.6(T3)(H0) +231.9($H24^2$) +182.8(Tm3)(Hm24) +36.3(S168) +194.8(T24)(S24) -119.7(T0)(S24) -23.7(delTm24) +46.8(delHm24)
2	2098.0 -409.2(T24) +240.5(Tm1) +1426.2(T24)(H0) +844.8($H24^2$) +301.0(T48) -801.1(Tm24)(Hm1) -332.3(H24) -9.4(S72) +255.7(T24)(S24) -150.6(delTm24)(delHm24)
3	1946.4 +826.9(T24)(H24)(S24) +250.6(H72) +684.6(Tm3)(Hm1) +4.2(H24) -145.9(H168) +17.0(S168) -71.4(T24)(S24) +89.3(S48)
4	1929.6 +610.6(T24)(H24)(S24) +1438.6(Tm3) -1663.8(T3) -472.8($H0^3$) +1095.4(T0)(H0) +269.9($H24^2$) +148.9(T72) +24.1(Tm3)(Hm24) +11.3(S168) +47.7(S48)
5	1974.8 -9.2(H0)(S24) +166.5($H24^2$) +819.8(T24)(H3) -6.5(T168) +189.5(Tm1)(Hm24) +78.6(S72) +45.6($S24^2$)
6	2196.8 -114.2(H24)(S24) +297.7(H0)(S24) +652.9(T0)(H24) +290.3(Tm3)(Hm1) +27.8(T168) +140.7(Hm24)(Sm1) +120.7(S48) -74.5(S3) -125.9(T3)(S0) -141.5(delHm24)
7	2487.6 +105.1(H0)(S3) +304.9(Tm3) +233.7($T24^2$) +340.8($H0^2$) +282.9(H24)(S3) -339.2(T0) -217.1($H3^2$) +152.5(H72) +463.1(T24)(H0) -135.8(H168) -14.7(T168) -333.7(Hm24)(Sm1) +266.6(Tm3)(Hm24)
8	2782.6 +486.2(T24)(H24)(S24) +679.6(T3) -181.5($H0^3$) -588.3(Tm1) +1367.2(T0)(H0) +99.0($H24^2$) +92.9(H48) +275.3(Tm24)(Hm1) -843.7(T0)(H3) -170.2(Hm24)(Sm1) +219.5(Tm3)(Hm24) -154.0(Tm24) +91.2(T0)(S3)
9	2936.0 -126.9(Hm1)(Sm3) +427.9(T24)(H24)(S24) +279.4($H0^2$) +76.1(H72) +108.6(H0) -51.8(T3)(H3) +466.3(Tm1)(Hm3) +200.7(Tm24)(Hm3) +65.9(H24)(S0) +161.8(Tm24)(Sm3)
10	2697.1 -144.0(H0)(S3) +201.5(T24)(H24)(S24) +207.6($H0^3$) +182.0($T24^2$) +375.9($T0^3$) -568.3(T0)(H0) +399.6(T3)(H24) +145.9(H72) +187.1(T48) -102.3(H48) +1675.6(T0)(H3) -1163.5(Tm1)(Hm3) +352.7(Hm24) -61.2(Tm24) +85.0(T0)(S3) +54.4(S3) +454.5(delHm24)
11	3164.7 +3.1(Hm1)(Sm3) +476.3($H0^2$) +81.3($T24^2$)($H24^2$) +701.0($T0^2$) +579.2(T3)(H24) +142.5(T48) -60.9(H24) +170.8(H168) -106.2(T168) -20.5(S72) +2.7(T24)(S3) -147.4(delTm24) +58.5($S0^3$)
12	3075.5 -62.4(H3)(S3) +462.5($T0^3$) +381.7(T3)(H0) +840.7(T3)(H24) +210.3(T48) -52.0(Hm3)(Sm1) +228.5(Hm3) +156.9(Tm1)(Hm24) -26.7(Tm24) +108.2(delHm24)
13	2699.6 -334.9(H24)(S3) +429.4($T0^3$) +482.7($T24^2$)($H24^2$) +359.5(H48) +157.5(Tm24)(Hm1) +152.4(H24)(S0) +1142.6(Tm3)(Hm24) -62.0(Tm1)(Hm24) +956.0(Tm1)(Sm3) -593.2(T3)(S3) +98.0(T24)(S3) -80.6(S3) +711.8(delHm24)
14	2932.5 -126.7(H3)(S3) +405.5($H0^3$) +1653.5(T3)(H24) +223.6(T48) -61.1(H3)(S0) -147.7(Hm24)(Sm1) +546.9(T24)(S3) -67.9($S3^2$) -52.7(S0) +171.9(delHm24) +442.6(delTm24)(delHm24)
15	2465.2 +1071.9(Tm3) -33.5(T24) -1125.6(Tm1) +742.0($T0^2$) +1273.6(T3)(H24) +185.3(T48) -310.6(Hm24)(Sm3) -287.0(Hm3) +184.4(H0)(S0) +868.6(Hm24) -10.5(Tm1)(Sm3) +867.3(delHm24) +134.2(delTm24)(delHm24)
16	2745.1 -59.9(T24) +777.7($T0^3$) +1640.5(T24)(H24) +33.4(S24) +116.1($S3^2$) +618.8(delHm24)
17	2834.2 +1441.1($T3^2$) -145.0(H24)(S24) +670.5($T24^2$)($H24^2$) +457.3($H24^2$) +427.0(Hm3)(Sm1) +140.8(H168) +243.6(Tm24)(Sm3) -755.9(T3)(S3) +215.9($S24^2$) +233.6(T3)(S0) +314.3(delHm24)
18	2731.7 +1239.1(Tm1) -139.6(H24)(S3) +858.9($T0^3$) -1137.6(T0) +1382.8(T24)(H24) +142.2(H24) +187.7($S24^2$) +49.1(Tm24)(Sm1) -108.7(Sm1) +466.3(delHm24) +68.3($S0^3$)
19	2988.9 +835.1(Tm3) -416.2($H0^2$) -1027.6(Tm1) +993.9($T0^2$) +2234.6(T24)(H0) +257.1(H48) -591.7(Tm24)(Hm1) +50.7(T72) -111.1(T168)
20	3117.9 +388.1(T3) +1123.5($T0^3$) -2212.4(T0)(H0) +2791.2(T24)(H0) +51.5(H48) +252.2(Hm1) -352.2(Tm24) -206.3(S168)
21	3023.4 +73.1(T3) +835.4($T0^2$) +1048.3(T24)(H0) -211.2(T48) -161.3(S168) -842.6(T3)(S24) +976.7(T24)(S24) +82.0(S48)
22	2839.4 +810.4($T0^3$) +645.2($T24^2$)($H24^2$) +172.3(T0) +477.2(Tm24)(Hm1) +167.2($S24^2$)
23	2552.8 -114.2(Tm3) +788.9($T24^2$)($H24^2$) +638.3($T0^2$) +67.9(H72) +408.6(T24)(H0) +9.7(S72) +256.7(T0)(S24) +113.6(S48)
24	2330.1 -289.6($H0^3$) +405.5(H0)(S24) +126.3(Tm1) +599.6($T0^3$) +509.7($T24^2$)($H24^2$) -1.1(T0) -174.1(Tm24)(Hm1) +684.8(Tm24)(Hm3)

APPENDIX D

REGIONAL PREDICTION RESULTS FOR THE STEPWISE AND GENETIC ALGORITHM-BASED VARIABLE SELECTION METHODS

Table D.1. Hourly mean absolute percent error (MAPE) – regional; Variable selection method - step

	Tr - 05, Pr - 06	**Tr - 05, Pr - 07**	**Tr - 06, Pr - 05**	**Tr - 06, Pr - 07**	**Tr - 07, Pr - 05**	**Tr - 07, Pr - 06**	**Hrly Avg**
Hour 1	2.55	4.15	2.95	3.16	4.52	4.35	3.61
Hour 2	2.21	4.11	2.38	3.29	4.38	4.19	3.43
Hour 3	2.06	3.51	2.00	4.94	3.63	4.03	3.36
Hour 4	1.86	3.74	2.06	4.12	3.20	4.00	3.17
Hour 5	1.67	3.62	1.72	3.64	3.25	3.62	2.92
Hour 6	1.62	2.56	1.74	3.56	2.51	3.19	2.53
Hour 7	2.34	2.99	1.95	3.24	2.94	2.75	2.70
Hour 8	2.71	3.72	3.88	1.88	2.76	2.41	2.89
Hour 9	2.35	2.36	2.26	1.50	2.48	1.74	2.12
Hour 10	2.47	2.90	2.01	1.41	1.97	1.69	2.08
Hour 11	2.29	2.67	2.89	1.83	2.40	1.98	2.34
Hour 12	3.16	3.19	4.02	1.86	2.46	2.74	2.91
Hour 13	2.98	2.18	3.03	2.18	2.55	2.79	2.62
Hour 14	4.21	2.99	4.91	2.96	2.60	2.58	3.37
Hour 15	3.50	2.77	3.34	2.87	2.87	2.45	2.97
Hour 16	3.04	2.22	3.12	2.65	2.76	2.79	2.76
Hour 17	2.87	2.47	3.49	2.66	2.49	2.52	2.75
Hour 18	2.62	2.45	3.72	2.61	2.49	2.39	2.71
Hour 19	3.98	3.11	3.35	2.72	2.96	2.50	3.10
Hour 20	3.36	2.90	4.42	3.52	3.43	3.02	3.44
Hour 21	2.43	2.33	3.26	2.45	2.81	2.30	2.60
Hour 22	2.66	2.83	3.17	2.26	2.94	2.33	2.70
Hour 23	1.97	2.41	3.09	1.93	2.82	2.35	2.43
Hour 24	2.05	2.03	2.65	1.60	3.05	2.37	2.29
Average	2.62	2.93	2.98	2.70	2.93	2.80	2.83

Table D.2. Mean absolute percent error (MAPE) during top 5 load hours – regional; Variable selection method - step

	Tr - 05, Pr - 06	**Tr - 05, Pr - 07**	**Tr - 06, Pr - 05**	**Tr - 06, Pr - 07**	**Tr - 07, Pr - 05**	**Tr - 07, Pr - 06**	**Hrly Avg**
Hour 1	3.19	2.19	3.76	1.90	7.28	7.12	4.24
Hour 2	4.58	1.58	2.62	2.45	7.03	7.04	4.22
Hour 3	4.24	2.96	1.33	5.57	6.17	7.10	4.56
Hour 4	3.91	3.30	2.03	4.41	5.80	7.92	4.56
Hour 5	3.82	3.78	1.02	5.76	5.47	7.76	4.60
Hour 6	4.13	2.25	1.20	5.95	4.73	7.63	4.31
Hour 7	1.54	3.88	2.86	5.69	5.06	7.88	4.48
Hour 8	0.60	4.67	4.08	2.41	6.00	7.09	4.14
Hour 9	1.20	3.62	1.84	2.15	4.13	4.26	2.87
Hour 10	2.04	3.75	1.77	1.85	2.12	3.81	2.56
Hour 11	2.16	3.79	3.89	2.15	4.05	3.17	3.20
Hour 12	3.37	4.34	5.30	3.05	3.13	2.94	3.69
Hour 13	3.27	2.19	4.73	2.22	4.46	2.77	3.28
Hour 14	5.06	1.98	6.91	4.01	4.15	2.34	4.07
Hour 15	3.83	1.37	4.59	2.50	2.98	2.91	3.03
Hour 16	4.63	2.28	3.40	2.58	2.52	3.47	3.15
Hour 17	3.26	3.29	4.21	2.01	1.69	3.07	2.92
Hour 18	3.63	1.40	3.03	2.58	1.31	4.44	2.73
Hour 19	4.56	2.40	2.56	2.62	2.89	5.28	3.38
Hour 20	4.00	2.14	4.94	4.42	4.17	6.99	4.44
Hour 21	4.16	1.83	3.95	1.92	4.15	6.65	3.78
Hour 22	3.97	2.47	3.32	1.85	3.76	6.09	3.58
Hour 23	4.88	2.38	3.58	1.55	3.45	5.68	3.59
Hour 24	5.35	1.87	4.68	2.87	4.81	6.23	4.30
Average	3.56	2.74	3.40	3.10	4.22	5.40	3.74

Table D.3. Hourly mean absolute percent error (MAPE) – regional; Variable selection method - GA

	Tr - 05, Pr - 06	**Tr - 05, Pr - 07**	**Tr - 06, Pr - 05**	**Tr - 06, Pr - 07**	**Tr - 07, Pr - 05**	**Tr - 07, Pr - 06**	**Hrly Avg**
Hour 1	2.41	2.87	2.06	3.12	3.08	2.90	2.74
Hour 2	2.23	2.89	2.01	2.31	2.73	2.93	2.52
Hour 3	2.02	2.63	1.97	3.07	2.54	2.57	2.47
Hour 4	1.81	2.71	1.80	3.03	2.81	2.41	2.43
Hour 5	1.79	2.99	1.76	2.78	2.65	2.45	2.40
Hour 6	1.55	2.48	1.43	2.78	2.60	2.53	2.23
Hour 7	1.82	2.09	1.79	2.36	2.40	2.33	2.13
Hour 8	1.58	1.82	1.86	1.87	2.15	2.10	1.90
Hour 9	1.41	1.54	1.87	1.06	2.11	1.54	1.59
Hour 10	1.28	1.37	1.68	1.21	1.91	1.54	1.50
Hour 11	1.51	1.44	1.76	1.34	2.06	1.54	1.61
Hour 12	1.79	1.56	1.77	1.45	1.89	1.72	1.70
Hour 13	1.97	1.66	2.30	1.74	2.08	1.94	1.95
Hour 14	2.06	2.09	2.53	1.99	2.58	2.14	2.23
Hour 15	2.39	1.93	2.64	2.06	2.59	2.17	2.30
Hour 16	2.24	1.99	2.54	1.98	2.61	2.08	2.24
Hour 17	2.36	2.28	2.62	2.08	2.51	2.15	2.33
Hour 18	2.71	2.33	2.64	2.11	2.46	2.38	2.44
Hour 19	2.50	2.37	2.73	2.62	2.35	2.48	2.51
Hour 20	2.78	2.60	2.96	2.68	2.84	2.76	2.77
Hour 21	2.21	2.33	2.46	2.10	2.58	2.21	2.32
Hour 22	2.29	2.04	2.59	1.78	2.56	2.22	2.25
Hour 23	2.34	1.94	2.32	1.63	2.29	2.19	2.12
Hour 24	2.07	1.67	2.01	1.38	2.12	2.00	1.88
Average	2.05	2.15	2.17	2.11	2.44	2.22	2.19

Table D.4. Mean absolute percent error (MAPE) during top 5 load hours – regional; Variable selection method - GA

	Tr - 05, Pr - 06	**Tr - 05, Pr - 07**	**Tr - 06, Pr - 05**	**Tr - 06, Pr - 07**	**Tr - 07, Pr - 05**	**Tr - 07, Pr - 06**	**Hrly Avg**
Hour 1	4.57	1.66	2.20	2.38	6.16	7.17	4.02
Hour 2	3.94	1.23	1.86	2.30	5.08	6.69	3.52
Hour 3	4.45	1.10	1.39	2.42	4.76	6.33	3.41
Hour 4	3.95	1.22	1.31	2.73	5.87	6.56	3.61
Hour 5	4.15	1.56	1.73	2.13	5.51	7.04	3.69
Hour 6	4.21	1.35	1.32	3.05	5.21	7.81	3.82
Hour 7	4.09	1.76	2.55	3.40	5.20	6.91	3.98
Hour 8	3.22	1.07	1.69	1.64	5.27	7.02	3.32
Hour 9	1.24	1.49	2.30	0.99	4.45	4.57	2.51
Hour 10	1.75	1.56	1.24	1.92	3.10	4.35	2.32
Hour 11	1.77	1.96	2.10	2.15	3.80	3.59	2.56
Hour 12	1.73	2.10	2.27	2.26	2.81	2.40	2.26
Hour 13	2.05	1.98	3.52	2.62	2.60	2.61	2.56
Hour 14	2.28	1.24	3.82	1.89	4.04	3.06	2.72
Hour 15	2.95	0.58	3.91	2.70	3.45	3.00	2.76
Hour 16	2.67	1.51	2.56	2.05	3.09	2.63	2.42
Hour 17	2.58	1.91	3.07	1.71	2.32	3.20	2.46
Hour 18	4.27	1.87	2.23	2.14	2.02	5.07	2.93
Hour 19	4.81	2.06	3.47	3.26	2.26	5.35	3.53
Hour 20	4.58	2.35	3.19	2.79	2.89	5.89	3.61
Hour 21	4.66	1.46	2.89	1.67	3.55	5.65	3.31
Hour 22	4.24	1.95	3.74	1.92	3.30	5.33	3.42
Hour 23	4.65	1.85	3.50	2.52	3.03	5.22	3.46
Hour 24	4.76	1.04	3.22	1.80	3.84	5.50	3.36
Average	3.48	1.58	2.54	2.27	3.90	5.12	3.15

APPENDIX E

REGRESSION RESULTS FOR STEPWISE VARIABLE SELECTION METHOD

Table E.1. Hourly mean absolute percent error (MAPE); Load zone - NEMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.62	3.17	3.17	4.17	3.17	3.98	3.38
Hour 2	2.72	3.02	3.53	4.01	3.19	4.06	3.42
Hour 3	4.53	3.93	2.86	4.08	3.14	3.99	3.75
Hour 4	2.80	3.01	3.26	3.64	3.30	4.27	3.38
Hour 5	2.79	3.21	3.65	4.05	3.39	4.69	3.63
Hour 6	4.05	3.59	2.61	3.92	3.46	4.30	3.66
Hour 7	2.75	2.66	2.90	3.54	2.93	3.75	3.09
Hour 8	2.89	2.36	2.40	2.73	3.36	2.73	2.74
Hour 9	2.23	2.32	2.33	2.39	2.86	2.79	2.48
Hour 10	2.01	1.90	2.27	2.41	4.73	3.35	2.78
Hour 11	2.24	2.53	4.34	3.73	3.37	4.34	3.43
Hour 12	2.02	2.81	3.65	2.91	4.66	4.53	3.43
Hour 13	2.24	2.63	2.81	4.20	5.14	6.28	3.88
Hour 14	2.49	3.30	4.20	6.24	3.73	4.98	4.16
Hour 15	3.14	6.29	3.12	5.08	3.22	6.04	4.48
Hour 16	3.33	5.26	3.51	5.24	4.32	5.37	4.51
Hour 17	3.02	5.10	4.30	6.51	5.02	6.99	5.16
Hour 18	4.82	3.76	4.22	6.57	6.38	4.22	5.00
Hour 19	6.25	3.94	4.31	6.24	4.11	4.50	4.89
Hour 20	3.69	4.78	4.96	7.36	5.16	4.13	5.01
Hour 21	2.63	5.00	3.44	5.05	4.52	5.75	4.40
Hour 22	3.63	4.14	2.86	3.66	3.42	5.21	3.82
Hour 23	2.19	3.13	4.59	5.60	3.53	4.17	3.87
Hour 24	2.87	3.39	2.95	4.99	3.44	3.84	3.58
Average	3.08	3.55	3.43	4.51	3.90	4.51	3.83

Table E.2. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - NEMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.90	3.57	3.35	5.86	2.59	7.54	4.47
Hour 2	4.23	3.52	3.75	3.40	2.58	7.57	4.18
Hour 3	7.01	1.45	2.71	4.68	2.28	7.46	4.26
Hour 4	4.40	3.04	2.96	3.47	1.60	7.96	3.90
Hour 5	4.17	4.54	4.02	7.36	2.00	7.75	4.97
Hour 6	6.68	1.22	1.55	5.07	2.38	8.95	4.31
Hour 7	3.78	2.37	1.62	5.68	2.76	8.15	4.06
Hour 8	3.58	2.39	1.71	3.84	1.97	5.09	3.10
Hour 9	1.36	2.40	3.76	2.13	3.47	3.55	2.78
Hour 10	2.42	1.24	2.88	2.28	3.62	4.70	2.86
Hour 11	1.59	2.29	4.59	5.11	3.74	3.94	3.55
Hour 12	1.54	2.63	3.34	2.87	3.32	1.90	2.60
Hour 13	1.88	1.23	3.84	6.49	2.57	0.80	2.80
Hour 14	2.11	3.89	5.87	10.55	3.51	1.00	4.49
Hour 15	1.38	9.13	4.12	7.40	5.28	9.92	6.21
Hour 16	4.98	5.75	5.19	8.25	4.09	5.68	5.66
Hour 17	2.30	5.92	6.04	10.40	3.43	5.59	5.61
Hour 18	4.62	3.10	6.17	11.54	4.23	2.88	5.42
Hour 19	7.61	3.73	5.87	10.45	1.62	3.77	5.51
Hour 20	1.27	5.29	9.33	12.14	2.45	3.94	5.74
Hour 21	3.49	6.20	5.47	7.60	2.82	1.40	4.50
Hour 22	0.85	3.52	4.18	5.59	4.53	3.15	3.64
Hour 23	3.01	3.25	4.99	6.55	3.01	2.14	3.82
Hour 24	3.14	4.26	5.16	10.39	3.22	2.58	4.79
Average	3.39	3.58	4.27	6.63	3.04	4.89	4.30

Table E.3. Hourly mean absolute percent error (MAPE); Load zone - SEMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	4.87	5.84	5.51	4.05	5.72	4.01	5.00
Hour 2	4.61	6.14	5.06	3.90	5.53	3.85	4.85
Hour 3	4.21	5.94	3.80	4.02	4.66	3.95	4.43
Hour 4	3.89	5.35	3.59	3.92	4.01	3.79	4.09
Hour 5	3.67	5.21	3.12	3.86	3.60	3.61	3.84
Hour 6	2.94	4.30	3.74	3.88	4.24	3.40	3.75
Hour 7	4.09	5.46	4.76	3.99	4.52	3.78	4.43
Hour 8	2.74	5.33	4.81	2.91	4.65	3.57	4.00
Hour 9	3.08	4.35	4.43	2.57	4.10	3.38	3.65
Hour 10	2.97	4.88	4.16	2.50	3.77	3.29	3.60
Hour 11	3.85	3.85	4.49	3.34	3.85	3.08	3.74
Hour 12	4.41	5.18	4.81	3.52	4.52	3.58	4.34
Hour 13	4.60	5.60	6.33	4.23	4.80	4.20	4.96
Hour 14	5.42	4.56	5.67	4.51	4.48	5.50	5.02
Hour 15	4.33	4.25	5.79	4.62	4.52	5.08	4.76
Hour 16	4.87	4.13	5.74	5.10	4.50	5.89	5.04
Hour 17	5.01	4.37	5.80	5.04	4.82	4.82	4.98
Hour 18	5.71	5.15	6.47	4.89	4.38	4.97	5.26
Hour 19	7.55	7.65	6.04	4.05	4.67	4.10	5.68
Hour 20	4.60	5.35	6.49	4.78	4.57	3.73	4.92
Hour 21	5.00	3.74	5.32	3.36	4.05	3.60	4.18
Hour 22	7.04	7.43	6.44	3.38	4.25	3.07	5.27
Hour 23	4.32	4.32	6.96	3.43	4.83	5.62	4.91
Hour 24	4.55	3.70	5.93	3.63	5.37	3.84	4.50
Average	4.51	5.09	5.22	3.90	4.52	4.07	4.55

Table E.4. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - SEMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.93	4.02	7.27	6.27	6.03	5.71	5.54
Hour 2	3.67	4.31	5.97	5.65	7.41	5.99	5.50
Hour 3	4.45	3.52	3.36	5.62	5.30	7.64	4.98
Hour 4	4.49	3.11	4.48	5.09	5.39	7.86	5.07
Hour 5	4.09	3.47	3.87	5.77	5.19	8.89	5.21
Hour 6	3.30	3.91	1.66	6.73	5.98	9.24	5.14
Hour 7	1.98	2.75	4.46	3.74	5.83	9.64	4.73
Hour 8	2.45	3.82	3.50	2.94	8.19	9.62	5.09
Hour 9	2.23	2.13	6.57	1.75	7.17	8.91	4.79
Hour 10	2.48	3.09	5.60	2.92	6.99	6.94	4.67
Hour 11	2.20	3.87	5.75	3.63	5.90	4.16	4.25
Hour 12	2.45	4.13	6.29	4.73	7.26	6.73	5.27
Hour 13	2.20	3.96	5.99	5.57	6.71	6.40	5.14
Hour 14	3.26	3.49	5.75	3.49	5.56	6.35	4.65
Hour 15	2.98	4.05	6.46	3.64	4.30	3.11	4.09
Hour 16	3.41	4.65	5.67	6.14	5.01	4.35	4.87
Hour 17	5.38	3.95	6.89	4.48	6.10	4.03	5.14
Hour 18	5.45	4.98	6.79	5.21	4.37	4.79	5.26
Hour 19	5.99	6.08	5.26	4.86	3.38	5.90	5.25
Hour 20	6.98	3.97	5.04	5.59	4.50	6.37	5.41
Hour 21	7.71	4.56	3.78	4.24	3.01	9.35	5.44
Hour 22	3.95	4.48	7.23	5.35	5.19	7.88	5.68
Hour 23	6.69	1.78	5.72	5.50	3.27	6.24	4.87
Hour 24	3.69	2.46	9.03	7.78	5.59	4.11	5.44
Average	3.97	3.77	5.52	4.86	5.57	6.68	5.06

Table E.5. Hourly mean absolute percent error (MAPE); Load zone - WCMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.51	6.64	4.77	5.61	6.55	5.08	5.36
Hour 2	3.31	5.78	4.60	5.23	6.46	4.82	5.03
Hour 3	3.16	5.53	4.60	8.49	6.13	4.52	5.40
Hour 4	3.15	6.13	3.94	4.94	5.47	4.29	4.65
Hour 5	2.97	6.43	3.74	4.58	5.03	4.33	4.51
Hour 6	2.80	3.84	3.36	4.16	4.23	3.58	3.66
Hour 7	2.74	3.44	3.34	3.44	3.39	2.71	3.18
Hour 8	3.09	4.79	4.93	3.02	4.02	2.86	3.78
Hour 9	2.56	3.21	3.53	2.58	4.00	2.07	2.99
Hour 10	3.99	3.63	3.99	3.01	2.88	2.50	3.33
Hour 11	3.93	4.42	3.75	2.72	3.35	2.93	3.52
Hour 12	3.13	2.65	5.52	3.50	3.26	3.15	3.53
Hour 13	2.78	2.84	4.10	2.90	3.75	3.01	3.23
Hour 14	5.42	4.02	6.30	3.89	4.56	2.47	4.44
Hour 15	7.36	5.63	3.89	3.06	5.81	2.66	4.74
Hour 16	3.45	2.77	6.30	4.65	3.84	2.85	3.97
Hour 17	3.46	2.77	3.62	2.77	3.37	2.60	3.10
Hour 18	3.57	3.18	6.75	4.01	3.59	3.20	4.05
Hour 19	4.72	4.63	4.03	3.23	4.08	3.28	3.99
Hour 20	3.72	3.34	3.99	3.53	4.13	3.64	3.72
Hour 21	3.53	3.26	3.57	4.48	3.52	3.38	3.62
Hour 22	2.74	2.81	3.94	3.07	4.38	3.70	3.44
Hour 23	3.43	3.71	3.77	2.74	3.77	2.94	3.39
Hour 24	4.27	4.71	3.65	2.97	3.43	2.71	3.62
Average	3.62	4.17	4.33	3.86	4.29	3.30	3.93

Table E.6. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - WCMA; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.99	4.41	4.13	3.99	11.05	8.40	6.33
Hour 2	5.96	3.44	3.76	6.19	10.38	6.37	6.02
Hour 3	5.97	3.13	3.36	9.22	10.51	6.95	6.52
Hour 4	5.02	4.15	3.73	4.45	10.84	9.42	6.27
Hour 5	4.42	5.05	2.72	5.23	9.12	9.38	5.98
Hour 6	5.28	2.13	3.81	6.99	9.62	9.78	6.27
Hour 7	4.92	3.32	5.71	5.42	7.15	6.04	5.43
Hour 8	1.96	5.73	6.89	2.50	8.37	7.63	5.51
Hour 9	2.79	3.92	4.63	2.05	8.06	5.04	4.41
Hour 10	7.25	4.18	4.37	2.27	4.60	6.51	4.86
Hour 11	8.12	4.93	4.23	2.75	4.30	7.06	5.23
Hour 12	4.74	1.92	5.56	5.20	2.51	2.92	3.81
Hour 13	4.66	2.46	6.25	4.71	6.06	5.23	4.89
Hour 14	7.47	5.15	8.29	1.70	6.57	2.86	5.34
Hour 15	8.29	2.56	4.96	1.91	6.94	3.14	4.63
Hour 16	5.03	1.75	4.70	5.54	4.99	4.35	4.39
Hour 17	4.97	1.53	2.96	1.76	3.95	3.76	3.15
Hour 18	3.77	1.90	7.79	3.97	5.01	4.28	4.45
Hour 19	4.01	3.18	5.02	4.33	6.15	7.03	4.95
Hour 20	3.04	2.92	6.56	4.42	6.83	8.39	5.36
Hour 21	3.09	3.20	3.51	6.44	6.29	6.70	4.87
Hour 22	5.36	3.07	5.15	2.69	5.36	7.13	4.79
Hour 23	2.88	3.22	4.39	2.43	5.45	5.95	4.05
Hour 24	4.82	4.86	5.48	2.06	5.77	6.77	4.96
Average	4.99	3.42	4.91	4.09	6.91	6.29	5.10

Table E.7. Hourly mean absolute percent error (MAPE); Load zone - CT; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.69	6.70	7.09	4.82	8.02	6.18	6.08
Hour 2	3.30	6.49	4.06	5.06	7.87	5.51	5.38
Hour 3	2.88	6.59	3.35	9.34	5.88	5.74	5.63
Hour 4	2.84	6.25	3.36	9.18	5.69	5.54	5.48
Hour 5	2.66	5.75	3.52	5.95	6.08	4.74	4.78
Hour 6	2.46	5.08	3.41	7.28	4.65	4.31	4.53
Hour 7	4.88	5.59	2.77	7.55	4.68	3.73	4.87
Hour 8	5.71	6.41	7.07	2.51	4.43	3.67	4.97
Hour 9	5.56	3.88	2.84	2.54	4.20	2.00	3.50
Hour 10	5.75	5.65	2.40	2.18	3.49	1.99	3.58
Hour 11	4.45	4.81	3.28	2.24	3.34	2.72	3.47
Hour 12	8.46	9.63	6.61	2.57	3.39	3.71	5.73
Hour 13	6.15	3.35	4.27	4.69	5.18	3.46	4.52
Hour 14	7.62	8.81	9.23	3.58	4.86	3.87	6.33
Hour 15	8.62	10.19	4.85	4.89	4.99	4.47	6.33
Hour 16	5.45	5.91	4.92	5.19	8.38	3.55	5.57
Hour 17	6.23	6.54	5.37	5.20	5.54	3.23	5.35
Hour 18	4.55	4.13	4.72	4.25	5.55	3.91	4.52
Hour 19	4.94	5.48	4.76	6.05	5.81	4.90	5.32
Hour 20	4.38	5.94	6.08	4.50	6.29	4.89	5.34
Hour 21	3.97	4.46	4.46	3.38	4.67	5.02	4.33
Hour 22	3.58	4.73	5.28	4.76	4.48	3.26	4.35
Hour 23	3.25	5.12	3.88	4.44	4.59	3.00	4.05
Hour 24	3.38	4.28	3.40	3.65	6.14	4.54	4.23
Average	4.78	5.91	4.62	4.82	5.34	4.08	4.93

Table E.8. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - CT; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.51	5.35	7.81	5.77	10.05	8.47	7.16
Hour 2	9.47	4.05	4.55	8.10	10.56	7.86	7.43
Hour 3	4.79	6.08	3.41	14.27	9.11	9.81	7.91
Hour 4	4.67	5.88	2.39	12.80	8.81	10.58	7.52
Hour 5	4.23	6.07	4.21	9.38	9.23	8.46	6.93
Hour 6	4.85	4.38	4.99	11.45	8.77	7.15	6.93
Hour 7	4.05	6.06	3.17	9.94	7.43	7.13	6.30
Hour 8	4.89	6.97	11.11	2.79	7.52	7.95	6.87
Hour 9	6.00	5.11	2.71	4.27	5.50	3.23	4.47
Hour 10	7.23	8.19	2.18	3.77	4.56	2.74	4.78
Hour 11	7.33	7.68	3.80	2.11	4.84	2.58	4.72
Hour 12	9.37	13.26	8.26	2.73	4.95	2.54	6.85
Hour 13	7.44	4.06	2.08	6.07	6.94	2.86	4.91
Hour 14	8.93	12.23	11.54	2.65	5.10	2.51	7.16
Hour 15	9.78	12.50	3.97	6.03	7.53	5.98	7.63
Hour 16	6.81	7.73	3.66	6.45	9.26	4.58	6.42
Hour 17	6.80	9.94	3.84	6.63	6.69	4.32	6.37
Hour 18	6.67	4.56	3.51	5.28	7.82	4.76	5.43
Hour 19	4.68	6.19	2.34	8.04	6.65	6.07	5.66
Hour 20	5.27	2.90	8.08	3.72	9.68	8.75	6.40
Hour 21	6.09	5.20	6.99	2.27	8.86	8.81	6.37
Hour 22	5.29	4.82	7.18	4.88	7.12	7.44	6.12
Hour 23	7.49	4.78	4.80	4.00	6.76	5.36	5.53
Hour 24	8.75	4.14	4.66	4.50	6.46	7.96	6.08
Average	6.52	6.59	5.05	6.16	7.51	6.16	6.33

Table E.9. Hourly mean absolute percent error (MAPE); Load zone - RI; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	4.18	4.80	5.82	3.87	3.99	4.47	4.52
Hour 2	4.85	5.96	4.90	3.90	4.93	5.24	4.96
Hour 3	4.19	4.36	4.24	4.06	4.34	4.41	4.27
Hour 4	3.88	3.85	3.98	3.77	4.71	4.26	4.08
Hour 5	3.46	3.66	4.36	4.90	3.96	4.57	4.15
Hour 6	3.18	3.11	4.62	4.68	3.73	5.24	4.09
Hour 7	3.08	4.17	4.26	4.27	4.13	4.48	4.06
Hour 8	2.94	3.61	5.35	3.11	4.61	3.28	3.82
Hour 9	3.86	3.70	3.50	2.62	4.13	2.88	3.45
Hour 10	3.56	4.14	3.56	3.35	4.03	3.42	3.68
Hour 11	3.77	4.62	5.71	4.16	4.06	3.57	4.32
Hour 12	4.51	5.03	5.17	4.44	3.88	4.60	4.60
Hour 13	4.90	5.01	9.33	5.85	4.23	5.17	5.75
Hour 14	6.94	5.44	6.66	5.27	4.77	5.38	5.74
Hour 15	6.70	4.78	6.26	6.27	4.49	5.03	5.59
Hour 16	6.62	5.08	6.68	6.27	5.63	4.78	5.84
Hour 17	5.88	3.94	7.37	7.21	7.06	4.74	6.03
Hour 18	5.12	4.54	5.13	4.98	6.29	5.29	5.22
Hour 19	6.29	5.16	5.06	4.65	5.93	4.29	5.23
Hour 20	5.23	4.61	5.52	4.07	5.75	4.86	5.01
Hour 21	5.74	4.41	5.31	5.42	4.95	3.42	4.87
Hour 22	3.17	4.39	5.30	3.52	5.44	3.77	4.26
Hour 23	3.39	4.57	4.13	5.07	3.94	4.59	4.28
Hour 24	4.41	3.54	5.62	4.65	3.90	3.79	4.32
Average	4.58	4.44	5.33	4.60	4.70	4.40	4.67

Table E.10. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - RI; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	4.06	1.63	4.48	5.94	2.32	5.75	4.03
Hour 2	3.64	1.03	4.46	4.84	5.35	7.88	4.53
Hour 3	6.03	2.89	3.81	4.07	4.68	6.71	4.70
Hour 4	6.79	3.82	3.49	3.37	5.20	6.49	4.86
Hour 5	7.64	3.62	4.47	4.93	4.05	9.15	5.64
Hour 6	6.52	2.13	4.70	2.92	2.40	9.60	4.71
Hour 7	3.40	2.57	4.66	3.72	6.27	9.49	5.02
Hour 8	4.58	1.44	6.10	1.69	7.92	7.03	4.80
Hour 9	2.83	2.43	2.54	2.73	5.28	5.19	3.50
Hour 10	2.36	2.12	4.54	4.49	4.65	3.69	3.64
Hour 11	3.69	2.20	3.08	4.83	3.48	3.00	3.38
Hour 12	2.02	3.50	4.58	3.99	2.48	4.68	3.54
Hour 13	3.00	2.04	7.63	5.84	2.08	4.20	4.13
Hour 14	2.40	2.84	9.14	6.92	4.48	4.48	5.05
Hour 15	3.34	1.95	6.94	9.79	4.57	2.93	4.92
Hour 16	2.28	1.93	6.90	6.61	5.10	4.24	4.51
Hour 17	2.07	2.41	4.62	5.78	4.64	2.96	3.75
Hour 18	3.90	3.29	6.05	5.34	3.07	3.16	4.14
Hour 19	4.05	3.56	4.49	3.94	4.10	3.92	4.01
Hour 20	4.60	3.86	4.42	3.25	6.69	6.67	4.91
Hour 21	4.77	3.14	5.89	7.54	7.63	9.16	6.35
Hour 22	6.25	4.89	7.66	4.20	5.87	7.76	6.10
Hour 23	5.53	3.60	3.51	5.45	3.18	10.16	5.24
Hour 24	7.60	1.75	4.31	6.58	3.17	7.33	5.12
Average	4.31	2.69	5.10	4.95	4.53	6.07	4.61

Table E.11. Hourly mean absolute percent error (MAPE); Load zone - ME; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.06	3.29	3.42	4.69	3.85	3.51	3.64
Hour 2	3.49	4.41	2.50	3.69	4.14	3.27	3.58
Hour 3	3.61	3.91	2.44	3.47	4.26	3.19	3.48
Hour 4	3.29	3.39	3.11	2.83	3.68	2.93	3.20
Hour 5	2.95	3.09	2.73	2.73	3.95	3.13	3.10
Hour 6	2.41	3.05	2.61	2.39	3.09	2.52	2.68
Hour 7	4.10	3.45	3.03	2.45	4.42	2.40	3.31
Hour 8	3.58	4.26	3.43	2.14	3.59	2.04	3.17
Hour 9	2.69	3.38	2.34	2.19	4.16	2.32	2.84
Hour 10	2.89	3.57	2.83	2.64	3.28	2.31	2.92
Hour 11	3.08	4.43	2.73	2.85	3.42	2.30	3.14
Hour 12	2.58	3.22	3.40	3.10	3.56	2.51	3.06
Hour 13	2.58	3.11	3.81	3.56	3.50	2.41	3.16
Hour 14	3.57	3.61	3.97	3.61	2.84	3.07	3.45
Hour 15	3.62	3.45	4.01	3.75	2.93	2.68	3.41
Hour 16	3.19	3.18	3.88	4.23	2.81	2.76	3.34
Hour 17	3.22	3.24	4.21	4.18	2.82	2.81	3.41
Hour 18	2.82	3.11	3.91	4.49	3.48	3.32	3.52
Hour 19	2.99	3.32	3.18	3.41	3.46	3.38	3.29
Hour 20	4.22	3.31	3.92	3.01	4.69	3.50	3.77
Hour 21	2.95	3.01	3.09	3.15	3.67	2.65	3.09
Hour 22	2.93	2.64	3.20	3.04	2.85	2.59	2.88
Hour 23	3.05	3.17	3.22	3.19	3.29	2.85	3.13
Hour 24	3.18	3.16	3.40	2.84	3.18	2.82	3.09
Average	3.17	3.41	3.26	3.23	3.54	2.80	3.24

Table E.12. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - ME; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.99	4.12	3.75	5.06	8.59	8.02	5.92
Hour 2	4.61	4.89	3.15	4.06	8.89	7.17	5.46
Hour 3	5.34	5.23	3.48	4.63	9.21	7.34	5.87
Hour 4	6.21	3.32	5.01	5.34	6.85	7.47	5.70
Hour 5	5.26	4.43	4.58	5.91	6.61	7.45	5.71
Hour 6	5.05	5.58	5.00	3.70	6.47	5.23	5.17
Hour 7	3.71	3.78	5.44	2.24	7.47	4.80	4.57
Hour 8	5.99	9.56	7.29	2.91	6.50	4.03	6.05
Hour 9	4.67	3.23	6.29	2.80	6.01	5.05	4.67
Hour 10	2.94	4.12	7.04	3.06	7.14	4.99	4.88
Hour 11	2.10	3.98	6.35	3.21	7.72	5.12	4.75
Hour 12	2.48	4.04	6.18	4.48	7.30	4.23	4.79
Hour 13	2.62	3.70	7.82	5.58	6.36	2.89	4.83
Hour 14	3.20	2.20	8.53	6.78	5.54	4.07	5.05
Hour 15	4.16	2.81	9.39	7.01	5.33	3.74	5.41
Hour 16	3.04	3.13	9.00	9.02	4.99	3.77	5.49
Hour 17	3.45	3.41	9.64	7.56	3.66	3.94	5.28
Hour 18	3.24	4.11	10.02	8.63	4.47	5.02	5.92
Hour 19	2.39	3.76	8.11	5.76	5.32	4.13	4.91
Hour 20	3.66	2.02	8.07	4.33	6.89	3.41	4.73
Hour 21	4.05	1.76	5.70	3.56	5.66	4.11	4.14
Hour 22	3.28	1.83	6.09	3.47	3.85	3.62	3.69
Hour 23	3.67	3.20	5.41	2.81	5.39	3.49	3.99
Hour 24	4.89	2.90	6.16	3.07	3.93	4.12	4.18
Average	4.00	3.80	6.56	4.79	6.26	4.88	5.05

Table E.13. Hourly mean absolute percent error (MAPE); Load zone - NH; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.75	3.79	4.33	5.84	3.67	5.65	4.34
Hour 2	2.91	3.48	3.00	4.91	2.89	5.61	3.80
Hour 3	3.13	3.31	3.16	5.25	2.94	4.97	3.79
Hour 4	3.51	3.08	3.12	5.29	2.77	4.77	3.76
Hour 5	3.17	3.09	2.98	4.72	2.60	4.09	3.44
Hour 6	2.60	2.40	2.79	3.26	2.30	3.05	2.73
Hour 7	2.43	2.92	2.90	3.54	2.58	3.14	2.92
Hour 8	1.90	3.69	2.25	3.31	2.49	3.05	2.78
Hour 9	1.43	2.81	2.33	3.16	2.04	2.00	2.29
Hour 10	2.11	2.51	2.79	2.33	2.73	2.09	2.43
Hour 11	2.33	2.59	2.95	2.18	3.01	2.44	2.58
Hour 12	3.13	2.72	2.74	2.97	3.35	2.13	2.84
Hour 13	2.74	3.25	3.75	2.97	3.38	2.34	3.07
Hour 14	3.27	3.59	2.91	2.86	3.15	3.06	3.14
Hour 15	4.63	4.43	3.22	3.29	3.91	2.64	3.69
Hour 16	5.43	5.11	3.00	2.96	7.31	3.01	4.47
Hour 17	3.17	4.11	3.25	2.67	4.70	2.78	3.45
Hour 18	3.64	3.79	3.46	2.86	4.27	2.27	3.38
Hour 19	3.20	3.30	4.23	3.89	3.87	3.12	3.60
Hour 20	3.99	4.30	3.75	3.77	4.24	3.24	3.88
Hour 21	3.59	3.22	3.77	3.84	3.44	2.50	3.39
Hour 22	2.11	2.27	2.52	2.56	2.40	2.53	2.40
Hour 23	4.01	2.85	3.00	3.14	2.46	3.13	3.10
Hour 24	3.61	2.72	2.80	2.73	2.89	3.31	3.01
Average	3.12	3.31	3.13	3.51	3.31	3.20	3.26

Table E.14. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - NH; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.70	1.96	3.97	6.09	4.80	10.95	5.58
Hour 2	6.75	2.25	2.20	5.64	2.81	12.17	5.30
Hour 3	7.04	3.14	2.35	7.16	3.26	12.70	5.94
Hour 4	7.40	3.00	2.15	7.01	1.77	11.27	5.43
Hour 5	7.43	2.86	2.39	6.81	3.73	12.17	5.90
Hour 6	5.14	2.85	5.61	4.46	3.78	10.34	5.36
Hour 7	5.51	4.05	6.07	6.69	5.16	9.64	6.19
Hour 8	2.41	5.17	4.64	6.44	4.93	8.31	5.32
Hour 9	1.76	4.86	3.39	4.74	2.42	3.27	3.41
Hour 10	1.52	3.99	2.44	2.77	3.77	1.61	2.68
Hour 11	1.25	3.70	1.74	2.71	2.96	4.79	2.86
Hour 12	2.21	3.20	2.47	3.43	2.89	3.73	2.99
Hour 13	1.19	4.97	6.45	7.24	1.66	2.94	4.07
Hour 14	3.66	4.06	4.08	5.25	1.77	3.41	3.71
Hour 15	3.70	5.79	5.09	6.30	4.08	6.20	5.19
Hour 16	2.00	4.73	2.06	5.56	3.46	3.06	3.48
Hour 17	4.95	4.81	2.21	4.05	7.66	2.86	4.42
Hour 18	5.13	2.54	2.87	3.45	6.06	0.97	3.50
Hour 19	3.40	2.11	4.23	2.13	4.16	3.19	3.20
Hour 20	1.84	3.58	3.88	4.07	3.72	0.96	3.01
Hour 21	2.15	3.44	3.52	2.88	2.06	1.71	2.63
Hour 22	4.10	1.89	2.45	1.78	1.39	2.11	2.29
Hour 23	4.31	2.88	3.00	2.30	1.92	3.51	2.99
Hour 24	4.51	2.76	2.75	2.87	2.04	3.73	3.11
Average	3.96	3.52	3.42	4.66	3.43	5.65	4.11

Table E.15. Hourly mean absolute percent error (MAPE); Load zone - VT; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.56	4.22	3.76	3.71	5.12	4.50	3.98
Hour 2	2.38	4.17	3.57	4.80	4.78	4.68	4.07
Hour 3	2.42	4.15	3.46	4.67	4.00	3.35	3.67
Hour 4	2.17	3.76	3.14	3.53	3.66	2.90	3.19
Hour 5	1.93	3.32	3.04	3.06	3.36	2.58	2.88
Hour 6	2.45	3.29	3.50	2.94	2.82	3.38	3.06
Hour 7	2.22	2.61	3.41	2.14	3.19	2.12	2.62
Hour 8	3.30	2.50	2.32	1.99	2.57	1.96	2.44
Hour 9	1.57	1.86	2.79	2.01	2.58	1.73	2.09
Hour 10	1.92	2.20	2.69	2.02	1.99	1.96	2.13
Hour 11	1.79	2.34	2.68	2.34	1.89	1.82	2.14
Hour 12	2.29	2.27	2.59	2.27	2.44	1.93	2.30
Hour 13	2.61	2.47	2.86	2.57	2.66	2.87	2.67
Hour 14	2.89	2.82	2.67	2.39	2.40	2.13	2.55
Hour 15	2.49	2.84	2.65	2.91	2.90	2.29	2.68
Hour 16	2.72	2.93	3.19	3.17	2.99	2.37	2.90
Hour 17	2.48	3.22	3.19	3.23	3.11	2.49	2.95
Hour 18	2.05	2.92	3.37	3.93	3.14	2.72	3.02
Hour 19	2.78	3.01	3.39	3.13	3.10	2.34	2.96
Hour 20	3.21	3.32	3.74	3.34	3.70	3.52	3.47
Hour 21	2.70	3.33	3.41	2.82	3.29	2.59	3.02
Hour 22	2.59	3.72	3.09	3.02	3.35	2.44	3.04
Hour 23	2.41	3.25	3.43	2.84	3.18	2.00	2.85
Hour 24	2.31	3.13	3.51	2.95	3.59	2.34	2.97
Average	2.43	3.07	3.14	2.99	3.16	2.62	2.90

Table E.16. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - VT; Variable selection method - step

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.59	5.33	5.21	3.47	10.12	10.14	6.14
Hour 2	2.24	5.44	3.77	3.70	9.68	10.52	5.89
Hour 3	1.56	4.59	3.68	2.69	10.25	8.79	5.26
Hour 4	1.56	4.81	4.67	2.97	9.45	8.73	5.36
Hour 5	1.13	4.61	4.55	2.83	8.70	7.80	4.94
Hour 6	5.76	4.02	8.75	5.95	7.65	9.70	6.97
Hour 7	2.39	3.91	6.09	4.50	6.80	2.63	4.39
Hour 8	3.29	1.47	2.72	2.08	6.06	4.71	3.39
Hour 9	3.57	1.81	6.08	2.93	6.25	3.63	4.04
Hour 10	1.78	1.23	6.52	1.55	4.19	3.31	3.10
Hour 11	1.81	3.26	6.14	1.66	3.03	1.99	2.98
Hour 12	2.21	1.74	5.68	3.82	3.17	1.95	3.10
Hour 13	2.49	3.59	4.74	4.95	5.58	1.78	3.86
Hour 14	2.43	4.26	7.37	3.68	2.71	2.19	3.77
Hour 15	3.78	3.23	4.73	4.90	6.08	2.06	4.13
Hour 16	3.85	2.61	4.98	4.30	5.81	2.53	4.01
Hour 17	2.12	3.76	7.07	5.21	6.85	2.96	4.66
Hour 18	1.49	1.92	3.31	2.68	3.56	2.29	2.54
Hour 19	1.34	2.03	4.88	1.72	4.79	1.53	2.71
Hour 20	1.77	3.73	5.53	4.56	4.67	3.35	3.94
Hour 21	2.38	3.66	5.77	4.21	5.61	3.15	4.13
Hour 22	3.91	2.63	3.74	2.95	5.47	2.81	3.58
Hour 23	3.12	2.85	4.97	2.30	7.87	3.11	4.04
Hour 24	2.34	2.64	5.78	3.89	7.33	2.86	4.14
Average	2.54	3.30	5.28	3.48	6.32	4.35	4.21

APPENDIX F

REGRESSION RESULTS FOR GENETIC ALGORITHM VARIABLE SELECTION METHOD

Table F.1. Hourly mean absolute percent error (MAPE); Load zone - NEMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.52	3.11	2.91	3.65	2.64	2.94	2.96
Hour 2	2.26	2.69	2.49	3.39	2.54	2.97	2.72
Hour 3	2.59	2.87	3.03	3.72	2.57	3.26	3.01
Hour 4	2.25	3.03	2.46	3.66	2.45	2.61	2.74
Hour 5	2.46	2.81	2.60	3.71	2.67	3.14	2.90
Hour 6	1.81	2.97	2.89	4.06	2.77	2.93	2.91
Hour 7	2.43	2.00	3.02	3.11	2.31	2.73	2.60
Hour 8	2.02	1.93	2.59	2.44	2.13	2.36	2.25
Hour 9	1.65	1.84	2.40	2.17	2.20	2.26	2.09
Hour 10	1.80	1.75	2.09	2.19	2.22	2.06	2.02
Hour 11	1.71	2.09	2.14	2.44	2.57	2.20	2.19
Hour 12	1.69	2.20	2.12	2.43	2.67	2.68	2.30
Hour 13	2.09	2.08	2.30	2.52	2.58	2.32	2.32
Hour 14	1.99	2.92	2.35	3.07	2.89	2.43	2.61
Hour 15	3.06	2.85	2.67	3.09	3.11	3.13	2.99
Hour 16	2.65	2.67	3.34	2.91	3.39	3.01	2.99
Hour 17	4.06	3.00	3.82	3.58	3.49	3.35	3.55
Hour 18	4.16	3.28	3.63	4.18	3.33	3.98	3.76
Hour 19	3.08	3.68	2.99	4.94	3.18	3.71	3.60
Hour 20	3.59	3.71	3.22	4.61	3.57	3.92	3.77
Hour 21	3.00	3.32	3.06	3.18	2.75	3.26	3.10
Hour 22	2.86	2.98	2.89	2.97	3.07	3.12	2.98
Hour 23	2.79	2.56	2.84	2.57	2.76	2.91	2.74
Hour 24	2.74	2.36	2.56	2.30	2.70	2.87	2.59
Average	2.55	2.70	2.77	3.20	2.77	2.92	2.82

Table F.2. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - NEMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.17	3.09	2.60	3.06	2.18	5.85	3.66
Hour 2	3.57	2.12	2.56	2.10	2.61	5.69	3.11
Hour 3	5.97	2.66	1.61	3.39	5.33	8.04	4.50
Hour 4	3.97	2.09	1.90	3.71	5.30	7.69	4.11
Hour 5	5.30	2.46	3.00	2.89	1.76	8.59	4.00
Hour 6	5.18	1.83	3.91	3.99	4.18	8.99	4.68
Hour 7	5.91	0.57	3.84	3.76	3.48	7.51	4.18
Hour 8	2.93	1.71	2.02	3.60	4.03	5.49	3.30
Hour 9	3.32	1.09	3.71	2.54	4.54	5.02	3.37
Hour 10	3.14	1.25	2.80	2.92	3.01	4.61	2.96
Hour 11	1.08	2.16	3.62	2.04	2.79	2.23	2.32
Hour 12	1.45	1.74	3.32	2.52	2.47	2.41	2.32
Hour 13	2.15	2.02	3.72	3.85	3.58	2.24	2.93
Hour 14	2.38	2.39	3.01	4.43	2.28	2.18	2.78
Hour 15	1.90	2.33	1.68	2.75	1.83	1.35	1.97
Hour 16	3.36	1.53	2.18	2.13	1.59	3.10	2.31
Hour 17	4.46	1.19	5.36	5.98	1.15	2.00	3.36
Hour 18	2.60	5.75	4.52	8.54	1.79	2.13	4.22
Hour 19	3.18	4.92	4.59	8.23	2.03	3.93	4.48
Hour 20	2.77	3.77	4.20	5.43	2.20	2.98	3.56
Hour 21	4.90	2.77	3.53	3.13	2.93	4.08	3.56
Hour 22	3.59	3.17	3.08	3.50	2.42	1.92	2.95
Hour 23	1.77	4.08	3.81	5.10	2.50	1.46	3.12
Hour 24	1.80	2.48	2.55	3.00	3.86	2.38	2.68
Average	3.41	2.47	3.21	3.86	2.91	4.24	3.35

Table F.3. Hourly mean absolute percent error (MAPE); Load zone - SEMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.56	4.07	3.79	3.56	3.69	3.64	3.72
Hour 2	3.36	4.17	3.72	3.41	3.19	3.02	3.48
Hour 3	3.37	3.48	3.40	3.76	3.25	2.67	3.32
Hour 4	2.82	3.54	3.49	3.18	2.99	2.77	3.13
Hour 5	2.54	3.66	2.93	3.57	2.77	3.34	3.14
Hour 6	2.88	3.70	2.63	3.97	3.10	3.05	3.22
Hour 7	2.96	3.07	4.29	3.75	2.91	3.07	3.34
Hour 8	2.30	2.91	3.30	3.32	2.98	2.91	2.95
Hour 9	2.51	3.51	2.57	2.60	3.25	2.53	2.83
Hour 10	2.59	3.37	2.71	2.76	2.84	2.55	2.80
Hour 11	2.98	3.25	2.87	2.66	3.30	3.04	3.02
Hour 12	2.95	2.87	3.63	3.10	3.17	3.27	3.16
Hour 13	4.15	3.57	3.87	3.33	3.49	3.74	3.69
Hour 14	4.39	3.64	3.96	3.65	3.78	4.16	3.93
Hour 15	4.70	3.48	4.45	3.30	3.70	4.06	3.95
Hour 16	4.33	3.38	4.39	3.71	3.80	3.98	3.93
Hour 17	3.83	3.60	4.21	4.47	3.78	3.81	3.95
Hour 18	4.15	3.39	3.89	3.41	3.79	3.80	3.74
Hour 19	4.05	4.05	4.12	3.79	4.32	3.78	4.02
Hour 20	4.11	3.65	3.99	3.45	3.83	3.71	3.79
Hour 21	3.65	3.72	4.29	3.16	4.03	3.17	3.67
Hour 22	4.25	3.51	4.84	3.19	3.80	3.25	3.81
Hour 23	4.06	4.03	3.65	3.31	3.47	3.77	3.72
Hour 24	3.67	3.44	3.13	3.60	3.45	4.28	3.59
Average	3.51	3.54	3.67	3.42	3.44	3.39	3.50

Table F.4. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - SEMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.93	3.72	3.79	3.40	3.31	6.90	4.51
Hour 2	4.29	4.14	2.98	5.24	2.83	3.91	3.90
Hour 3	5.62	5.19	1.73	4.97	1.42	5.71	4.11
Hour 4	4.25	2.79	3.75	4.19	5.19	6.57	4.46
Hour 5	6.48	2.12	2.78	3.71	4.61	10.43	5.02
Hour 6	4.66	3.15	2.21	4.77	6.39	9.85	5.17
Hour 7	7.12	3.86	2.36	4.11	4.64	9.81	5.32
Hour 8	5.88	1.06	2.11	3.76	4.75	10.36	4.65
Hour 9	3.71	1.89	1.74	3.22	5.63	8.42	4.10
Hour 10	3.00	2.34	1.04	3.30	4.14	7.48	3.55
Hour 11	2.97	3.74	2.88	3.76	5.13	6.36	4.14
Hour 12	5.43	4.32	2.89	4.68	2.81	3.27	3.90
Hour 13	2.81	5.81	3.59	6.33	2.90	6.28	4.62
Hour 14	2.40	3.43	1.90	4.09	2.91	5.15	3.31
Hour 15	3.98	4.43	4.46	2.30	2.72	3.75	3.61
Hour 16	3.48	6.29	3.42	4.93	4.88	4.84	4.64
Hour 17	4.64	4.51	2.87	5.21	3.94	6.60	4.63
Hour 18	4.83	3.17	3.45	3.69	4.00	5.88	4.17
Hour 19	7.68	2.48	3.91	3.91	5.48	9.24	5.45
Hour 20	7.39	3.61	4.88	4.64	4.27	7.86	5.44
Hour 21	8.82	3.31	4.40	4.18	3.69	9.95	5.72
Hour 22	7.13	3.12	5.39	4.67	2.78	9.35	5.41
Hour 23	5.86	2.72	3.06	4.57	2.24	7.01	4.24
Hour 24	3.33	4.06	1.99	4.51	2.32	4.78	3.50
Average	5.07	3.55	3.07	4.26	3.87	7.07	4.48

Table F.5. Hourly mean absolute percent error (MAPE); Load zone - WCMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.43	4.39	4.61	5.17	4.86	4.25	4.45
Hour 2	3.11	4.85	4.06	4.63	4.48	3.97	4.18
Hour 3	2.83	5.32	3.63	4.37	4.31	3.91	4.06
Hour 4	2.50	5.00	3.26	4.32	4.95	3.51	3.92
Hour 5	2.85	4.17	3.71	3.18	4.23	2.79	3.49
Hour 6	2.28	3.50	2.65	3.20	4.00	2.79	3.07
Hour 7	2.69	2.97	2.39	3.20	3.30	2.76	2.88
Hour 8	2.51	2.86	2.29	2.43	3.33	2.48	2.65
Hour 9	1.92	2.13	2.03	2.01	3.38	1.92	2.23
Hour 10	1.69	1.96	2.13	1.86	2.49	2.17	2.05
Hour 11	1.87	1.86	2.48	2.33	2.82	1.78	2.19
Hour 12	2.37	2.11	3.09	2.05	2.58	2.05	2.37
Hour 13	2.30	2.17	2.93	2.54	2.79	2.10	2.47
Hour 14	3.21	2.45	3.20	2.32	3.14	2.64	2.83
Hour 15	2.92	2.58	3.37	2.52	3.38	2.37	2.86
Hour 16	2.80	2.65	3.27	2.51	3.25	2.72	2.87
Hour 17	3.10	2.49	3.23	2.64	3.26	3.37	3.01
Hour 18	3.23	2.67	3.57	3.07	3.26	2.94	3.12
Hour 19	3.02	2.60	3.61	3.21	3.27	3.06	3.13
Hour 20	3.50	2.84	3.63	2.84	3.64	3.32	3.30
Hour 21	2.91	2.79	3.15	2.71	3.20	2.68	2.91
Hour 22	2.83	2.77	3.36	2.80	3.43	2.89	3.01
Hour 23	2.48	2.79	3.39	2.32	3.34	2.74	2.84
Hour 24	2.44	2.73	3.17	2.08	3.30	2.47	2.70
Average	2.70	3.03	3.17	2.93	3.50	2.82	3.03

Table F.6. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - WCMA; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.37	3.33	4.57	2.69	6.99	7.79	5.12
Hour 2	4.59	2.90	3.70	3.12	7.24	6.71	4.71
Hour 3	4.64	2.08	3.74	2.53	5.60	4.96	3.93
Hour 4	3.76	3.55	3.15	4.63	8.21	7.16	5.08
Hour 5	4.79	1.83	3.75	1.53	6.76	6.98	4.27
Hour 6	4.57	1.05	2.50	1.52	7.12	6.64	3.90
Hour 7	3.48	3.08	3.54	5.81	6.61	6.04	4.76
Hour 8	2.46	3.30	3.89	2.27	7.30	7.10	4.39
Hour 9	1.05	1.99	2.45	1.46	6.85	3.84	2.94
Hour 10	1.34	1.35	2.70	1.03	4.60	5.30	2.72
Hour 11	1.59	1.35	2.80	2.59	3.88	2.98	2.53
Hour 12	2.02	1.52	3.18	1.66	3.51	2.61	2.42
Hour 13	2.22	2.47	3.83	0.91	2.64	2.34	2.40
Hour 14	2.90	2.89	5.49	1.35	3.95	2.80	3.23
Hour 15	3.91	2.49	4.92	2.86	4.75	2.43	3.56
Hour 16	2.43	1.66	3.86	1.78	2.90	2.60	2.54
Hour 17	4.12	3.05	4.59	2.61	3.47	3.28	3.52
Hour 18	5.20	2.57	4.75	3.03	3.72	4.85	4.02
Hour 19	3.84	1.27	6.12	5.25	4.85	5.48	4.47
Hour 20	2.47	3.11	4.77	3.35	5.50	5.56	4.13
Hour 21	3.49	2.95	4.35	2.98	4.77	5.28	3.97
Hour 22	3.69	2.19	4.70	2.69	6.00	6.33	4.27
Hour 23	3.73	2.73	4.35	2.15	5.79	5.66	4.07
Hour 24	3.59	1.71	4.51	1.76	6.25	5.81	3.94
Average	3.39	2.35	4.01	2.56	5.39	5.02	3.79

Table F.7. Hourly mean absolute percent error (MAPE); Load zone - CT; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.41	5.05	2.67	4.89	5.20	3.89	4.19
Hour 2	3.60	5.16	3.40	3.25	4.26	4.44	4.02
Hour 3	3.53	4.73	3.19	4.89	4.25	3.60	4.03
Hour 4	2.76	4.22	2.80	5.02	4.59	3.28	3.78
Hour 5	2.75	5.56	2.81	4.16	5.24	3.44	3.99
Hour 6	2.76	4.60	2.68	4.04	4.19	4.14	3.73
Hour 7	2.58	4.54	2.98	3.44	4.61	3.21	3.56
Hour 8	2.55	3.53	3.15	3.47	3.54	3.05	3.22
Hour 9	2.72	2.40	3.26	1.53	3.05	2.10	2.51
Hour 10	2.53	2.21	3.28	1.54	3.43	1.88	2.48
Hour 11	2.29	2.05	2.69	1.88	3.37	2.21	2.41
Hour 12	2.82	2.29	2.76	2.13	2.73	2.56	2.55
Hour 13	3.74	2.99	4.18	2.25	3.57	2.83	3.26
Hour 14	3.81	4.79	4.36	3.59	4.84	3.04	4.07
Hour 15	3.49	5.06	4.82	3.53	4.59	3.10	4.10
Hour 16	5.10	3.53	4.56	3.69	4.94	3.27	4.18
Hour 17	3.23	4.07	3.91	3.61	5.47	3.52	3.97
Hour 18	3.39	4.23	4.01	3.44	4.59	3.30	3.83
Hour 19	3.06	3.79	4.18	3.71	3.92	2.91	3.59
Hour 20	3.38	4.17	4.59	3.78	4.63	3.25	3.97
Hour 21	2.77	3.99	3.59	3.18	4.19	3.35	3.51
Hour 22	3.29	3.45	3.89	2.91	4.12	3.10	3.46
Hour 23	3.22	3.46	3.39	2.61	3.72	3.04	3.24
Hour 24	2.71	3.30	3.61	2.57	3.69	3.30	3.20
Average	3.14	3.88	3.53	3.30	4.20	3.16	3.53

Table F.8. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - CT; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	4.71	4.81	4.59	7.32	9.02	8.60	6.51
Hour 2	5.27	5.40	3.31	6.87	6.58	8.90	6.05
Hour 3	4.74	3.42	3.62	7.44	7.11	7.95	5.72
Hour 4	6.14	3.13	3.35	7.68	7.56	7.78	5.94
Hour 5	4.83	5.39	3.29	4.47	8.89	6.73	5.60
Hour 6	5.20	3.68	3.13	3.87	7.26	9.40	5.42
Hour 7	3.45	5.25	5.53	4.79	9.93	6.25	5.87
Hour 8	4.32	4.78	3.40	4.53	6.77	8.05	5.31
Hour 9	3.16	2.98	3.64	2.04	4.35	3.31	3.25
Hour 10	2.67	3.06	3.59	2.17	4.93	3.57	3.33
Hour 11	2.13	3.08	2.82	1.89	5.12	3.48	3.09
Hour 12	1.49	2.43	3.49	2.08	4.01	1.39	2.48
Hour 13	4.22	4.42	6.14	3.02	3.66	2.51	4.00
Hour 14	4.45	5.85	5.44	3.67	7.05	3.37	4.97
Hour 15	3.83	6.49	9.48	2.19	8.68	4.60	5.88
Hour 16	4.66	3.81	4.68	2.11	6.30	3.65	4.20
Hour 17	2.34	3.18	3.54	2.58	6.69	4.35	3.78
Hour 18	2.66	5.72	5.67	5.04	7.52	4.98	5.27
Hour 19	3.37	5.06	6.03	4.56	4.72	3.31	4.51
Hour 20	4.04	4.77	7.56	3.72	7.33	4.93	5.39
Hour 21	4.25	4.51	5.28	2.96	6.75	5.14	4.81
Hour 22	5.02	5.53	5.44	3.89	6.72	5.39	5.33
Hour 23	6.00	3.59	5.36	3.33	6.68	6.41	5.23
Hour 24	6.48	3.24	5.99	3.29	6.16	6.55	5.28
Average	4.14	4.32	4.77	3.98	6.66	5.44	4.88

Table F.9. Hourly mean absolute percent error (MAPE); Load zone - RI; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.51	3.64	4.07	3.40	3.89	3.59	3.68
Hour 2	3.65	3.60	3.73	3.63	3.90	3.34	3.64
Hour 3	3.42	2.92	3.92	3.80	3.09	3.51	3.44
Hour 4	3.40	3.01	3.16	3.78	3.36	3.58	3.38
Hour 5	3.45	3.64	3.28	3.61	3.54	3.58	3.52
Hour 6	3.08	3.25	3.68	3.64	3.20	3.39	3.37
Hour 7	3.08	2.65	3.90	3.52	3.20	3.13	3.24
Hour 8	2.89	3.02	3.20	3.28	2.87	3.24	3.08
Hour 9	2.45	2.77	3.34	2.52	3.34	2.92	2.89
Hour 10	2.79	2.60	3.03	2.67	3.43	3.18	2.95
Hour 11	3.41	3.01	3.47	3.05	3.58	3.86	3.40
Hour 12	4.14	3.24	3.43	3.53	3.28	4.07	3.61
Hour 13	3.88	3.21	4.06	3.74	3.45	4.12	3.75
Hour 14	4.98	3.89	4.69	4.37	3.63	4.58	4.36
Hour 15	5.05	4.37	4.23	3.49	4.31	4.53	4.33
Hour 16	5.22	4.05	5.03	4.04	4.27	4.51	4.52
Hour 17	4.68	4.26	4.67	3.45	4.26	3.97	4.21
Hour 18	4.49	3.61	4.55	3.54	3.98	3.89	4.01
Hour 19	4.55	4.14	4.23	3.80	4.45	4.31	4.25
Hour 20	3.99	4.24	4.38	4.01	4.31	3.95	4.15
Hour 21	3.76	3.51	3.89	3.72	3.89	3.71	3.75
Hour 22	3.66	3.27	3.94	3.56	4.14	3.52	3.68
Hour 23	3.56	3.15	3.85	2.81	3.77	3.34	3.41
Hour 24	3.48	2.73	3.76	3.23	3.72	3.21	3.36
Average	3.77	3.41	3.90	3.51	3.70	3.71	3.67

Table F.10. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - RI; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	5.24	4.22	4.80	4.54	5.13	6.27	5.03
Hour 2	5.91	3.20	4.52	3.84	5.40	6.71	4.93
Hour 3	5.54	4.62	4.50	1.45	4.68	7.59	4.73
Hour 4	5.04	3.58	2.39	2.15	4.74	6.28	4.03
Hour 5	5.29	2.76	3.37	2.99	4.32	7.61	4.39
Hour 6	6.75	2.28	2.74	2.55	5.11	9.14	4.76
Hour 7	9.68	5.39	3.31	3.39	5.42	8.08	5.88
Hour 8	3.43	3.29	2.20	5.28	3.14	7.18	4.09
Hour 9	4.43	3.34	1.94	3.36	3.59	6.65	3.88
Hour 10	3.45	4.45	2.65	4.50	2.51	6.20	3.96
Hour 11	4.09	4.44	3.70	4.66	4.38	8.39	4.94
Hour 12	2.09	2.82	3.73	4.09	2.88	5.44	3.51
Hour 13	4.29	5.22	3.47	6.04	2.82	5.02	4.48
Hour 14	4.18	3.19	5.76	4.89	4.77	5.04	4.64
Hour 15	3.27	2.32	6.74	4.50	6.97	4.97	4.80
Hour 16	3.94	4.10	5.16	3.47	4.81	5.02	4.42
Hour 17	3.62	4.93	3.78	3.31	3.22	3.28	3.69
Hour 18	4.49	2.85	2.89	1.88	3.36	5.96	3.57
Hour 19	7.82	3.03	2.24	3.44	3.05	8.68	4.71
Hour 20	9.42	2.76	3.39	4.27	3.91	10.02	5.63
Hour 21	7.08	3.46	6.30	3.56	5.69	9.53	5.94
Hour 22	6.60	3.76	6.15	3.68	6.26	9.40	5.98
Hour 23	6.43	3.58	2.91	2.90	3.39	6.78	4.33
Hour 24	3.97	4.31	2.33	3.10	2.45	5.34	3.58
Average	5.25	3.66	3.79	3.66	4.25	6.86	4.58

Table F.11. Hourly mean absolute percent error (MAPE); Load zone - ME; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.92	2.56	2.56	3.06	2.38	3.38	2.81
Hour 2	2.82	3.12	2.41	3.34	2.95	2.91	2.92
Hour 3	3.05	2.90	2.70	2.89	2.68	2.86	2.85
Hour 4	2.70	2.89	2.31	2.71	2.76	2.70	2.68
Hour 5	2.42	2.73	1.86	2.63	2.68	2.59	2.49
Hour 6	2.27	2.78	1.82	2.49	2.64	2.40	2.40
Hour 7	2.11	2.69	2.26	2.46	2.82	2.26	2.43
Hour 8	2.56	2.41	2.38	2.53	2.37	2.31	2.43
Hour 9	2.26	2.49	2.17	2.45	2.52	2.43	2.39
Hour 10	2.20	2.41	2.29	2.08	2.62	2.31	2.32
Hour 11	2.51	2.24	2.67	2.65	2.32	2.76	2.52
Hour 12	2.47	2.76	2.37	2.25	2.75	2.39	2.50
Hour 13	2.38	2.33	2.82	2.70	2.07	2.33	2.44
Hour 14	2.42	2.56	2.91	2.50	2.68	2.56	2.61
Hour 15	2.32	2.44	2.73	2.43	2.41	2.47	2.47
Hour 16	2.74	2.69	2.52	2.68	2.26	2.43	2.55
Hour 17	2.53	2.79	2.38	2.68	2.45	2.61	2.58
Hour 18	2.43	2.69	2.62	2.77	2.61	2.87	2.67
Hour 19	2.70	2.68	2.49	2.40	2.46	2.74	2.58
Hour 20	3.26	3.12	3.09	3.42	2.59	3.31	3.13
Hour 21	2.58	2.76	2.20	3.20	2.19	2.56	2.58
Hour 22	2.47	2.42	2.66	2.34	2.37	2.49	2.46
Hour 23	2.78	2.27	2.53	2.34	2.35	2.77	2.50
Hour 24	2.65	2.34	2.70	2.39	2.69	2.57	2.56
Average	2.56	2.63	2.48	2.64	2.53	2.63	2.58

Table F.12. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - ME; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	6.12	1.49	4.20	2.26	4.83	8.38	4.55
Hour 2	5.83	3.34	3.99	2.55	7.00	7.28	5.00
Hour 3	5.73	2.03	5.31	2.21	6.30	7.28	4.81
Hour 4	5.51	2.74	4.29	2.33	6.46	7.04	4.73
Hour 5	5.39	2.52	4.73	3.20	6.81	6.60	4.88
Hour 6	5.52	3.64	3.23	3.40	6.28	6.55	4.77
Hour 7	5.32	4.23	3.79	1.35	5.22	6.30	4.37
Hour 8	5.03	1.70	6.07	1.80	5.93	5.94	4.41
Hour 9	5.17	2.91	5.47	1.72	7.05	6.43	4.79
Hour 10	3.76	1.55	5.06	1.74	4.61	5.74	3.74
Hour 11	5.28	1.68	4.65	2.52	4.70	6.53	4.23
Hour 12	2.44	2.90	5.14	2.36	5.86	4.51	3.87
Hour 13	2.70	2.42	5.96	2.92	3.85	3.52	3.56
Hour 14	2.40	2.41	7.90	4.20	4.92	4.16	4.33
Hour 15	2.10	1.78	5.62	3.82	4.16	3.35	3.47
Hour 16	1.91	1.83	5.39	3.88	3.79	3.37	3.36
Hour 17	3.74	1.61	4.55	2.15	3.57	4.28	3.32
Hour 18	3.26	2.92	5.40	5.69	3.36	4.13	4.13
Hour 19	3.39	1.91	4.61	1.90	4.08	3.36	3.21
Hour 20	3.92	2.63	2.86	3.49	3.49	4.37	3.46
Hour 21	3.18	1.83	3.87	4.11	4.13	3.42	3.42
Hour 22	3.41	1.03	4.77	1.93	4.25	2.65	3.01
Hour 23	4.09	2.49	5.51	3.44	4.44	3.81	3.96
Hour 24	3.94	2.59	4.85	3.50	4.26	3.06	3.70
Average	4.13	2.34	4.88	2.85	4.97	5.09	4.04

Table F.13. Hourly mean absolute percent error (MAPE); Load zone - NH; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.92	4.00	3.32	4.29	3.56	2.91	3.50
Hour 2	2.57	3.62	3.75	4.00	3.34	3.37	3.44
Hour 3	2.37	3.51	3.29	4.58	2.89	3.35	3.33
Hour 4	2.42	3.00	2.88	4.00	2.83	3.11	3.04
Hour 5	2.28	2.61	2.52	3.26	2.47	2.54	2.61
Hour 6	1.88	2.85	2.47	3.40	2.16	3.26	2.67
Hour 7	2.76	2.10	2.81	3.19	2.21	2.84	2.65
Hour 8	2.07	2.14	2.07	1.83	2.01	1.80	1.98
Hour 9	1.31	1.90	1.82	1.78	1.87	1.55	1.71
Hour 10	1.57	2.04	2.18	2.28	2.38	1.80	2.04
Hour 11	1.71	2.32	2.41	2.00	2.43	1.56	2.07
Hour 12	1.70	2.13	2.47	2.31	2.59	1.49	2.11
Hour 13	2.02	2.40	2.63	2.37	2.58	1.62	2.27
Hour 14	2.04	2.41	2.74	2.23	2.65	2.44	2.42
Hour 15	2.32	2.31	2.80	2.88	2.62	2.39	2.55
Hour 16	2.36	2.69	2.72	3.09	2.75	2.40	2.67
Hour 17	2.83	2.42	2.91	2.78	2.55	2.67	2.69
Hour 18	2.46	2.86	2.43	2.79	2.98	2.83	2.72
Hour 19	2.41	2.77	2.68	2.92	2.97	2.33	2.68
Hour 20	2.89	2.92	2.88	3.21	3.22	2.84	2.99
Hour 21	2.56	3.06	2.84	3.23	2.82	2.59	2.85
Hour 22	2.43	2.31	2.32	2.49	2.29	2.78	2.44
Hour 23	2.39	2.33	2.24	2.46	2.24	2.53	2.36
Hour 24	2.24	2.31	2.42	2.65	2.30	2.40	2.39
Average	2.27	2.63	2.65	2.92	2.61	2.47	2.59

Table F.14. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - NH; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	6.05	1.86	2.40	3.13	5.43	9.09	4.66
Hour 2	6.76	2.34	3.31	1.34	4.63	10.37	4.79
Hour 3	6.87	2.19	4.49	4.48	4.54	10.21	5.46
Hour 4	6.67	1.07	3.01	3.40	5.06	10.13	4.89
Hour 5	6.15	1.65	3.40	2.42	3.24	8.26	4.19
Hour 6	5.16	3.15	3.87	4.52	3.27	9.06	4.84
Hour 7	6.53	1.07	5.43	6.39	1.82	7.42	4.78
Hour 8	2.08	2.17	1.23	1.09	2.95	3.47	2.16
Hour 9	1.46	2.36	1.65	2.36	2.83	3.27	2.32
Hour 10	2.33	2.47	2.23	3.12	3.22	2.38	2.63
Hour 11	1.50	2.79	2.32	2.92	3.56	2.96	2.68
Hour 12	1.51	2.28	2.42	2.40	2.83	1.71	2.19
Hour 13	2.42	2.94	5.16	3.11	4.59	1.88	3.35
Hour 14	2.26	1.85	4.49	3.05	3.09	1.90	2.77
Hour 15	3.11	1.35	3.86	3.79	2.77	2.61	2.91
Hour 16	3.21	1.71	2.78	5.27	2.49	2.94	3.07
Hour 17	2.24	1.82	2.61	3.40	2.35	1.74	2.36
Hour 18	4.70	2.28	1.82	1.77	3.51	4.70	3.13
Hour 19	2.54	2.68	2.91	4.18	2.38	1.41	2.68
Hour 20	2.27	0.72	1.86	3.74	2.79	2.02	2.24
Hour 21	3.09	1.90	1.59	2.15	3.39	1.72	2.31
Hour 22	2.04	2.06	2.00	2.30	2.15	2.33	2.15
Hour 23	3.26	2.34	2.16	1.95	3.29	3.13	2.69
Hour 24	3.50	2.69	3.22	3.63	3.05	4.00	3.35
Average	3.65	2.07	2.93	3.16	3.30	4.53	3.27

Table F.15. Hourly mean absolute percent error (MAPE); Load zone - VT; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	2.73	3.99	3.62	3.89	3.93	2.82	3.49
Hour 2	2.59	4.15	3.48	3.56	4.01	2.51	3.38
Hour 3	2.28	3.75	3.38	3.46	3.87	3.11	3.31
Hour 4	2.32	3.45	3.23	3.34	3.60	2.68	3.10
Hour 5	2.07	3.39	2.93	3.06	3.54	2.21	2.87
Hour 6	2.05	2.87	2.55	2.97	2.86	2.40	2.62
Hour 7	1.96	2.25	2.20	2.58	2.53	2.08	2.27
Hour 8	1.63	1.75	1.76	1.65	1.98	1.67	1.74
Hour 9	1.31	1.61	1.62	1.66	1.67	1.55	1.57
Hour 10	1.35	1.85	1.81	2.01	1.78	1.43	1.71
Hour 11	1.57	1.81	1.88	2.00	1.76	1.54	1.76
Hour 12	1.67	1.78	2.22	2.19	1.95	1.59	1.90
Hour 13	1.98	2.05	2.21	2.21	2.03	1.89	2.06
Hour 14	2.07	2.27	2.24	2.36	2.09	1.80	2.14
Hour 15	1.98	2.42	2.39	2.37	2.36	1.60	2.19
Hour 16	1.85	2.41	2.53	2.51	2.40	2.00	2.28
Hour 17	1.92	2.55	2.44	2.72	2.54	1.93	2.35
Hour 18	1.89	3.05	2.47	2.71	2.73	2.03	2.48
Hour 19	2.39	2.67	2.68	2.63	2.66	2.21	2.54
Hour 20	2.79	2.50	3.04	2.72	3.09	2.75	2.82
Hour 21	2.39	2.73	3.23	2.97	3.19	2.39	2.82
Hour 22	2.17	2.92	2.80	2.69	2.94	2.02	2.59
Hour 23	2.00	2.94	3.07	2.87	3.01	2.06	2.66
Hour 24	2.31	2.90	3.21	2.90	3.23	2.00	2.76
Average	2.05	2.67	2.63	2.67	2.74	2.09	2.48

Table F.16. Mean absolute percent error (MAPE) during top 5 load hours; Load zone - VT; Variable selection method - GA

	Tr - 05, Pr - 06	Tr - 05, Pr - 07	Tr - 06, Pr - 05	Tr - 06, Pr - 07	Tr - 07, Pr - 05	Tr - 07, Pr - 06	Hrly Avg
Hour 1	3.15	3.36	5.09	3.55	10.06	7.77	5.50
Hour 2	2.08	3.65	5.54	4.15	9.82	5.77	5.17
Hour 3	2.51	3.63	4.61	3.08	9.82	8.77	5.40
Hour 4	3.31	3.67	3.43	4.63	8.31	7.91	5.21
Hour 5	1.69	3.54	4.43	3.06	9.23	7.03	4.83
Hour 6	5.45	3.31	3.73	2.92	7.36	7.66	5.07
Hour 7	2.78	3.06	2.12	2.90	5.11	4.10	3.35
Hour 8	1.73	1.50	2.48	2.53	4.06	3.32	2.60
Hour 9	3.02	0.97	2.74	1.49	3.04	3.25	2.42
Hour 10	2.04	1.98	3.05	1.89	2.75	2.19	2.32
Hour 11	1.87	2.87	3.16	2.41	2.99	1.62	2.49
Hour 12	1.73	0.90	4.60	3.20	2.80	1.67	2.48
Hour 13	1.65	1.87	3.48	2.70	3.28	1.64	2.44
Hour 14	1.96	2.53	4.78	4.47	4.71	1.52	3.33
Hour 15	2.08	2.43	5.00	3.83	4.71	2.08	3.36
Hour 16	1.77	2.64	5.11	3.51	5.05	2.06	3.36
Hour 17	1.08	2.53	4.52	3.29	4.90	1.48	2.97
Hour 18	3.31	2.87	2.49	2.72	2.79	2.38	2.76
Hour 19	2.18	2.30	4.19	2.52	3.52	2.84	2.92
Hour 20	2.14	2.18	3.57	2.86	4.62	3.39	3.13
Hour 21	4.13	4.55	2.66	3.73	3.68	4.09	3.81
Hour 22	3.37	2.97	4.29	2.21	5.16	3.11	3.52
Hour 23	2.23	2.54	4.81	2.65	6.40	3.06	3.61
Hour 24	1.99	2.83	4.53	3.68	6.66	3.49	3.86
Average	2.47	2.69	3.93	3.08	5.45	3.84	3.58

APPENDIX G

PLOTS OF β_0

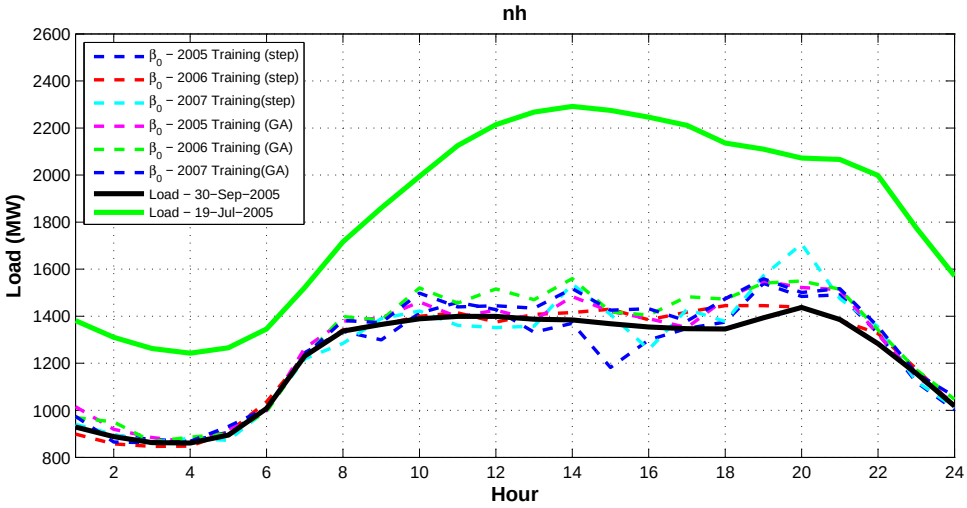

Figure G.1. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the NH load zone.

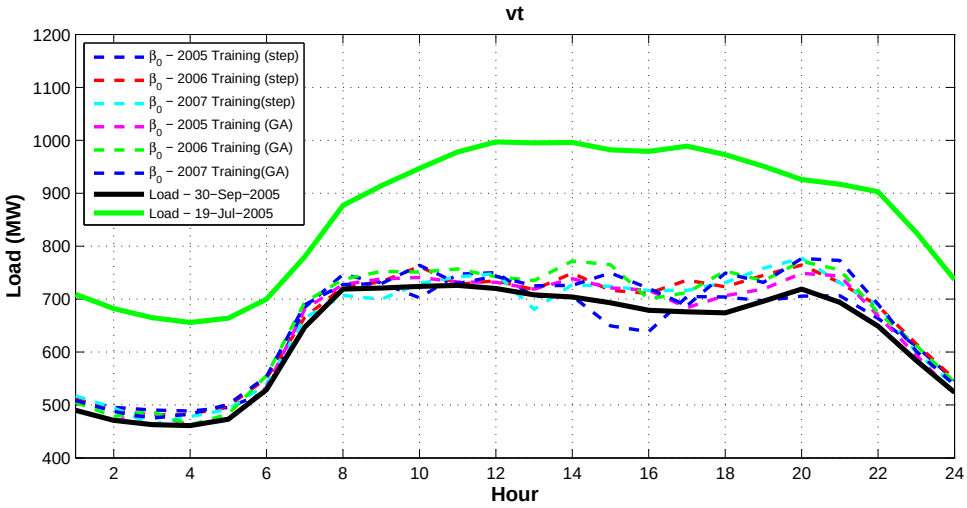

Figure G.2. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the VT load zone.

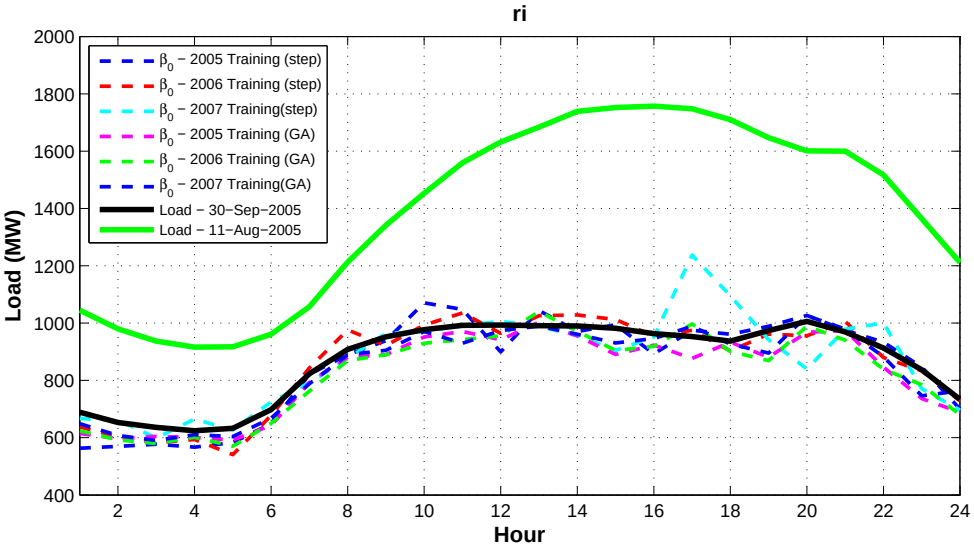

Figure G.3. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the RI load zone.

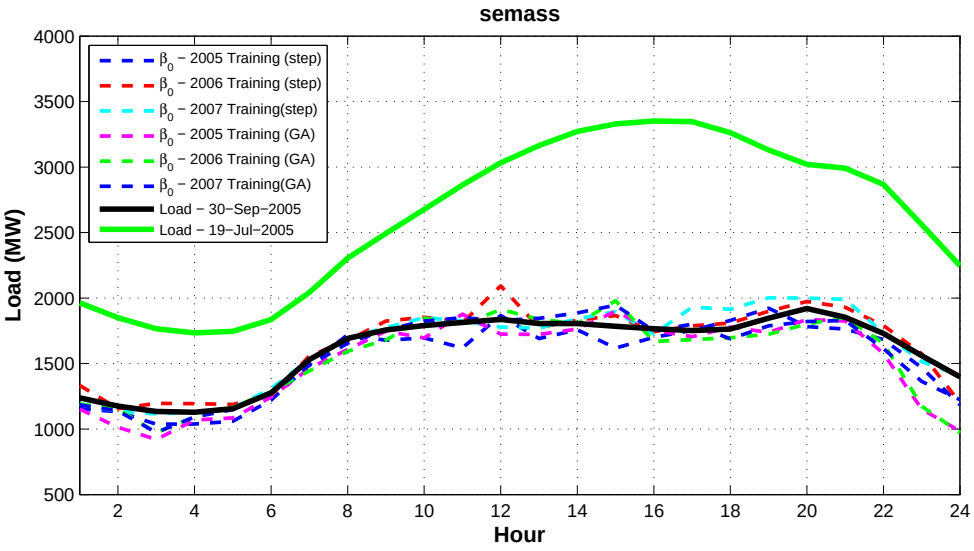

Figure G.4. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the SEMA load zone.

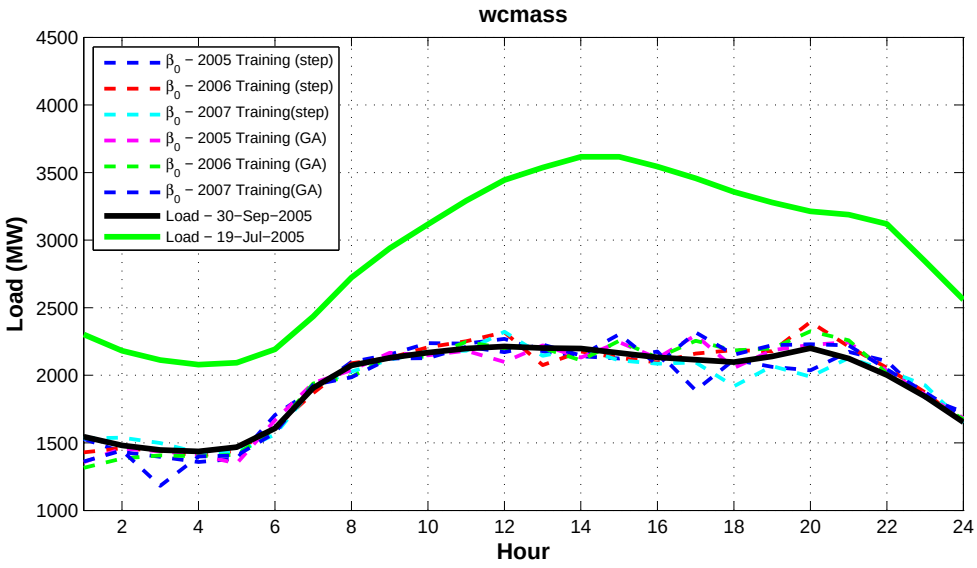

Figure G.5. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the WCMA load zone.

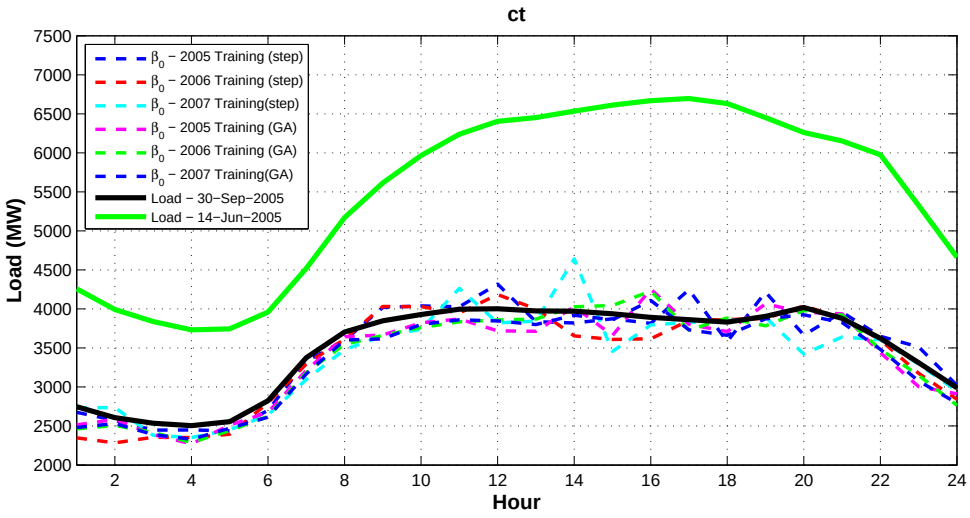

Figure G.6. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the CT load zone.

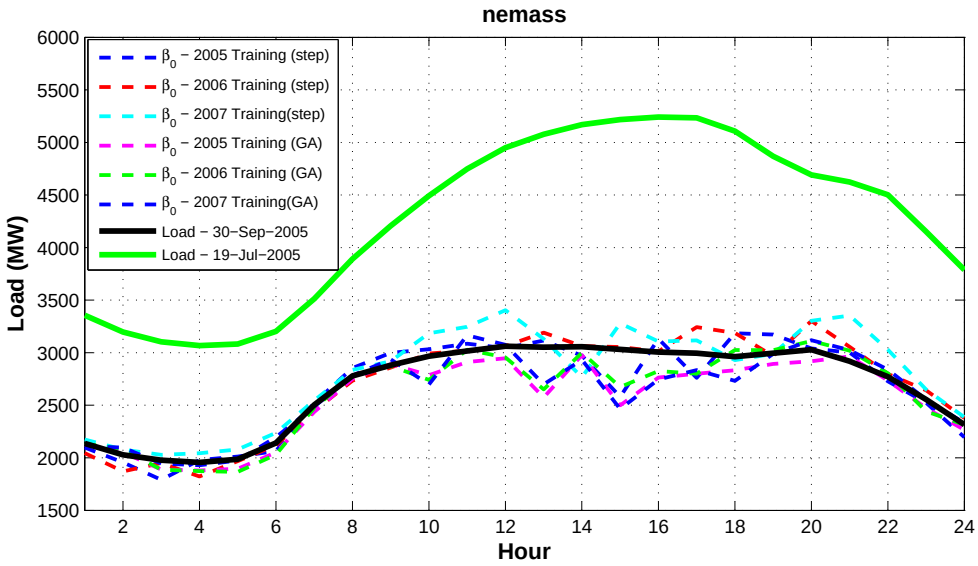

Figure G.7. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the NEMA load zone.

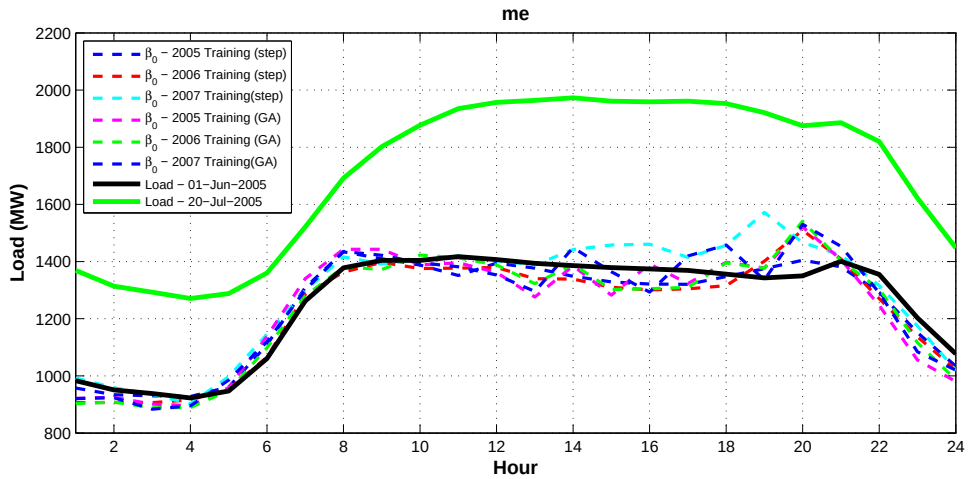

Figure G.8. Plots of β_0 for all training years compared with the 2005 maximum and minimum daily average loads for the ME load zone.

APPENDIX H

MATLAB M-FILES

```
function  [bestVars]= speedyGA(GAzone,GAhr,popSize)
% SpeedyGA is a vectorized implementation of a Simple Genetic ...
    Algorithm in Matlab
% Version 1.3
% Copyright (C) 2007, 2008, 2009  Keki Burjorjee
% Created and tested under Matlab 7 (R14).

%  Licensed under the Apache License, Version 2.0 (the ...
    "License"); you may
%  not use this file except in compliance with the License. You ...
    may obtain
%  a copy of the License at
%
%  http://www.apache.org/licenses/LICENSE-2.0
%
%  Unless required by applicable law or agreed to in writing, ...
    software
%  distributed under the License is distributed on an "AS IS" BASIS,
%  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express ...
    or implied.
%  See the License for the specific language governing ...
    permissions and
%  limitations under the License.

%  Acknowledgement of the author (Keki Burjorjee) is requested, ...
    but not required,
%  in any publication that presents results obtained by using ...
    this script

%  Without Sigma Scaling, Stochastic Universal Sampling, and the ...
    generation of mask
%  repositories, SpeedyGA faithfully implements the specification ...
    of a simple genetic
%  algorithm given on pages 10,11 of M. Mitchell's book An ...
    Introduction to
%  Genetic Algorithms, MIT Press, 1996). Selection is fitness
%  proportionate.
%clear
%clc
%clf
%close all

% GAzone='nemass';
% GAhr=7;


% load('variables');

% if sun==0;
% vars = cellstr('');
% for i=1:length(variables);
%     string=char(variables(i));
%     Sm1=strfind(string,'Sm1');
```

```
%     Sm3=strfind(string,'Sm3');
%     Sm24=strfind(string,'Sm24');
%     S3=strfind(string,'S3');
%     S0=strfind(string,'S0');
%     if isempty(Sm1) && isempty(Sm3) && isempty(Sm24) && ...
    isempty(S3) && isempty(S0);
%         [vars]=[vars variables(i)];
%     end
%     vars=vars(2:length(vars));
% end
% else vars=variables;

yr1=2006;
yr2=2005;
yr3=2007;


[sun,vars]=anysun(GAzone,GAhr,yr1);
tic;

len=length(vars);                    % The length of the genomes
%popSize=800;                % The size of the population (must be ...
    an even number)
maxGens=6;               % The maximum number of generations ...
    allowed in a run
probCrossover=.3;           % The probability of crossing over.
probMutation=0.03;         % The mutation probability (per bit)
sigmaScalingFlag=1;         % Sigma Scaling is described on pg 168 ...
    of M. Mitchell's
                            % GA book. It often improves GA ...
                               performance.
sigmaScalingCoeff=0.3;        % Higher values => less fitness pressure

SUSFlag=1;                  % 1 => Use Stochastic Universal ...
    Sampling (pg 168 of
                            %      M. Mitchell's GA book)
                            % 0 => Do not use Stochastic Universal ...
                               Sampling
                            %      Stochastic Universal Sampling ...
                               almost always
                            %      improves performance

crossoverType=2;            % 0 => no crossover
                            % 1 => 1pt crossover
                            % 2 => uniform crossover

visualizationFlag=1;        % 0 => don't visualize bit frequencies
                            % 1 => visualize bit frequencies

verboseFlag=1;              % 1 => display details of each generation
                            % 0 => run quietly

useMaskRepositoriesFlag=1; % 1 => draw uniform crossover and ...
    mutation masks from
```

```
88                              %      a pregenerated repository of ...
                                    randomly generated bits.
89                              %      Significantly improves the ...
                                    speed of the code with
90                              %      no apparent changes in the ...
                                    behavior of
91                              %      the SGA
92                              % 0 => generate uniform crossover and ...
                                    mutation
93                              %      masks on the fly. Slower.
94
95
96
97
98
99  % crossover masks to use if crossoverType==0.
100 %mutationOnlycrossmasks=false(popSize,len);
101
102 % pre-generate two repositories of random binary digits from ...
       which the
103 % the masks used in mutation and uniform crossover will be picked.
104 % maskReposFactor determines the size of these repositories.
105
106 maskReposFactor=5;
107 uniformCrossmaskRepos=rand(popSize/2,(len+1)*maskReposFactor)<0.1;%
108
109 mutmaskRepos=rand(popSize,(len+1)*maskReposFactor)<probMutation;
110
111 % preallocate vectors for recording the average and maximum ...
       fitness in each
112 % generation
113 avgFitnessHist=zeros(1,maxGens+1);
114 maxFitnessHist=zeros(1,maxGens+1);
115
116
117 eliteIndiv=[];
118 eliteFitness=-realmax;
119
120
121 % the population is a popSize by len matrix of randomly generated ...
       boolean
122 % values
123
124 pop=rand(popSize,len)<.025;%<---setting lower threshold yield ...
       less variables selected
125
126
127 for gen=0:maxGens
128
129     % evaluate the fitness of the population. The vector of ...
          fitness values
130     % returned  must be of dimensions 1 x popSize.
131     %%%%%%%%%%%%%%%%%%%   fitnessVals=oneMax(pop); original
132     %fitnessVals=[];
```

```
133 %     for c=1:popSize;
134 %         indprime=+(pop(c,:)≠0);%turn logical array into double
135 %         ind=find(indprime);%make index of activated variables
136 %         variablesUsed = [variables(ind)];
137 %         fitnessVals(c) = ...
    [MLRga(variablesUsed,'nemass',15)];%call function with indexed ...
    variables
138 %     end
139      for c=1:popSize;
140          ind=find(pop(c,:));
141          if isempty(ind);
142              pop(c,:)=ones(1,len);%if no variables show up, make lots!
143          end
144          variablesUsed = [vars(ind)];
145          [MAE1]=hindzoneGA(variablesUsed,GAzone,GAhr,yr1,yr2);%call ...
                function with indexed variables
146          [MAE2]=hindzoneGA(variablesUsed,GAzone,GAhr,yr1,yr3);%call ...
                function with indexed variables
147          [MAE3]=hindzoneGA(variablesUsed,GAzone,GAhr,yr2,yr1);%
148          [MAE4]=hindzoneGA(variablesUsed,GAzone,GAhr,yr2,yr3);%
149          [MAE5]=hindzoneGA(variablesUsed,GAzone,GAhr,yr3,yr1);%
150          [MAE6]=hindzoneGA(variablesUsed,GAzone,GAhr,yr3,yr2);%
151          fitnessVals(c) = mean([MAE1 MAE2 MAE3 MAE4 MAE5 MAE6]);
152      end
153
154      [maxFitnessHist(1,gen+1),maxIndex]=max(fitnessVals);
155      avgFitnessHist(1,gen+1)=mean(fitnessVals);
156      if eliteFitness<maxFitnessHist(gen+1)
157          eliteFitness=maxFitnessHist(gen+1);
158          eliteIndiv=pop(maxIndex,:);
159      end
160
161      % display the generation number, the average Fitness of the ...
            population,
162      % and the maximum fitness of any individual in the population
163      if verboseFlag
164          display(['gen=' num2str(gen,'%.3d') '   avgFitness=' ...
165              num2str(avgFitnessHist(1,gen+1),'%3.3f') '   ...
                    maxFitness=' ...
166              num2str(maxFitnessHist(1,gen+1),'%3.3f')]);
167      end
168      % Conditionally perform bit-frequency visualization
169 %     if visualizationFlag
170 %         figure(1)
171 %         set (gcf, 'color', 'w');
172 %         hold off
173 %         bitFreqs=sum(pop)/popSize;
174 %         plot(1:len,bitFreqs, '.');
175 %         axis([0 len 0 1]);
176 %         title(['Generation = ' num2str(gen) ', Average Fitness ...
    = ' sprintf('%0.3f', avgFitnessHist(1,gen+1))]);
177 %         ylabel('Frequency of the Bit 1');
178 %         xlabel('Locus');
179 %         drawnow;
```

```
%     end

    % Conditionally perform sigma scaling
    if sigmaScalingFlag
        sigma=std(fitnessVals);
        if sigma~=0;
            fitnessVals=1+(fitnessVals-mean(fitnessVals))/...
            (sigmaScalingCoeff*sigma);
            fitnessVals(fitnessVals<=0)=0;
        else
            fitnessVals=ones(popSize,1);
        end
    end


    % Normalize the fitness values and then create an array with the
    % cumulative normalized fitness values (the last value in ...
       this array
    % will be 1)
    cumNormFitnessVals=cumsum(fitnessVals/sum(fitnessVals));

    % Use fitness proportional selection with Stochastic ...
       Universal or Roulette
    % Wheel Sampling to determine the indices of the parents
    % of all crossover operations
    if SUSFlag
        markers=rand(1,1)+[1:popSize]/popSize;
        markers(markers>1)=markers(markers>1)-1;
    else
        markers=rand(1,popSize);
    end
    [temp parentIndices]=histc(markers,[0 cumNormFitnessVals]);
    parentIndices=parentIndices(randperm(popSize));

    % deterimine the first parents of each mating pair

    firstParents=pop(parentIndices(1:popSize/2),:);%


    % determine the second parents of each mating pair
    secondParents=pop(parentIndices(popSize/2+1:end),:);

    % create crossover masks
    if crossoverType==0
        masks=mutationOnlycrossmasks;
    elseif crossoverType==1
        masks=false(popSize/2, len);
        temp=ceil(rand(popSize/2,1)*(len-1));
        for i=1:popSize/2
            masks(i,1:temp(i))=true;
        end
    else
        if useMaskRepositoriesFlag
            temp=floor(rand*len*(maskReposFactor-1));
```

```
232             masks=uniformCrossmaskRepos(:,temp+1:temp+len);
233         else
234             masks=rand(popSize/2, len)<.1;
235         end
236     end
237
238     % determine which parent pairs to leave uncrossed
239     reprodIndices=rand(popSize/2,1)<1-probCrossover;
240     masks(reprodIndices,:)=false;
241
242     % implement crossover
243     firstKids=firstParents;
244     firstKids(masks)=secondParents(masks);
245     secondKids=secondParents;
246     secondKids(masks)=firstParents(masks);
247     pop=[firstKids; secondKids];
248
249     % implement mutation
250     if useMaskRepositoriesFlag
251         temp=floor(rand*len*(maskReposFactor-1));
252         masks=mutmaskRepos(:,temp+1:temp+len);
253     else
254         masks=rand(popSize, len)<probMutation;
255     end
256     pop=xor(pop,masks);
257 end
258
259
260 if verboseFlag
261     bestVars=vars(find(eliteIndiv));%variables associated with ...
            max fitness
262 end
263
264 toc;
265 end
```

BIBLIOGRAPHY

[1] Abu-El-Magd, Mohamed A., and Sinha, Naresh K. Short-Term Load Demand Modeling and Forecasting: A Review. *IEEE Transactions on Systems, Man, and Cybernetics 12*, 3 (May 1982), 370–382.

[2] Al-Hamadi, H M, and Soliman, S A. Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. *Electric Power Systems Research 68*, 1 (2004), 47–59.

[3] ASHRAE. *ASHRAE Handbook-Fundamentals*. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., 1997.

[4] Bendat, Julius S., and Piersol, Allan G. *Random Data: Analysis and Measurement Procedures*, 3rd ed. John Wiley & Sons, Inc., New York, 2000.

[5] Bengtsson, L., and Shukla, J. Integration of space and in situ observations to study global climate change. *Bulletin of the American Meteorological Society 69* (1988), 1130–1143.

[6] Bengtsson, L. et al. The need for a dynamical climate reanalysis. *Bulletin of the American Meteorological Society 88* (2007), 495–501.

[7] Bosilovich, Michael G., Robertson, Franklin R., and Chen, Junye. Global energy and water budgets in merra. *Journal of Climate early online release* (2011).

[8] Brunke, Michael A., Wang, Zhuo, Zeng, Xubin, Bosilovich, Michael, and Shie, Chung-Lin. An assessment of the uncertainties in ocean surface turbulent fluxes in 11 reanalysis, satellite-derived, and combined global data sets. *Journal of Climate early online release* (2011).

[9] Burjorjee, Keki. Speedyga: A fast simple genetic algorithm, May 2007.

[10] Draper, Norman Richard, and Smith, Harry. *Applied regression analysis*. Wiley, New York, 1981.

[11] Espinoza, M, Suykens, J A K, Belmans, R, and De Moor, B. Electric Load Forecasting. *Control Systems Magazine, IEEE 27*, 5 (2007), 43–57.

[12] Feinberg, Eugene A, Genethliou, Dora, Pai, M A, and Stankovic, Alex. *Load Forecasting*. Springer US, 2005, pp. 269–285.

[13] GE Energy Applications and Systems Engineering AWS Truepower, EnerNex Coporation. New England Wind Integration Study.

[14] GE Energy Applications and Systems Engineering, EnerNex Corporation, AWS Truewind. Technical requirements for wind generation interconnection and integration. Tech. rep.

[15] Hasheminia, Hamed, and Niaki, Seyed Taghi Akhavan. A genetic algorithm approach to find the best regression/econometric model among the candidates. *Applied Mathematics and Computation 183*, 1 (2006), 337 – 349.

[16] Henfridsson, Urban, Neimane, Viktoria, Strand, Kerstin, Kapper, Robert, Bernhoff, Hans, Danielsson, Oskar, Leijon, Mats, Sundberg, Jan, Thorburn, Karin, Ericsson, Ellerth, and Bergman, Karl. Wave energy potential in the baltic sea and the danish part of the north sea, with reflections on the skagerrak. *Renewable Energy 32*, 12 (2007), 2069 – 2084.

[17] Henson, William L W, McGowan, Jon G, and Manwell, James F. Utilizing Reanalysis and Synthesis Datasets in Wind Resource Characterization for Large-Scale Wind Integration, 2010.

[18] Hurrell, James W, Kushnir, Yochanan, Ottersen, Geir, and Visbeck, Martin. An Overview of the North Atlantic Oscillation. *Structure 134* (2003), 1–35.

[19] ISO New England, Inc. System Operating Procedures: Create Demand Forecast.

[20] ISO New England, Inc. WEM 101: Forecast & Scheduling - Reserve Adequacy Assessment.

[21] ISO New England, Inc. http://www.iso-ne.com/markets/hstdata/hourly/syslds_eei/index.htm, Nov. 2010.

[22] Kapetanios, George. Variable selection in regression models using nonstandard optimisation of information criteria. *Computational Statistics & Data Analysis 52*, 1 (2007), 4 – 15.

[23] Keane, Andrew, Milligan, Michael, Dent, Chris J, Hasche, Bernhard, Annunzio, Claudine D, Member, Student, Dragoon, Ken, Holttinen, Hannele, Samaan, Nader, Söder, Lennart, and O'Malley, Mark. Capacity Value of Wind Power. *IEEE Transactions on Power Systems* (2010), 1–9.

[24] Kenyon, Jesse, and Hegerl, Gabriele C. Influence of Modes of Climate Variability on Global Precipitation Extremes. *Journal of Climate 23*, 23 (2010), 6248–6262.

[25] Khan, M. J., and Iqbal, M. T. Wind energy resource map of newfoundland. *Renewable Energy 29*, 8 (2004), 1211 – 1221.

[26] Kirov, B. Long-term variations and interrelations of ENSO, NAO and solar activity. *Physics and Chemistry of the Earth, Parts A/B/C 27* (2002), 441–448.

[27] Kyriakides, Elias, and Polycarpou, Marios. *Short Term Electric Load Forecasting: A Tutorial*, vol. 35. Springer Berlin / Heidelberg, 2007, pp. 391–418.

[28] Manwell, J F, McGowan, J G, and Rogers, Anthony L. *Wind energy explained : theory, design and application*. Wiley, Chichester; New York, 2002.

[29] Marshall, John, Kushnir, Yochanan, Battisti, David, Chang, Ping, Czaja, Arnaud, Dickson, Robert, Hurrell, James, McCartney, Michael, Saravanan, R, and Visbeck, Martin. North Atlantic climate variability: phenomena, impacts and mechanisms. *International Journal of Climatology 21* (2001), 1863–1898.

[30] Mesinger, Fedor, DiMego, Geoff, Kalnay, Eugenia, Mitchell, Kenneth, Shafran, Perry C., Ebisuzaki, Wesley, Jovic, Duan, Woollen, Jack, Rogers, Eric, Berbery, Ernesto H., Ek, Michael B., Yun, Fan, Grumbine, Robert, Higgins, Wayne, Hong, Li, Ying, Lin, Manikin, Geoff, Parrish, David, and Wei, Shi. North american regional reanalysis. *Bulletin of the American Meteorological Society 87*, 3 (2006), 343 – 360.

[31] Mitchell, Melanie. *An Introduction to Genetic Algorithms*. The MIT Press, Cambridge, MA, 1996.

[32] Moler, Cleve B. *Numerical computing with MATLAB*. Society for Industrial and Applied Mathematics, Philadelphia, 2004.

[33] Montgomery, Douglas C., and Peck, Elizabeth A. *Introduction to linear regression analysis*. Wiley, New York, 2006.

[34] NASA – Global Modeling And Assimilation Office. File Specification for MERRA Products, 2008.

[35] NERC. Methods to model and calculate capacity contributions of variable generation for resource adequacy planning. Tech. rep., North American Electric Reliability Corporation, Princeton, NJ, March, 2011.

[36] Neter, John, and Wasserman, William. *Applied linear statistical models; regression, analysis of variance, and experimental designs*. Irwin, Boston, 1989.

[37] Pap, Judit M., Fox, Peter A., and Frohlich, Claus. *Solar variability and its effects on climate*. American Geophysical Union, Washington, DC, 2004.

[38] Petersen, Erik L, Mortensen, Niels G, Landberg, Lars, Højstrup, Jørgen, and Frank, Helmut P. Wind power meteorology. Part II: siting and models. *Wind Energy 1* (1998), 55–72.

[39] Proceedings of a Symposium on Building Applications of Heat Flux Transducers. *Calculation of thermal conductance based on measurements of heat flow rates in a flat roof using heat flux transducers* (Philadelphia, PA, 1985).

[40] Proceedings of the Eighth Symposium on Improving Building Systems in Hot and Humid Climates. *Disaggregating Cooling Energy Use of Commercial Buildings Into Sensible and Latent Fractions From Whole-Building Monitored Data: Methodology and Advantages* (Dallas, TX, 1992).

[41] Pryor, S C, Barthelmie, R J, and Riley, E S. Historical evolution of wind climates in the USA. *Journal of Physics: Conference Series 75* (2007), 12065.

[42] Pryor, S C, Barthelmie, R J, Young, D T, Takle, E S, Arritt, R W, Flory, D, Gutowski, W J, Nunes, a., and Roads, J. Wind speed trends over the contiguous United States. *Journal of Geophysical Research 114* (2009).

[43] Rienecker, Michele M., Suarez, Max J., Gelaro, Ronald, Todling, Ricardo, Bacmeister, Julio, Liu, Emily, Bosilovich, Michael G., Schubert, Siegfried D., Takacs, Lawrence, Kim, Gi-Kong, Bloom, Stephen, Chen, Junye, Collins, Douglas, Conaty, Austin, Silva, Arlindo da, Gu, Wei, Joiner, Joanna, Koster, Randal D., Lucchesi, Robert, Molod, Andrea, Owens, Tommy, Pawson, Steven, Pegion, Philip, Redder, Christopher R., Reichle, Rolf, Robertson, Franklin R., Ruddick, Albert G., Sienkiewicz, Meta, and Woollen, Jack. Merra - nasa's modern-era retrospective analysis for research and applications. *Journal of Climate early online release* (2011).

[44] Robertson, Franklin R., Bosilovich, Michael G., Chen, Junye, and Miller, Timothy L. The Effect of Satellite Observing System Changes on MERRA Water and Energy Fluxes. *Journal of Climate early online release* (2011).

[45] Schwartz, M. N., and George, R. L. (National Renewable Energy Lab). On the use of reanalysis data for wind resource assessment. *Conference: 11th Applied Climatology, Conference and American Meteorological Society, Dallas T. X.* (1998). (US).

[46] Soliman, S A, Persaud, S, El-Nagar, K, and El-Hawary, M E. Application of least absolute value parameter estimation based on linear programming to short-term load forecasting. *International Journal of Electrical Power & Energy Systems 19*, 3 (1997), 209–216.

[47] Soliman, Soliman Abdel-hady, and Al-Kandari, Ahmad M. *Electrical Load Forecasting: Modeling and Model Construction*. Elsevier Inc., Boston, MA, 2010.

[48] Sterne, Jonathan A. C., and Smith, George Davey. Sifting the evidence: What's wrong with significance tests? *BMJ: British Medical Journal 322*, 7280 (2001), pp. 226–231.

[49] Suarez, Max J (Editor). The GEOS-5 Data Assimilation System – Documentation of Versions 5.0.1, 5.1.0, and 5.2.0. 2008.

[50] The MathWorks. MATLAB User's Guide.

[51] U.S. Climate Change Science Program and the Subcommittee on Global Change Research. Reanalysis of historical climate data for key atmospheric features implications for attribution of causes of observed change. Tech. rep., U.S. Climate Change Science Program, 2008.

[52] Viles, H. Interannual, decadal and multidecadal scale climatic variability and geomorphology. *Earth-Science Reviews 61* (2003), 105–131.

[53] Zielinski, Gregory A., and Keim, Barry D. *New England weather, New England climate*. University of New Hampshire ; Published by University Press of New England, Hanover [N.H.]; Lebanon, NH, 2003.