Marquette University
e-Publications@Marquette

Master's Theses (2009 -)

Dissertations, Theses, and Professional Projects

# Transforming Analogous Time Series Data to Improve Natural Gas Demand Forecast Accuracy

Paul E. Kaefer
*Marquette University*

# TRANSFORMING ANALOGOUS TIME SERIES DATA TO IMPROVE NATURAL GAS DEMAND FORECAST ACCURACY

by

Paul Kaefer, B.S.

A Thesis Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

August 2015

# ABSTRACT
## TRANSFORMING ANALOGOUS TIME SERIES DATA TO IMPROVE NATURAL GAS DEMAND FORECAST ACCURACY

Paul Kaefer, B.S.

Marquette University, 2015

This work improves daily natural gas demand forecasting models for days with unusual weather patterns through the use of analogous data (also known as surrogate data). To develop accurate mathematical models, data are required that describe the system. When this data does not completely describe the system or all possible events in the system, alternative methods are used to account for this lack of information. Improved models can be built by supplementing the lack of data with data or models from sources where more information is available.

Time series forecasting involves building models using a set of historical data. When "enough" historical data are available, the set used to train models exhibits ample variation. This results in higher accuracy in GasDay$^{TM}$ natural gas demand forecasting models, since there is a wide range of history to describe. In real-world applications, this also means that the data are more realistic, due to the stochastic nature of real events. However, it is not always the case that "enough" historical data are available. This may be due to few years of available historical data, or a case where available data does not exhibit as much variation as desired.

By taking advantage of GasDay's many customers from various geographical locations, a large pool of data sets may be used to address this problem of insufficient data. Data from utilities of similar climate or gas use may be used to build useful models for other utilities. In other words, available data sets may be used as analogues or surrogates for building models for areas with insufficient data.

The results show that the use of surrogate data improves forecasting models. Notably, forecasts for days with unusual weather patterns are improved. By applying clever transformation methods and carefully selecting donor areas, the methods discussed in this thesis help GasDay to improve forecasts for natural gas demand across the United States.

# ACKNOWLEDGMENTS

Paul Kaefer, B.S.

I would like to thank everyone who directly or indirectly influenced my education, both in my time at Marquette University and previously. There are too many people to list who have contributed to where I am today.

A special thanks goes out to the many participants of the GasDay Lab. Notable are the graduate students who started before me and gave me advice to assist me on my adventure, including Hermine Akouemo, Tian Gao, James Gramz, Samson Kiware, James Lubow, Sanzad Siddique, and Steve Vitullo. An array of students involved in the project provided technical and moral support including, but certainly not limited to, Britt Ahlgrim, William Castedo, Ben Clark, Maral Fakoor, Brandon Howard, Babatunde Ishola, Calvin Jay, David Kaftan, Tim Kehoe, Zach Nordgren, Mohammad Saber, Nathan Wilson, and Nick Winninger. I wish them all the best in their current and future endeavors.

I could not have accomplished so much were it not for my many mentors and advisors. These include my committee, Drs. Ronald Brown, George Corliss, Stephen Merrill, and Richard Povinelli. These also include Paula Gallitz, Catherine Porter, and Thomas Quinn, as well as the many professors I've had along my journey at Marquette University. I would also like to thank the Marquette University Raynor Memorial Libraries for being an invaluable resource.

I would not be where I am today without my parents, Frederick and Jeanne. My journey began with them reading to me as a child, encouraging me in my pursuit of mathematics and engineering, and supporting me in all of my endeavors.

I dedicate this work to my siblings, Stephanie, Matthew, and Conor. They light up my world.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

## Building Good Forecasting Models

This chapter introduces the natural gas industry, the importance of developing good forecasts, and an overview of the work done at GasDay. The focus of this research is described in the context of the GasDay lab. A synopsis of our research is also presented.

### 1.1 The Natural Gas Industry

The United States of America consumes approximately 26% of the world's total annual natural gas consumption, according to the U.S. Energy Information Administration (EIA) [51]. Natural gas is used for a variety of purposes including space heating, cooking, water heating, clothing dryers, electric power generation, industrial processes, and as a vehicle fuel.

In the United States, natural gas futures are traded on the New York Mercantile Exchange (NYMEX). The "gas day" is defined as the 24-hour period that starts at 9 A.M. Central Time and ends at 9 A.M. the next calendar day [22]. Energy utilities have to predict the amount of gas they expect to sell or use in advance of each gas day. They are highly motivated to get their forecasts correct, as there are penalties for both low and high forecasts. If the utility does not nominate enough gas in advance, it may be forced to curtail certain customers or purchase the gas on the sometimes much more expensive spot market. If they have too much gas due to a high forecast, they may incur expensive storage costs or penalties for not withdrawing the gas from the pipeline. Many natural gas Local Distribution Companies (LDCs) do not have available storage [52].

The costs incurred by the utility are passed on to their customers, which encourages the utilities to develop good forecasts. In some cases, customers may be able to switch to an alternative energy source such as electricity or heating oil if they are experiencing high prices from their natural gas utility.

Marquette University GasDay is like a small business that is housed in a research lab that forecasts the daily demand for about 17.7% of the natural gas consumed by residential, commercial, and industrial customers in the United States. This market share has been achieved through years of research and work improving the methods used to forecast in such an interesting industry.

## 1.2 A Brief Introduction to Natural Gas Forecasting Methods

This section gives a brief overview of forecasting natural gas demand. The second chapter of this thesis delves further into forecasting methods and an overview of work that has been done in the field of gas demand forecasting.

Various methods are used to forecast gas demand. Linear models have been used, as mentioned in [6, 11, 47]. Box-Jenkins [7] methods are used by [40]. Artificial neural networks are used by [31, 52]. Other nonlinear models are used by [30]. Vitullo et al. [52] discuss various mathematical models in use, including those currently in use at the GasDay laboratory.

There are some interesting techniques in use in the energy domain. One is the nonlinear effect of temperature on demand. Demand is linearly modeled with temperature inputs, but only below a certain threshold temperature. People turn on their furnaces once the temperature outside is below a certain level. This concept is called the heating degree day (HDD) [13], similar to the degree day found in fields such as parasitology [33], and the temperature transformation done by [40], and is calculated as

$$\text{HDD} = \max(T_{ref} - T, 0) \ ,$$

where $T_{ref}$ is typically 65°F. By using this value instead of the raw temperature value, more accurate forecasting models are developed [44, 47].

Balestra and Nerlove [6] use consumer behavior to improve models of natural gas demand. In Section 3.2.4, we consider this and other features that are unique to natural gas time series and help improve our modelling techniques.

## 1.3 Forecasting at GasDay

GasDay's flagship product forecasts daily demand for time horizons of up to eight days for natural gas service territories known as operating areas. (Operating areas are also known as service territories, market areas, or zones; we use the term "operating area" for consistency.) An ensemble model [26] of linear regression models and artificial neural networks is used. These models are trained on historical data, so forecasts for unusual day types (which will be further explained in Chapters 2 and 3) depend on the available data from past events. Because not all gas utilities are able to provide many years of good data, surrogate data from other utilities is used to build forecasting models [20]. "Surrogate data" is also referred to as "analogous data" or "analogous time series" when the data comes from time series data sets.

## 1.4 Using Analogous Time Series in Forecasting Models

Chapter 3 provides mathematical details for our analogous data transformation algorithms. The high-level concept is to use data from other sources that exhibit

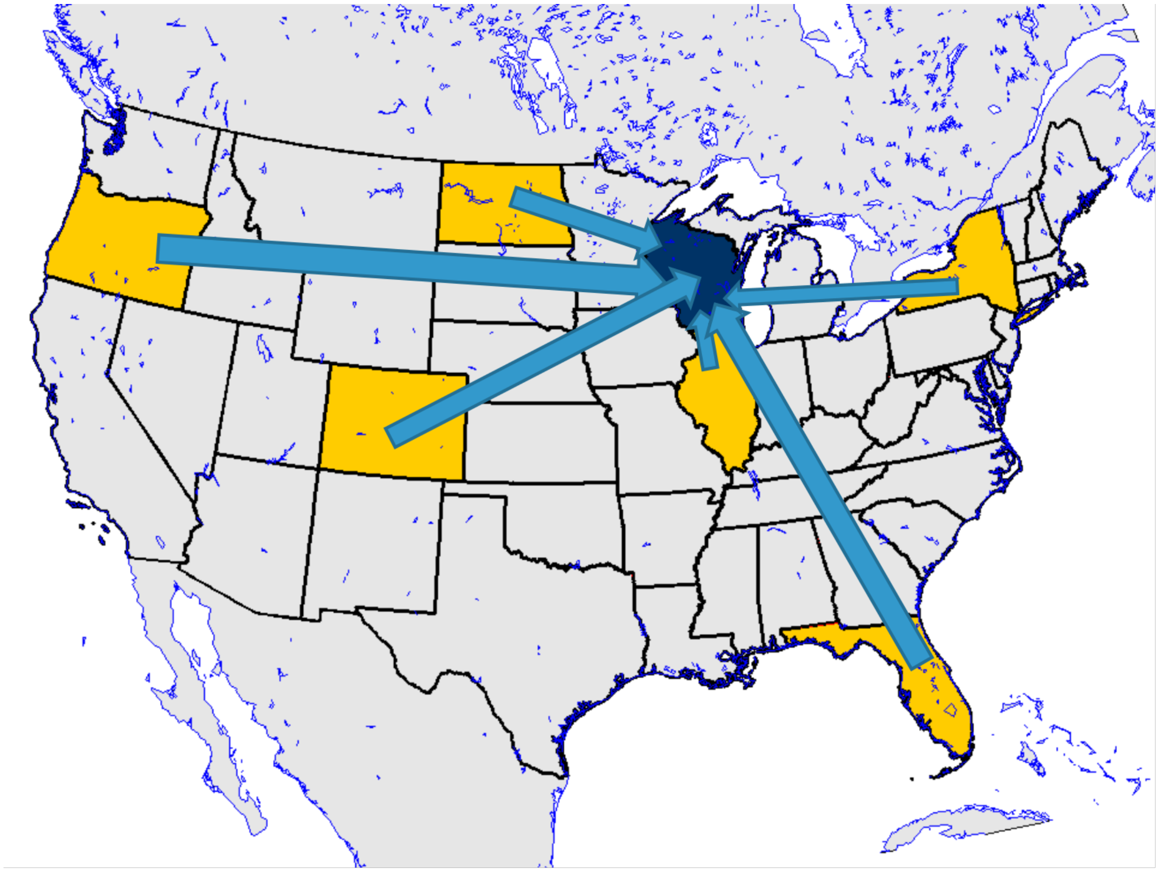

Figure 1.1: Data from different states may be used as surrogate data for the state of Wisconsin.

similar patterns to a given data set. Consider Figure 1.1. It may be the case that only a few years of historical data for the state of Wisconsin are available. If similar patterns are seen in Illinois, Oregon, etc., these data may be transformed to look like Wisconsin data. In this example, we refer to Illinois and Oregon as potential “donor” areas and to Wisconsin as the “target” area. Our research explores methods of determining good surrogate donors and methods of transforming data so they build better models.

If we consider data for different areas throughout the country, it is apparent that the areas are quite different due to various reasons including climate and customer bases. Figure 1.2 plots gas flow against (65 − temperature) for many areas in the country. By using several of these sets, it is obvious that a combined model will not forecast an individual area very well. Some of them have quite steep trends as the number of heating degree days increases (i.e., the HDD increases). In addition, some of these areas have been scaled to fit on the graph, as their typical flow values are 1000 times the plotted values.

It is apparent that work needs to be done in order for the data sets to be useful in building combined models.

Our work focuses on transforming these data sets to make them match the given target area for which we want to build better models. Our work is valuable because we scale and shift the data so that it has similar properties as the target area. However, we do not want to mangle the data or manipulate it in any way such that we cannot transform it back to the raw form. Through this process, we are able to transform donor data sets from any area in the country to make them look like they came from a different target area. This addresses the problem at hand by improving natural gas daily forecasts for unusual days where there would not otherwise be enough data.

Transformed surrogate data are used only for training models. Insufficient data is only a problem to the point that it leads to the risk of models that cannot

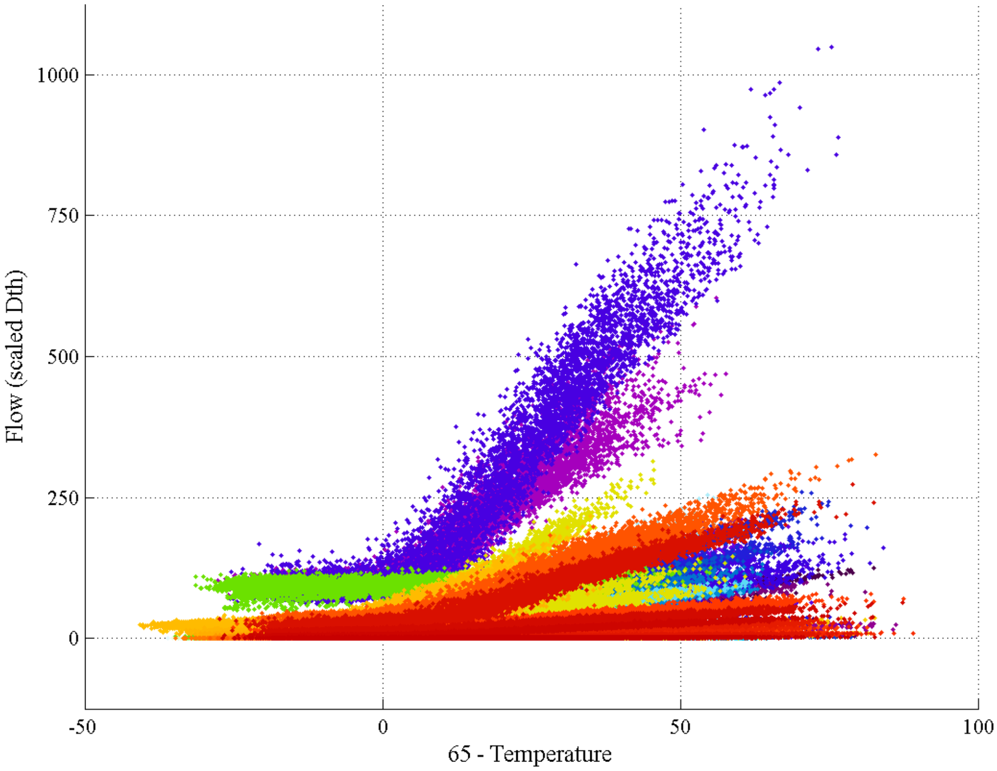

Figure 1.2: Data from many areas in the country.

accurately predict events that have not been seen in the past. The use of surrogate data enables us to train models that hypothetically produce better forecasts for these events. We then evaluate these models on testing data that originates solely from the target area. In practice, the models will not be used to forecast any of the donor areas, but rather the original target area.

Figures 1.3 and 1.4 illustrate this process. In the typical forecasting process, historical data are used to train models. These models accept exogenous variable input, and output forecasts, as seen in Figure 1.3. Analogous data are added for model training, as shown in Figure 1.4. Once again, exogenous variables are used by the models, and a different set of forecasts will be produced (since the models were trained on different data). We compare the forecasts from both to evaluate this research.

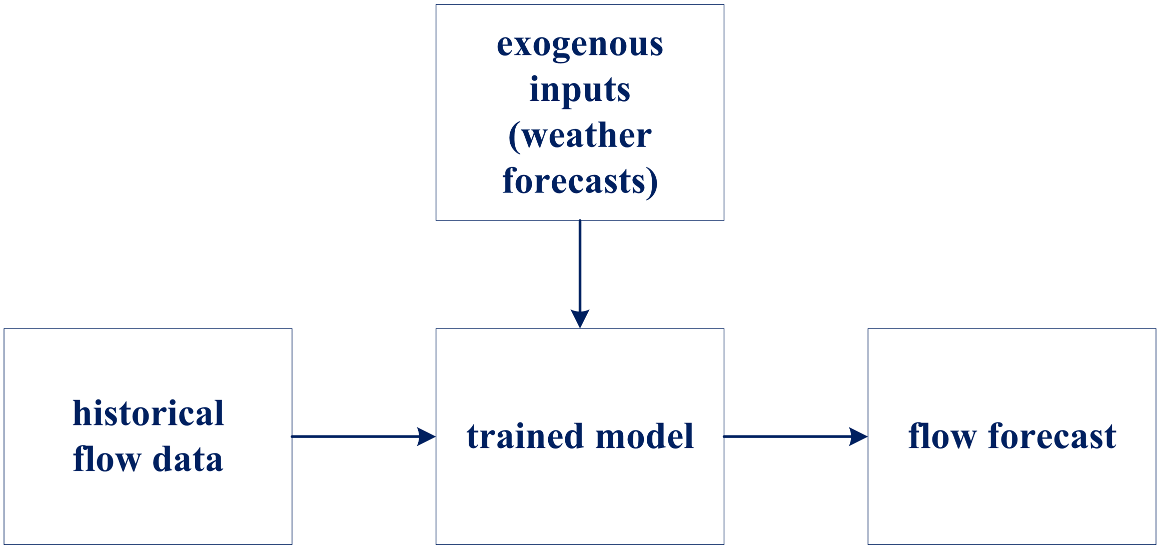

Figure 1.3: Forecasting process before analogous data are used.

## 1.5 Organization of this Thesis

The remainder of this thesis expands upon what has been explained in this chapter. In Chapter 2, we will discuss state-of-the-art natural gas demand forecasting and existing work addressing the problem of insufficient data. Chapter 3 will explain our analogous data transformation algorithm, including how we determine which data sets to transform and use in building models. The results from these models,

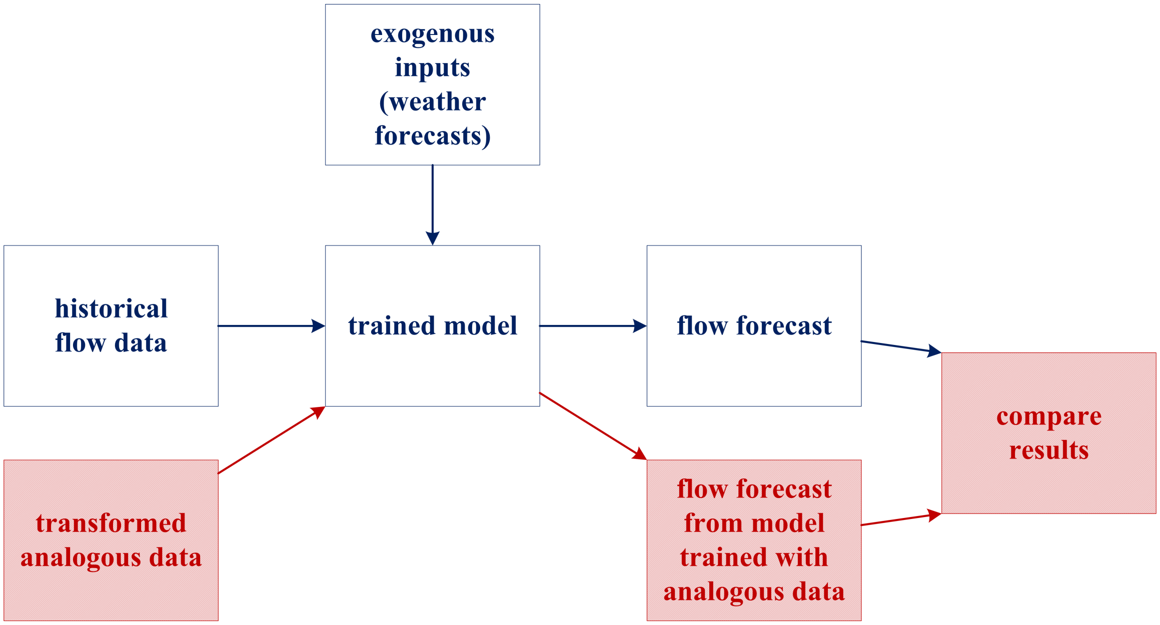
exogenous inputs (weather forecasts)
historical flow data
trained model
flow forecast
compare results
transformed analogous data
flow forecast from model trained with analogous data

Figure 1.4: Analogous data within the forecasting process. Our work adds the elements in red to the process in Figure 1.3.

compared against models built without analogous data, will be presented in Chapter 4. Finally, we discuss in Chapter 5 ideas for future research that are inspired by our work.

# CHAPTER 2

# Forecasting Models and Surrogate Data

The purpose of this chapter is to discuss state-of-the-art of forecasting and methods that are used to address the problem of not having sufficient data to develop a good forecast.

## 2.1 Forecasting Research

Much work has been done in the area of forecasting. De Gooijer and Hyndman provide an excellent review of various methods that are used in forecasting [16]. Soldo [44] and Vitullo, et al. [52] provide reviews of methods used specifically in the area of natural gas modeling and forecasting.

In the area of natural gas forecasting, research has been done around the world. Research in England by Piggott uses the concept of heating degree days to capture the not-exactly linear relationship between temperature and gas demand [40]. We use heating degree days (as explained in Chapter 1) for the same reason. In Iran, a Smooth Transition Auto-regression (STAR) model was used by Kani et al., as lagged variables capture nonlinear properties of natural gas demand [30]. Likewise, GasDay uses autoregressive terms. The impact of previous days on the current demand is the reason that many “unusual events” used here are multiple days long. Akkurt et al. [2] looked at different demand patterns in residential, commercial, and industrial groups in Turkey. While we do not explicitly separate an area’s data into these groups, we recognize that customer behavior differs in different areas, and the algorithm used to transform data sets takes into account different patterns.

Estimation of natural gas demand using ordinary least squares is done in [6]. This research uses least squares regression to validate the improvements from our algorithms. Linear regression is easy to do quickly and efficiently. Hence, we use it to test many different combinations of surrogate data to verify the accuracy of the results.

## 2.2 Insufficient Data

Since we want to predict demand on unusual days better than current practice, we wish we had more data than is available. If we have only one data point for a “once-in-30-years” event, it will only be a good predictor if a future such event behaves in exactly the same way, but events never happen in exactly the same way. Hence, it is very common that we do not have many data points for gas demand during unusual weather events. Hence, we search for more data.

Balestra and Nerlove [6, p. 607] mention having an insufficient amount of years of gas data given the number of estimated parameters to determine for their model. Mohri [37] and others also recognize that in many practical applications, there is insufficient data with which to both train and test models.

One method to combat having insufficient data uses hierarchical models. Models are built at different levels. The lowest level is the raw data. Parameters at different higher levels allow for generalization of levels below. Duncan, Gorr, and Szczypula [19] use hierarchical models to forecast government revenue for different school districts and find that they helped improve accuracy. Hierarchical forecasting also is done in the energy domain for electrical load forecasting [27]. In our research, we do not build hierarchical models, but similar to the previous research, we are able to construct an algorithm that describes a general natural gas demand time series set and then apply this algorithm to specific cases.

When our modeling application does not have enough data, we can supplement available data with data from other sources. To do this, we must find data that is similar in some way(s).

Dixon et al. [17] leverage auxiliary data for enhancing models for rare events. They look at several methods for getting this auxiliary data including aggregated data that corresponds to a different time scale and using probability distributions to generate data. We are also seeking more data to predict rare events. In some cases, gas utilities do provide data at different time scales. For example, they may have customer data aggregated into billing cycle frequency [12]. In this case, the characteristics of a rare event in that data are masked by the other days with which it is aggregated. The idea of using a probability distribution is interesting. However, the literature does not show a probability distribution that describes natural gas demand, or particularly gas demand during rare events.

Thomas [49] explores the use of data from similar products in estimating the market trends for a new product. He identifies similar products for which data are available and use the parameters from models for these similar products in modeling response to the new product. Likewise, we identify data sets that are similar to a target set. Rather than using the same coefficients from models, we transform the data so the coefficients of these models match the target area. We then use the transformed data, rather than the model parameters.

## 2.3 Determining “Similarity” in Data

Twenty plus years of natural gas forecasting research performed at the GasDay lab has resulted in the accumulation of domain knowledge that has contributed to this research. One important contribution is the familiarity with patterns in natural gas consumption time series. Essentially, we have a heuristic for how natural gas data should look. In addition, the patterns are seen independently in areas around the

United States with vastly different weather and climate patterns. As a result, we have found [10] that surrogate data works, at least by "guessing" through the use of a genetic algorithm which combination(s) of transformed data will perform well as surrogates.

Here, we make more educated guesses by comparing data sets, grouping similar sets, and using those in our surrogate data algorithms. Clustering is one method in which this may be performed.

The winner [42] of GEFCom 2012 [27] used $k$-means clustering to forecast wind power generation. The idea behind the clustering technique is mathematically grouping data based on known or calculated features. This could be Bayesian, as was done through prior domain knowledge by [19]. Even if there is no prior information, features can be extracted from the data and used to compare data sets for similarity.

Eigenfaces [50] and eigenvoices [34] are applications in facial recognition and speech processing, respectively, that involve recognizing features of the data that vary for individuals but are recognized as the same feature across the population. There are patterns in facial structure and in human voices that are computationally detectable, despite the differences that appear among individual examples of faces or voices.

Both eigenfaces and eigenvoices illustrate the idea that mathematical features are not necessarily the same features that are visible to the human eye. While the human eye recognizes eye color and nose shape, a computer recognizes mathematical similarity between the digital representation of two facial images. Likewise with natural gas data, there are features and patterns that require computational resources to detect. Experienced humans can recognize trend almost instantaneously, but it would take much more time staring at graphs or using a pencil and paper to realize other underlying patterns such as what happens in the days preceding a cold event.

Brown et al. [10] suggest the use of surrogate data in natural gas forecasting. Transforming data from operating areas that show similar customer or weather characteristics may be able to address problems that arise when the testing set has events or patterns not seen in the available training data. The GasDay lab currently uses surrogate data to help build accurate forecast models for extreme weather events. Similar to the application of eigenfaces and eigenvoices, this enables the lab to apply natural gas demand patterns that are found across data sets.

We use the $k$-Nearest-Neighbors ($k$NN) algorithm to select data sets with similar parameters of transformation. The $k$NN algorithm is well-known in the data mining community and is reviewed by Wu et al. [53].

Available literature shows several methods of dealing with insufficient data.

In many cases, there are not enough data for a specific area or model, but there are data from other sources that are similar in some way. Our lab historically has termed this the use of “surrogate” data, which we explain further in the next section.

## 2.4 Surrogate or Analogous Data

The terms “surrogate” and “analogous” appear in the literature in different contexts. Some of these contexts are explained briefly here to distinguish where our research fits.

Forrester, Sóbester, and Keane [23] present several uses of surrogate modelling techniques. One is to reduce computational complexity by using smaller surrogate sets. Another is for handling noisy or missing data. The use that most closely matches our research is the use of surrogate models to gain insight into the variables and results in a given problem. As explained later in this chapter, we want our surrogate data to improve forecasts in the case of variable combinations that have not been seen historically or have been seen very infrequently. By using surrogate data from other sources that exhibits similar patterns, we can create models that are better able to respond to unusual patterns in variables.

Surrogate data tests [24] are used to determine nonlinearity of a data set [41]. Often, they involve the creation of synthetic surrogates (with similar properties of the original data) that are used in statistical analysis. In some cases, “surrogate surrogates” are used to calibrate surrogate data tests for accuracy [48]. Surrogates used in our research are not synthetic or randomly generated, but rather transformed data from other sources. We use synthetic data sets to verify functionality of our code that performs the transformations. Another difference characterizing our research is that rather than use surrogates to examine properties of the original data, we use surrogates alongside this data. Rather than use surrogates to test our data, we use the original data and surrogates in combination and test the results from forecasting models.

Duncan, Gorr, and Szczypula [20] pooled data from analogous time series and found that the resulting group models improved forecast accuracy. This is a close fit to our research. By pooling or aggregating data from multiple sources, we improve accuracy in forecasts.

If we look back at Figure 1.2 from Chapter 1, it is apparent that we cannot simply pool natural gas data sets that have vastly different ranges and even patterns. Thus, we have to do something to make the data sets transformable. Fortunately, work has been done to this end. In mathematics, a canonical form of a matrix is used to describe the similarities that exist across matrices that initially look very different.

## 2.5 The Canonical Form of Data

The canonical form or Jordan form of a matrix is a way of representing a matrix in a reduced form such that it may be translated back to the original [43].

We do not use the canonical form, but it is a helpful concept when considering the problem of converting one data set to match features and parameters of another. If we consider both data sets the "in between" form that is a canonical form, we want our algorithm to be able to convert to and from this form seamlessly. There is no well-defined canonical form of gas data, but the thought exercise helps develop the concept that there are quantifiable similarities across natural gas demand data sets that help us convert a donor set to match a target set.

## 2.6 Current Practices at GasDay

Currently, GasDay uses surrogate data to forecast unusual day types [39], including near design days, days that are colder than normal, very windy or humid days, and the first heating and non-heating days of the year (i.e., days where people are just turning their furnaces on or off for the heating season). Specifically, GasDay considers the day types listed in Table 2.1.

Current practice [29] is to perform temperature scaling, so the HDD reference temperature and $95^{th}$ percentile temperatures of the donor and recipient area(s)

Table 2.1: Unusual day types at GasDay.

| |
|---|
| Coldest days |
| Days that are colder than normal weather |
| Days that are warmer than normal weather |
| Windiest days |
| Days that are much colder than the day before |
| Days that are much warmer than the day before |
| The first cold day of the year |
| The first warm day of the year |
| High humidity heating days |
| Low humidity heating days |
| Sunny heating days |
| Cloudy heating days |

match. This is shown in Figure 2.1. Next, similar scaling is done on the natural gas flow. As shown in Figure 2.2, when temperature and flow values are scaled, the donor area comes close to matching the target area. Finally, multi-dimensional scaling is done so various characteristics of the gas consumption in the donor area(s) match the target area. These characteristics include baseload (amount of gas used for non-temperature-related purposes, such as cooking, drying, and water heating), heatload (amount of gas used for space heating), day-of-week, and other cooling- and heating- degree day characteristics. These characteristics are not matched when the $5^{th}$ and $95^{th}$ percentile values of temperature and flow are matched.

Unusual day types are selected from the transformed surrogate data. A genetic algorithm then is used to determine the best combination(s) of surrogate donors. These are evaluated based on how well they backcast demand on unusual

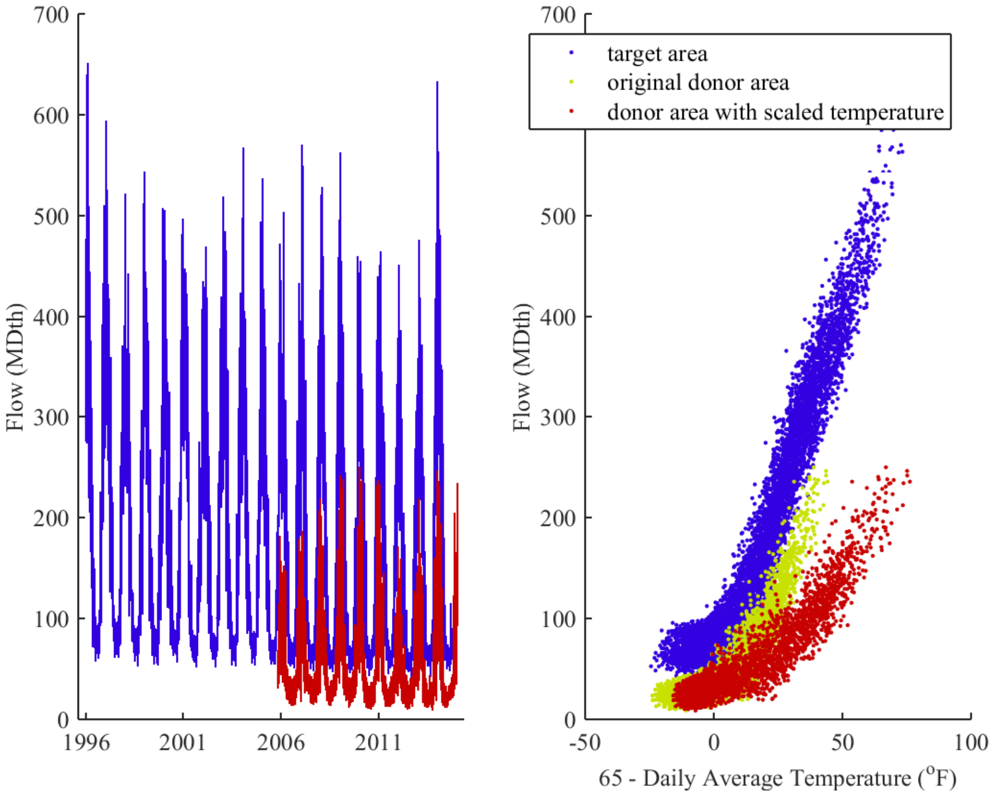

Figure 2.1: The temperature values for a donor area have been scaled to match those of the target area.

days for different test sets. GasDay considers several unusual day types, shown in Figure 2.3.

Winter periods tend to have different characteristics. For example, some winters are milder, whereas others seem to drag on and have multiple cold events. As a result, cross-validation of surrogate sets could be used. Since we cannot test how well the models with surrogate data will perform on events that have not yet happened, we can backcast several different winter periods. In this manner, we can be reasonably confident in the accuracy of a model if it performed well backcasting

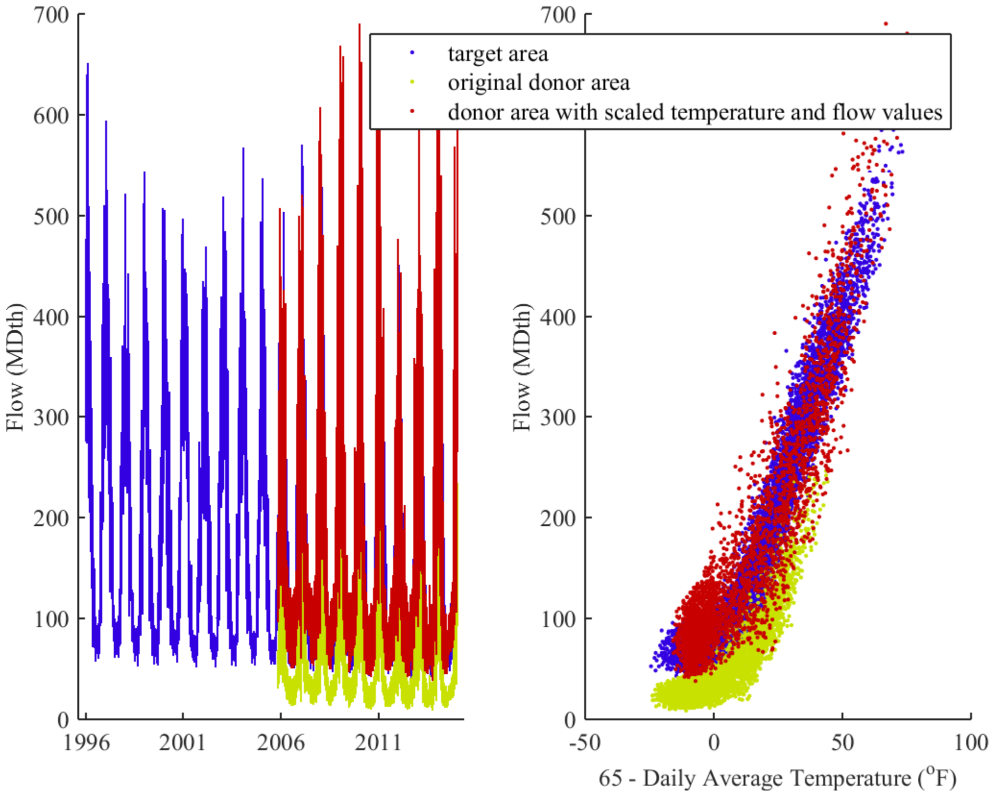

Figure 2.2: The temperature and flow values for a donor area have been scaled to match those of the target area.

several different “types” of winters. However, it is difficult to select which of historical winters the upcoming winter will be like. Hence, we perform ex ante forecasts of the 2013-2014 winter. This shows us that had our methods been used, we would have forecasted the past winter better than we did without using our surrogate data methods discussed in this thesis.

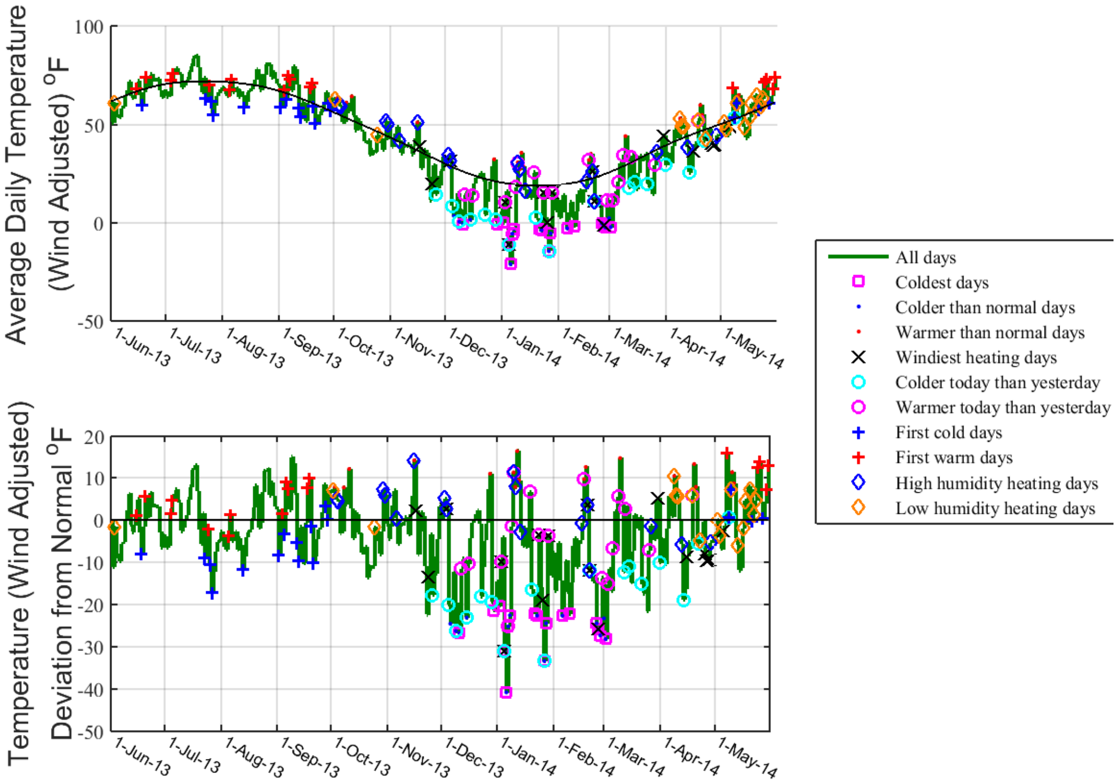

Figure 2.3: Unusual day types considered by GasDay.

## 2.7 Volatility in Time Series Data

Makridakis et al. [35] write, “It is believed that the greater the randomness in the data, the less important is the use of statistically sophisticated methods.” The goal of our research is not to create more sophisticated models. Rather, we can leverage available data to create a better model for areas that may have very random patterns, and not enough native data to model that inherent randomness in the data.

Rather than the term “randomness,” we use the term volatility. Volatility is a measure of variation in time series, for example in foreign currency [36] and stock market [45] applications. Different data sets exhibit different levels of volatility. This impacts how well donor data act as surrogates for other data sets. The methods of surrogate data transformation need to take this into account.

Duncan, Gorr, and Szczypula [20] found that time series with higher volatility benefited the most from methods that involved pooling data. Natural gas consumption data is a time series that is notably volatile. At different times during the year, we can expect there to be different levels of variation in the gas consumed. In the summer months, the baseload is expected to be fairly consistent. In summer, residential consumers tend to use gas primarily for cooking and water heating. As the temperature outside gets colder, heatload increases as people use gas to heat their homes. Since the weather in most of the United States is more volatile in the winter (as seen in Figure 2.4), the gas load is also more volatile (as seen in Figure 2.5). Since gas consumption data exhibits volatility, we expect that by pooling data through our surrogate data techniques, we will get more accurate forecasts.

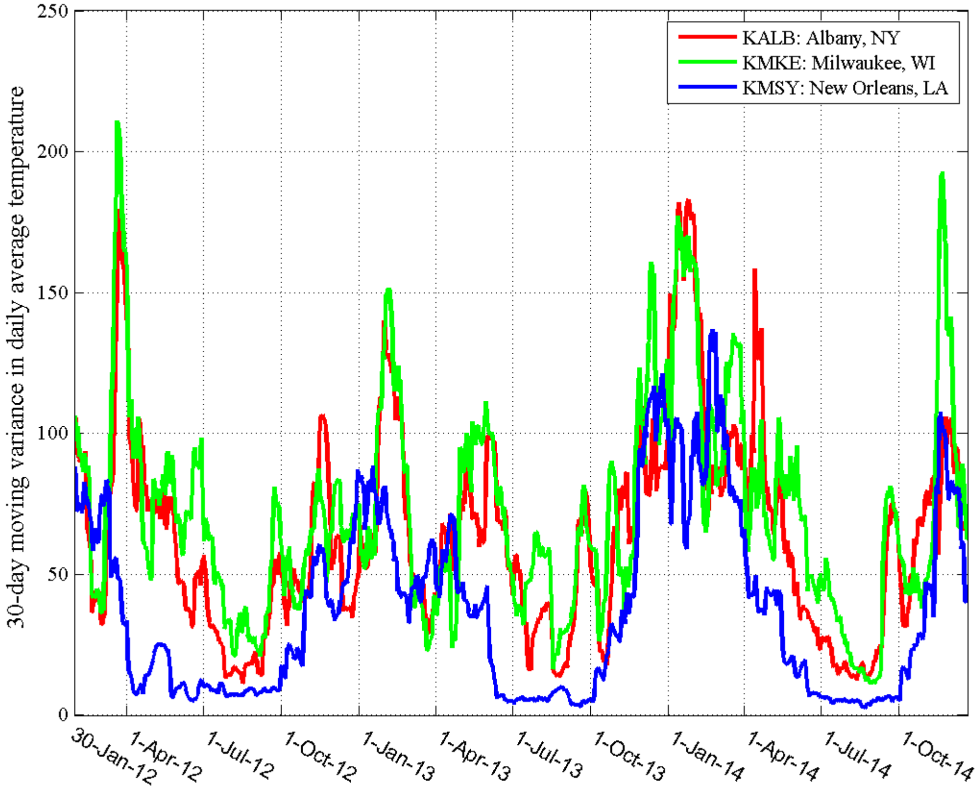

Figure 2.4: Volatility in daily average temperature for three different weather stations in the United States.

## 2.8 Using State-of-the-Art in Our Research

Our research leverages previous work in the areas described in this chapter. Our surrogate data algorithms transform time series data to be used to improve models specifically for days with unusual weather patterns. The next chapter describes the methods we use to apply the use of analogous data to the natural gas demand forecasting domain.

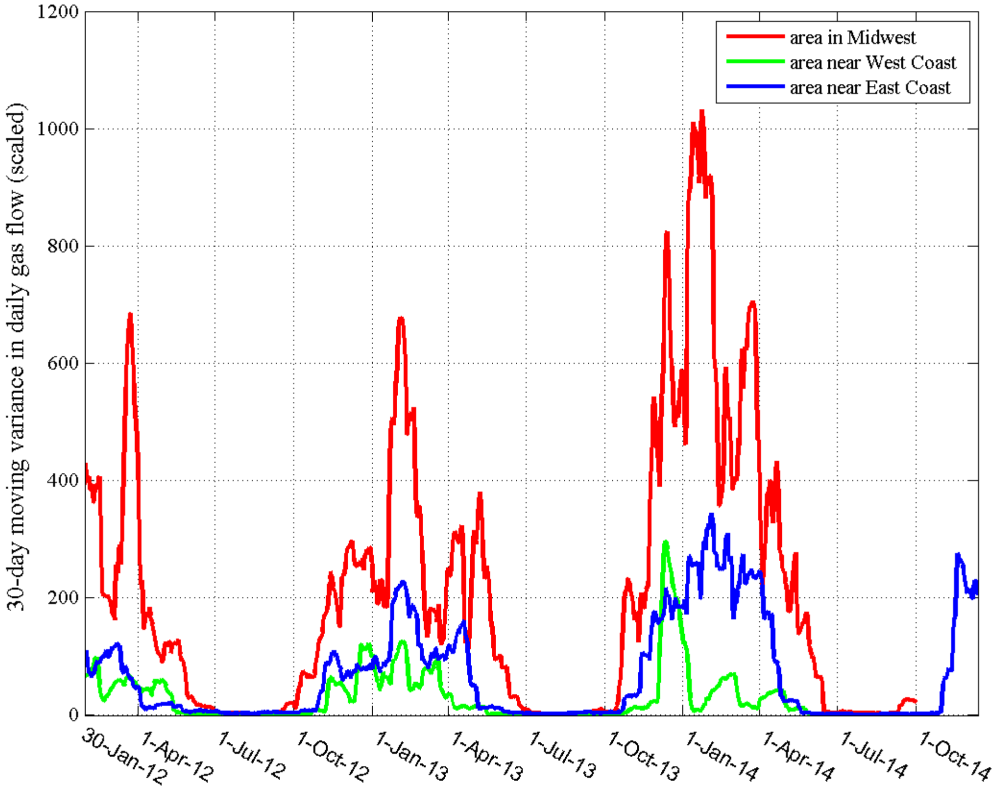

Figure 2.5: Volatility in daily gas flow for three operating areas in the U.S.

# CHAPTER 3

## Applying Surrogate Data to Improve Forecasts

This chapter discusses our methods of applying surrogate data to natural gas forecasting models. We discuss current practice, followed by the improved methods that are the focus of this thesis. For both, the methods require

1. transforming the data from its original (or donor) form to match the recipient (or target) data,
2. selecting a number of sets to use as surrogates, and
3. evaluating the resulting model with surrogate data.

Each operating area for which GasDay produces forecasts has sample data. This is generally several years of daily weather and gas flow data that are used to build a mathematical model. This model is used to make daily forecasts for the following calendar year. Notable is the winter, when gas demand is typically the highest. Weather variables are unknown until they actually happen, which is also when the data are most needed. This is because gas forecasts for the current day typically are made before the previous gas day is over.

The GasDay lab has accumulated extensive domain knowledge of over two decades of research and working with natural gas utilities. This knowledge is applied to the problem of how to develop good forecasts for days where we have little relevant historical data.

GasDay has data from more than 150 operating areas from local distribution companies (LDCs) from across the United States. The current practice [29] is to select data from unusual events from the various operating areas and transform them.

"Unusual events" are days with weather patterns that are abnormal, such as much colder than normal weather, that have higher flows than normal, or perhaps other nonlinear relationships with the weather. For example, days that are much warmer than the previous day are considered unusual days. When it gets unusually cold, a day may experience much higher flow than is expected simply because the previous day was much colder. Customer behavior [10] and thermodynamics [9] may account for the difference from a day with the same temperature that had a warmer previous day. Research is in progress to explain these events.

Once gas flow from a set of unusual day types has been transformed to match the target area, a genetic algorithm tries different combinations of transformed donor data, searching for the best combination of data that works as good surrogate data for building models to predict future unusual events. Good surrogate data contributes to models which forecast more accurately than a model trained where there are not many unusual data points. It is important to note that the original data (before transformations) has been preserved, so the transformations do not impact the integrity of the raw data sets.

Section 3.2.3 explains further what it means to transform data to match a set of features and make it appear that it originated from a different source. Consider Figure 3.1. The two time series look very different, as the flow in the red area is much higher than that of the blue area. However, similar patterns are visible, with gas flow being highest in winter and lowest in the summer for both areas. The goal of this chapter is to explain the methods with which we transform the donor to match the target.

## 3.1 The Brown Surrogate Algorithm

In this thesis, the current GasDay surrogate data algorithm is called the Brown Surrogate Algorithm. This section presents this algorithm. The algorithm is introduced and explained in [29].

The Brown Surrogate Algorithm has five steps:

1. Scale temperature of donor(s) to match the target area. Temperature values are scaled so they have properties similar to those seen in the target area.

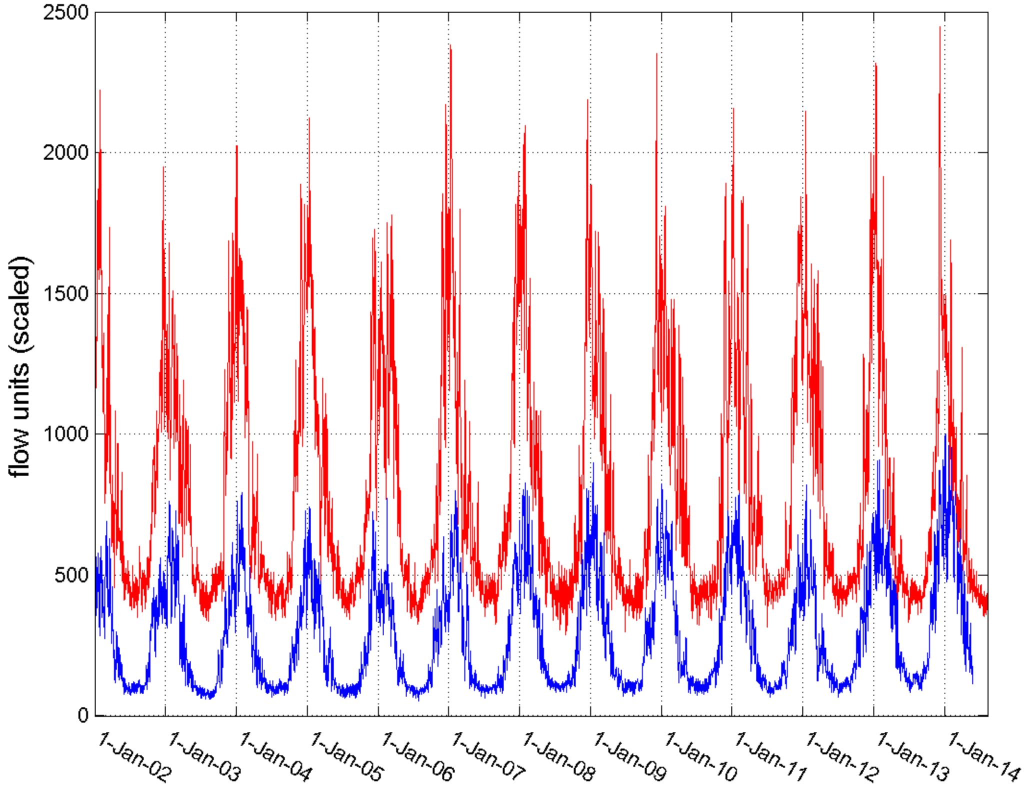

Figure 3.1: Time series of two operating areas from different places in the country. Proprietary data has been scaled for anonymity.

2. Scale natural gas flow data of donor(s) to match the target area. Similarly to the temperature scaling, this makes it look like the gas flow was originally from the target area.

3. Use a multi-dimensional detrending algorithm [14] so regression coefficients of the donor area(s) are close to those of the target area. This also makes historical data closely resemble recent data and provides the benefit of allowing all training data to behave as if they were from the most recent year.

As a result, positive or negative growth patterns are accommodated. This accounts for a multi-dimensional linear trend.

4. Identify and select unusual day types from Table 2.1.

5. Select and evaluate donors using a genetic algorithm.

Each of these are discussed in turn.

## 3.2 Improved Surrogate Selection and Transformation Algorithms

Our Universal Surrogate Algorithm improves upon current practice by accounting for prior information to select the best surrogate donors. The Brown Surrogate Algorithm was built under the assumption that a genetic algorithm would find a good set of surrogate data with which to build better models. Our new algorithm is "universal" in the sense that it may be used to transform any given daily natural gas consumption data set to match any other such set. It is less computationally intensive than the Brown Surrogate Algorithm, as it does not use a genetic algorithm, and thus may be used more universally, without costly time constraints.

In addition to improved selection methods, we have improved upon the transformation algorithm used to consider time series volatility.

The methods used in our research result in better accuracy of models over current practices. Less computational resources are required, which enables more surrogate data use and further research potential. We can transform a data set quickly and try new techniques instead of having to wait for a computationally-intensive genetic algorithm.

In our Universal Surrogate Algorithm, we improve on existing methods also by performing day-of-year scaling. In different areas of the country, winters have different start and end dates and durations. By scaling the day-of-year characteristics, we can simulate areas with different winter characteristics being similar to target areas.

Our Universal Surrogate Algorithm has the following steps, which we can contrast with the Brown Surrogate Algorithm:

1. Scale temperature of donor(s) to match the target area, just like the Brown Surrogate Algorithm
2. Scale natural gas flow data of donor(s) to match the target area, as we do in the Brown Surrogate Algorithm
3. Also scale by day-of-year to account for the fact that seasons start and end at different times during the year in different climates
4. Improve upon the multi-dimensional detrending algorithm [14] to include day-of-year terms

5. Identify and select unusual day types, as done in the Brown Surrogate Algorithm
6. Use natural gas time series features, rather than a genetic algorithm, to determine which donor set(s) to use

The following sections outline how data are converted (by further explaining steps 1-4 of the Universal Surrogate Algorithm), how the converted sets are selected to act as donors (details of steps 5 and 6), and how the resulting set is evaluated.

### 3.2.1 Conversion of Donor Data to Match Target Data

We scale temperature, flow, and day-of-year values of the donor area(s) to match the target. The temperature and flow scaling are performed by linear transformation. The general equation is given in Equation (3.1). $P_5$ and $P_{95}$ represent the $5^{th}$- and $95^{th}$-percentile values, respectively. The donor data, $x_{donor}$, is transformed to match the target data, $x_{target}$.

$$x_{\text{transformed}} = \frac{(x_{\text{donor}} - P_5(x_{\text{donor}})) * P_{95}(x_{\text{target}})}{P_{95}(x_{\text{donor}})} + P_5(x_{\text{target}}) \qquad (3.1)$$

Likewise, day-of-year values are transformed linearly so the median first warm and cold days match those of the target area, respectively. The date is converted to an integer from 1 (January 1) to 365 (December 31). A version of

Equation (3.1) is then used. Rather than values of $P_5$ and $P_{95}$, the median first cold and first warm days ($F_{\text{cold}}$ and $F_{\text{warm}}$, respectively) are matched. Before the median first warm day of the year for the donor area, we transform the date values using

$$\text{scaledDate} = \frac{\text{donorDate} * F_{\text{warm}}(\text{target})}{F_{\text{warm}}(\text{donor})}$$

Between the median first warm and cold days of the year, the transformation is

$$\text{scaledDate} = \frac{(\text{donorDate} - F_{\text{warm}}(\text{donor})) * (F_{\text{cold}}(\text{target}) - F_{\text{warm}}(\text{target}))}{F_{\text{cold}}(\text{donor}) - F_{\text{warm}}(\text{donor})} \\ + F_{\text{warm}}(\text{target})$$

Finally, between the median first cold day and the end of the year, we use

$$\text{scaledDate} = \frac{(\text{donorDate} - F_{\text{cold}}(\text{donor})) * (365 - F_{\text{cold}}(\text{target}))}{365 - F_{\text{cold}}(\text{donor})} \\ + F_{\text{cold}}(\text{target})$$

The transformed day numbers are then converted back into date values to be used in model training. An example of a day-of-year transformation from one area to match another is shown in Figure 3.2.

Next, we transform the data by applying a 17 parameter model (see

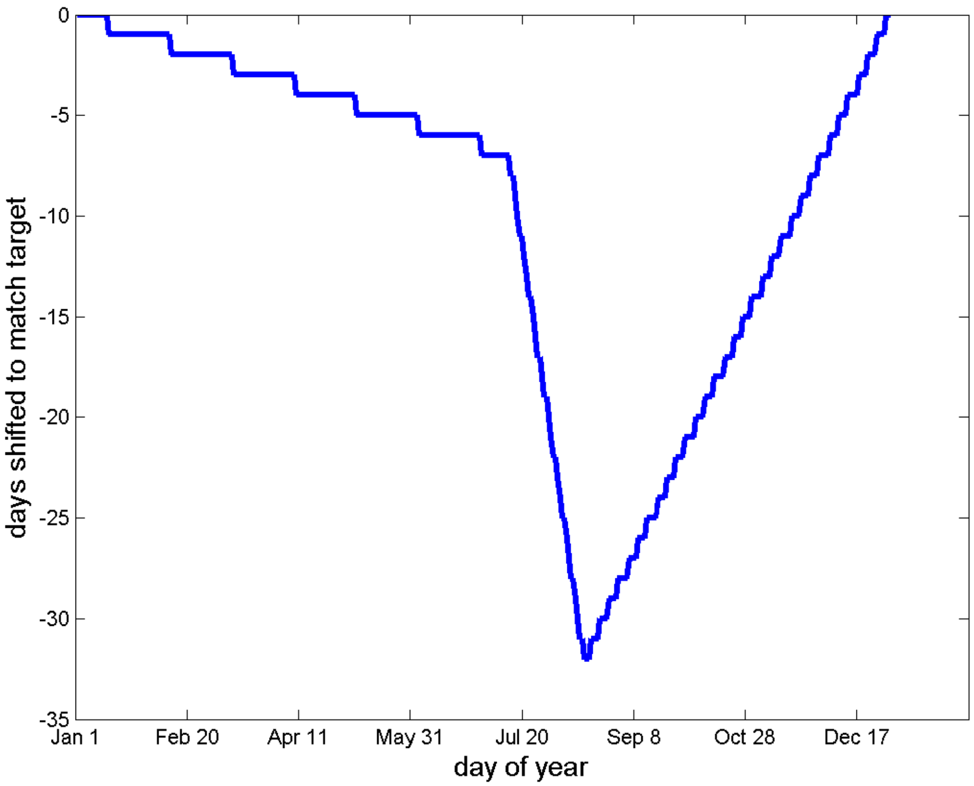

Figure 3.2: Number of days shifted in day-of-year transformation.

Table 3.1 for a complete list of terms),

$$\begin{aligned}\widehat{s} \;=\; & \beta_0 + \beta_1\text{HDD65} + \beta_2\text{HDD55} + \beta_3\text{CDD65} + \beta_4\Delta\text{MHDD} \\ & + \beta_5 \sin\left(\frac{2\pi\text{DOW}}{7}\right) + \beta_6 \cos\left(\frac{2\pi\text{DOW}}{7}\right) \\ & + \ldots \end{aligned} \tag{3.2}$$

Model 3.2 is fit to each set. By evaluating on each point, we can determine the value contributed by each variable in the model. Then we can scale these values by (target – donor) for each of the coefficients. This is performed on the donor data by multiplying each variable in Model 3.2 using the coefficients of the target area, rather than those of the scaled donor area.

Results of transforming data are shown in Table 3.1. After scaling by temperature and flow to match the 5% and 95% percentiles, the coefficients of the potential donor area are closer to those of the target area. After tuning based on our 17-parameter model, the coefficients match exactly. We also scale the day of the year values of the donor so the median first cold days and first warm days of the year match those of the target area. This is done because winters have different lengths and beginning dates with different climates seen across the country.

Data transformed in this way are more realistic than synthetic data, since they are not randomly generated. This preserves customer behavior characteristics and other properties of the unusual events for which we want improved forecasts. The transformations also preserve weather and flow variability that are difficult to simulate through random number generation.

### 3.2.2 Accounting for Volatility

We account for volatility by transforming data as surrogates that have similar coefficients of transformation from our 17-parameter model. This accounts for volatility by using surrogate data that exhibits similar patterns. For example,

Table 3.1: Surrogate data transformation model coefficients.

| Coefficient | Donor area before transformation | Donor area after temperature and flow scaling | Donor area after tuning with 17-parameter model | Target area |
|---|---:|---:|---:|---:|
| Constant | 15341.29 | 17723.42 | 18875.85 | 18875.85 |
| HDD65 | 830.61 | 569.37 | 582.77 | 582.77 |
| HDD55 | 1266.80 | 1280.68 | 1062.97 | 1062.97 |
| CDD65 | -21.82 | -23.70 | -30.24 | -30.24 |
| $\Delta$MHDD | -265.53 | -233.62 | -232.22 | -232.22 |
| sin(2 $\pi$ DOW/7) | 157.92 | 135.53 | -187.94 | -187.94 |
| cos(2 $\pi$ DOW/7) | -427.11 | -361.96 | -2704.41 | -2704.41 |
| sin(4 $\pi$ DOW/7) | 148.71 | 135.29 | -195.53 | -195.53 |
| cos(4 $\pi$ DOW/7) | -120.55 | -99.42 | -1065.36 | -1065.36 |
| sin(2 $\pi$ DOW/7) x HDD65 | 2.78 | 2.38 | 6.75 | 6.75 |
| cos(2 $\pi$ DOW/7) x HDD65 | -8.32 | -7.07 | -22.30 | -22.30 |
| sin(4 $\pi$ DOW/7) x HDD65 | -11.80 | -10.21 | -5.33 | -5.33 |
| cos(4 $\pi$ DOW/7) x HDD65 | -6.15 | -5.22 | -14.27 | -14.27 |
| sin(2 $\pi$ DOY/365) | 2796.92 | 2185.07 | 789.94 | 789.94 |
| cos(2 $\pi$ DOY/365) | 4178.69 | 3081.10 | 2258.32 | 2258.32 |
| sin(2 $\pi$ DOY/365) x HDD65 | 79.71 | 77.12 | 87.91 | 87.91 |
| cos(2 $\pi$ DOY/365) x HDD65 | 162.16 | 159.57 | 177.49 | 177.49 |

consider operating area A that is assumed to be identical to a different operating area B, with the exception that A is ten times the size of B. The coefficients for area A are 10 times the coefficients for area B. If we simply divide the data from area A by 10, we have the ideal set of surrogate data for area B. While we are not likely to find two areas that are so similar, we use nearest neighbor clustering techniques to reveal areas with similar coefficients.

GasDay accounts for volatility in forecasts by using ensemble models [26], which is also known as combining [16]. As Makridakis et al. explain [35], combining reduces forecast errors because techniques such as averaging reduce the volatility that is present in an individual model. For a given forecast, if one model is under-forecasting and one is over-forecasting, the resultant combined model is more accurate than either separately. Thus, volatility in the residual error is reduced through combining/ensembling.

### 3.2.3 Similarity Features and Donor Set Selection

There are a number of features of natural gas time series that we make surrogate data match via our transformation algorithm. They are explained in Table 3.2. However, there are a number of features that do not match when we transform surrogate data. Notable are temperature sensitivity and prior day weather sensitivity, which are explained in Section 3.2.4. Future research may consider methods of transforming data to match these features.

To determine which of the available data sets should be used as surrogates for a given target, we consider the values of the features discussed in Section 3.2.4. The features help us determine which data sets exhibit similar underlying relationships between weather variables and gas demand. The coefficients of transformation then allow the data to appear to have come from a different source. Future work may consider comparing the differences between the coefficients of the

Table 3.2: Features matched via our Surrogate Data Transformation Algorithm

| Feature | Brief Description | what we match |
|---|---|---|
| $s_k$ | natural gas flow for day $k$; amount of natural gas consumed that day, for that area | $5^{th}$- and $95^{th}$-percentile flows |
| Temperature | daily average temperature in °F used to calculate the HDD, CDD, and MHDD variables | $5^{th}$- and $95^{th}$-percentile temperatures; HDD reference temperature |
| Wind speed | Daily average wind speed in miles per hour | $5^{th}$- and $95^{th}$-percentile wind speeds |
| DOW | the day of the week (Sunday=1); used by itself and as a cross term with HDD65 | coefficients on 17+ parameter model |
| DOY | numeric day of the year | median first cold and first warm days of the year; coefficients on 17+ parameter model |

target and donor areas. Perhaps by normalizing these coefficients, we can extract more features for determining which donors will make good surrogates.

### 3.2.4 Natural Gas Time Series Features

Natural gas demand data sets exhibit some interesting characteristics that can be captured mathematically as attributes or features. Two such features are the Tenneti Index of Temperature Sensitivity [47] and Prior Day Weather Sensitivity.

The Tenneti Index of Temperature Sensitivity [47] is a quantitative measure that attempts to describe to what degree the weather variables temperature, wind,

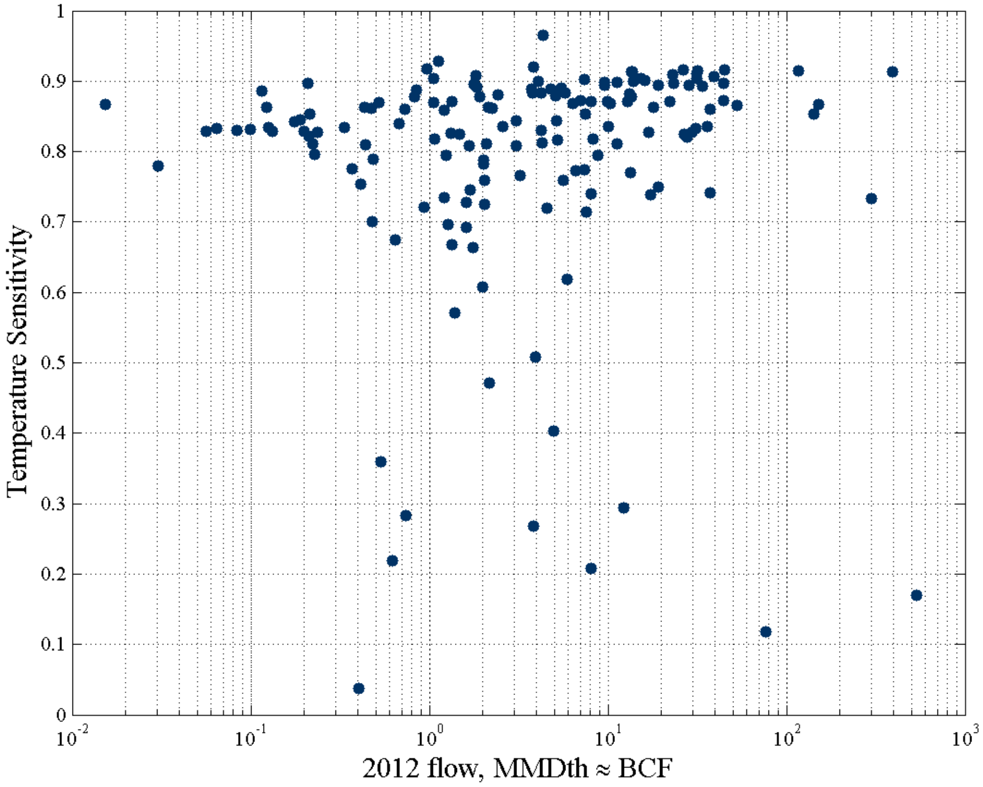

Figure 3.3: Tenneti Index of Temperature Sensitivity vs. aggregate annual flow.

and dew point impact natural gas demand forecasts. Tenneti's models were designed to quantify the impact of weather variables on natural gas demand models. The temperature sensitivity of an area ranges from 0 to 1. The greater the improvement by including temperature variables to the model, the closer the sensitivity is to one. Consider Figure 3.3. Operating areas around the United States have varying levels of sensitivity to weather variables, with the average being about 0.79. Similar to this sensitivity value is the F-test, which also is used to quantify the improvement of one model over another or the effects of some variables on the total [38].

We introduce the term “Prior Day Weather Sensitivity” to describe a similar quantitative measure. Rather than describing overall sensitivity of gas consumption in an area to weather factors, Prior Day Weather Sensitivity is a measure of how much consumption fluctuates day-to-day with change in temperature. Often, day-to-day natural gas use is not perfectly correlated with temperature fluctuations. This is because space heating is not the only use of natural gas and is partly due to consumer behavior [10]. It is also partly due to the thermodynamic lag effects [9] of home insulation and the thermal mass of a structure and its contents.

For example, consider the gas controller who, before the advent of computer models, drew a scatter plot of historical flow values to help predict the next day's flow. Domain knowledge led this controller to know that the next day flow is not a function of the next day's temperature alone. Rather, as Figure 3.4 illustrates, it is somewhere between no difference in flow and the expected difference due to the weather. In this situation, 0.3 was chosen as the Prior Day Weather Sensitivity. As Figure 3.4 shows, the red forecast point is 30% of the distance between today's forecast temperature and yesterday's point.

We can calculate an exact value of this constant for a given operating area. We do this by fitting a model to the historical data,

$$\widehat{s}_k = \beta_0 + \beta_1 \mathrm{HDD65}_k + \beta_2 \mathrm{HDD65}_{k-1}, \tag{3.3}$$

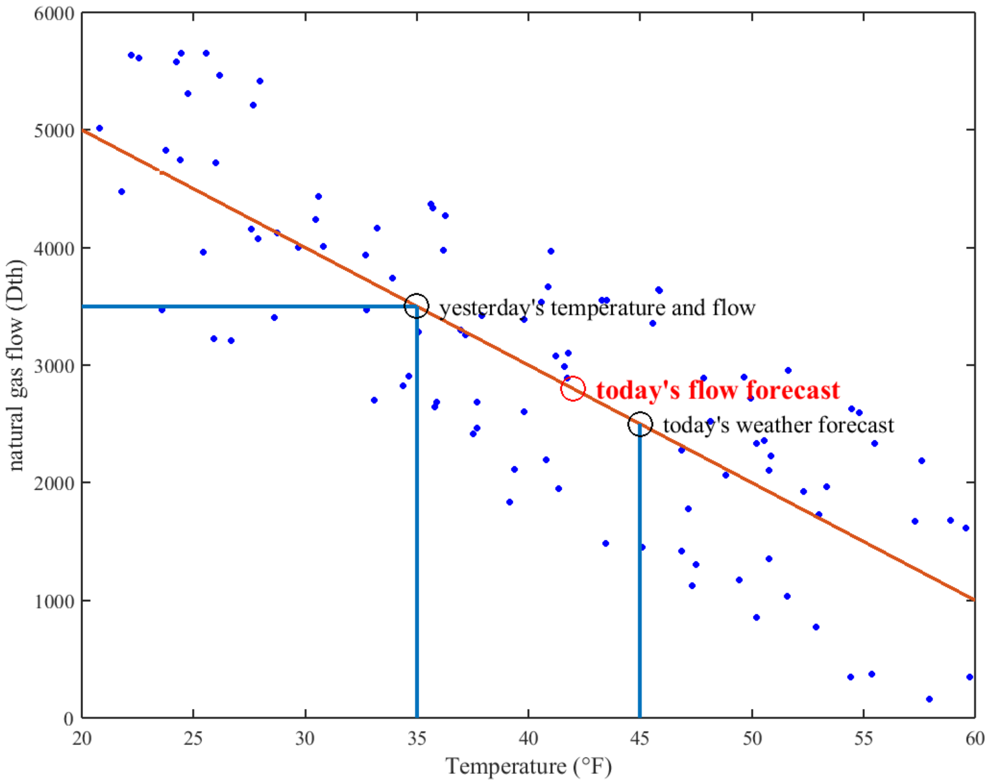

Figure 3.4: Prior Day Weather Sensitivity

which enables us to calculate

$$\text{Prior Day Weather Sensitivity} = -\frac{\beta_2}{\beta_1 + \beta_2}. \tag{3.4}$$

In principle, there is a risk of the denominator of Prior Day Weather Sensitivity being zero. If this occurs in one of GasDay's data sets, our code throws an exception. That did not happen in any of our experiments. Domain knowledge says that both coefficients will be greater than zero.

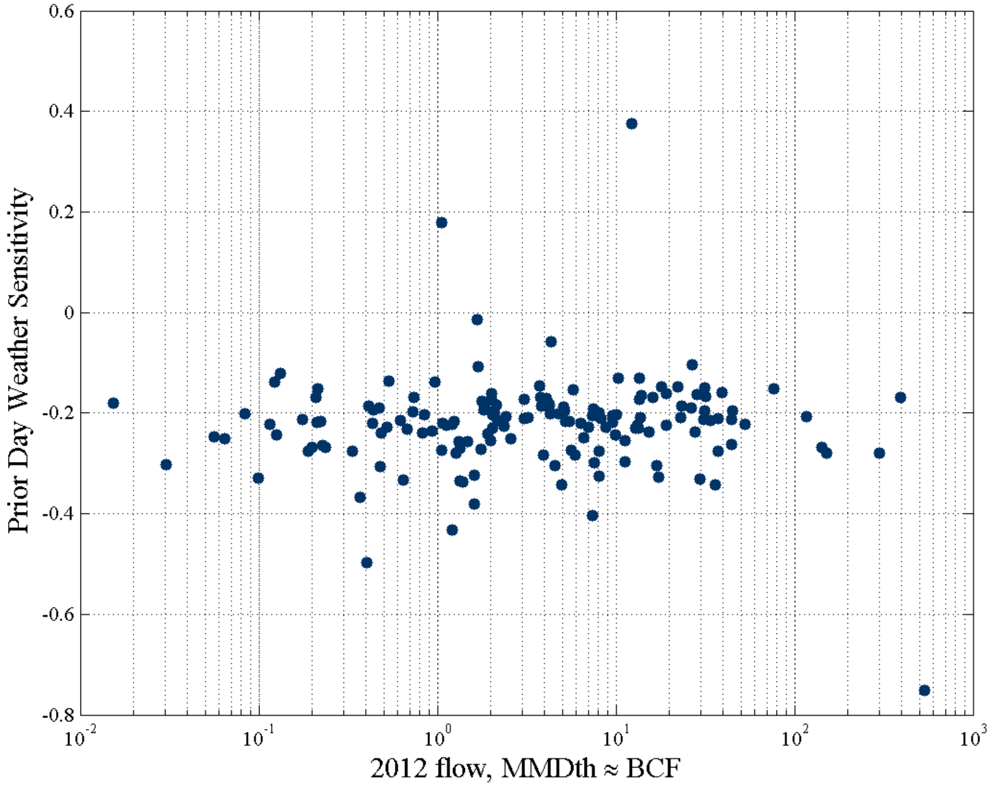

Figure 3.5: Prior Day Weather Sensitivity vs. aggregate annual flow.

Figure 3.5 plots the calculated value of Prior Day Weather Sensitivity by operating area against the annual flow in 2012 for that area. The average is approximately −0.22 for the operating areas in GasDay's database.

The Tenneti Index of Temperature Sensitivity and Prior Day Weather Sensitivity are two of many possible parameters that are used to characterize a GasDay data set. Parameters such as these are calculated to enable comparison of different data sets. We use a nearest neighbors clustering algorithm to determine which data sets are good surrogate data for building a model for a different area. In this research, we find the 12 nearest neighbors using the Tenneti Index of Temperature Sensitivity and Prior Day Weather Sensitivity. Future research may consider using other parameters for similarity and varying numbers of nearest neighbors.

The benefit of the features we use is that they may be generated in the same way for each available data set. Pseudocode for feature selection is shown in Algorithm 1: Surrogate Data Feature Selection.

**Algorithm 1** Surrogate Data Feature Selection

```
for each op area do
    if no weather data available then
        remove missing data range (shorten data set so there are no gaps)
    end if
    look at entire “good” (assumed clean) data set
    calculate (extract) features
    save results
end for
select 12 nearest neighbors as donor sets
```

We use these features to help determine which transformed surrogate sets will act as good surrogates for a target area. We then can use the transformed surrogate data to supplement available historical data in the target set. Since we are not trying to supplement available data for normal days, we specifically select data from unusual days (see Section 2.6) to use as surrogates. Once the original data and surrogates are aggregated to a form from which a model may be built, evaluation is performed as described in Section 3.2.5.

### 3.2.5 Evaluation Techniques

Many error metrics commonly are used in forecasting [16, 28]. We use MAPE and RMSE to evaluate the use of surrogate data in developing forecasting. We can determine appropriate validation methods by considering why we want to use surrogates.

### 3.2.6 Problems Forecasting Unusual Days

The goal of this research is improved forecasts for unusual events consisting of one or more days with patterns that have been seen at best rarely in historical data. These are often the days which gas transmission systems have been designed to handle at their peak. Brown [10] suggested using surrogate data from other operating areas to improve natural gas forecasting models for these uncommon days.

In real-world data, various factors contribute to noise in the data. These include inherent noise and measurement and reporting error. Chatfield [15, p. 217] talks about how outliers can cause major problems with data, and as such should be a focus of the initial examination of data (IDA). Akouemo [3], Kiware [32], and Tan et al. [46] discuss more reasons that data might require cleaning. This leads to outliers that in many cases fit the description of unusual events, as some outliers are particularly high or low values. We can distinguish outliers from unusual days by looking at the weather patterns [39]. We know that natural gas data has trends with respect to weather variables [14]. Unusual days will still reasonably follow this trend, while outliers will not.

We do not want to forecast outliers or have the outliers adversely affect our forecasts. To this end, we have assumed that the data has been cleaned, that suspected anomalous data have been detected and either removed or replaced with imputed data. The assumption of beginning with "clean" data is justified due to algorithms currently applied to GasDay data, e.g., [3, 4]. Cleaning removes *outliers*, so it is assumed that the desired unusual days data have not been removed. By nature, unusual days may seem like outliers, but the goal of previous work in this area is to remove true outliers. We acknowledge that by nature, these data sets will never be perfect, but the aforementioned previous work has proven to improve the accuracy of our models.

### 3.2.7 Validation of New Methods

As stated in the previous section, we want to forecast events unlike those in available past data.

It is difficult to validate our Universal Surrogate Algorithm, since we have not yet seen such events, and we do not know for sure when they might happen. As González-Rivera writes, “By their very nature, rare events occur very infrequently, but when they do occur, their consequences are catastrophic.” [25] In the natural gas industry, these rare events are often bitterly cold days. These are the days when residential customers want gas the most, so they can stay warm. Consequences of bad forecasts could include these customers being angry with their gas utilities, causing loss of customers, or perhaps lawsuits.

Since it is impractical to wait until a rare event occurs to see how our work performs, we have backcasted [5] events from the winter of 2013-2014. There were many cold events throughout the United States during that winter. By acting as if that winter had not yet occurred, we can see how our methods would have performed predicting these unusual events. This ex ante, out-of-sample testing enables us to verify that surrogate data works to forecast future events. We are able to demonstrate that adding surrogate data to the training set of data, forecasting model accuracy improves.

In this chapter, we discussed our methods of transforming data to act as surrogates and selecting a subset of the possible surrogate data to use. Results of validation of forecasting models tested with and without surrogate data are discussed in the next chapter.

# CHAPTER 4

## Evaluation of Our Improved Surrogate Data Methods

This chapter presents the results of our research into using surrogate data for improved forecasting models on days with unusual weather patterns. We illustrate results of the transformation algorithms described in the previous chapter. We then compare forecasts made with only the original raw data to forecasts made by augmenting the original data with surrogate data. This chapter illustrates that the methods explained in Chapter 3 improve forecasts on unusual days.

### 4.1 Transforming Surrogate Data

To illustrate the results of our transformation algorithm, we will consider an area near the East Coast. This area has about nine years of available historical data, which is less than many of the data sets in GasDay's databases. Hence, it is an ideal recipient of surrogate data to supplement what is available. Figure 4.1 shows the raw data for this target area (in red) and the first potential donor (in light blue), which is located in the Southwestern United States. The donor area has a smaller flow, and it does not get as cold as the target area. The donor area also has several years of historical data more than that of our initial target area.

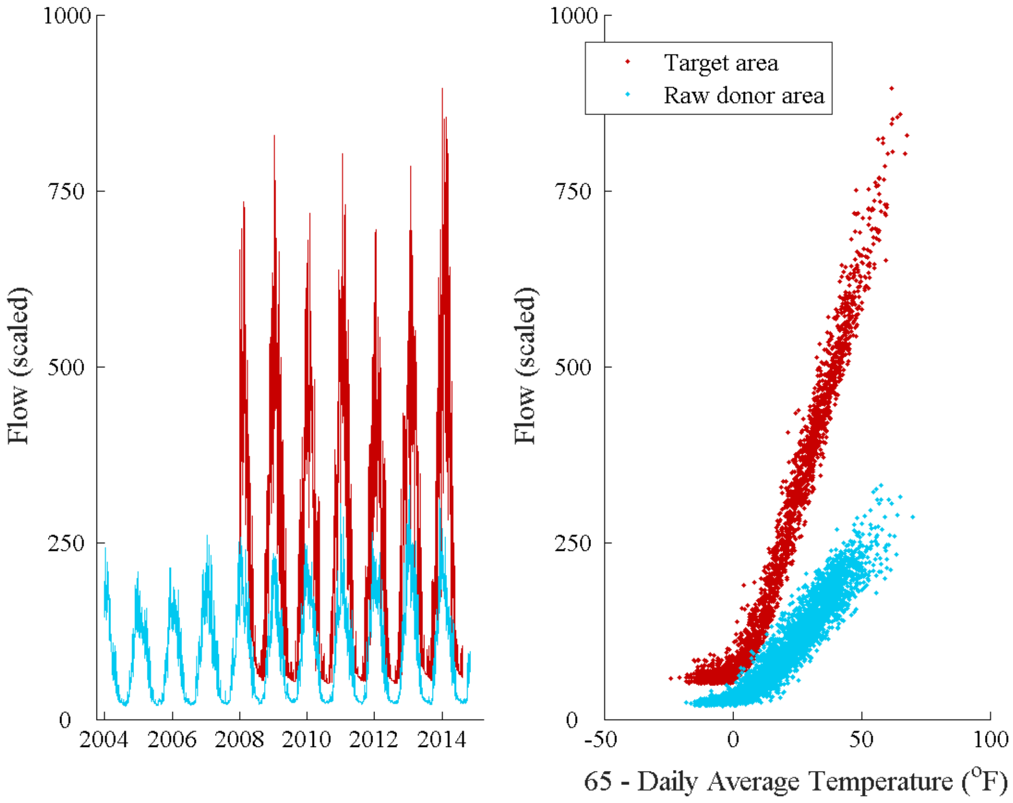

Figure 4.1: Target area from the East Coast and potential donor area from the Southwest.

Figure 4.2 shows the scaled donor area (in dark blue), which matches the target area's $5^{th}$- and $95^{th}$-percentile values for temperature and flow. The scaled donor area now fits in the same general window as the target area, but it shows more variation.

Figure 4.3 adds the fine-tuned transformed donor area data (in purple). This data looks much better; it almost could have originated in the target area (red) data set. It has been scaled to match the day-of-year characteristics of the target area,

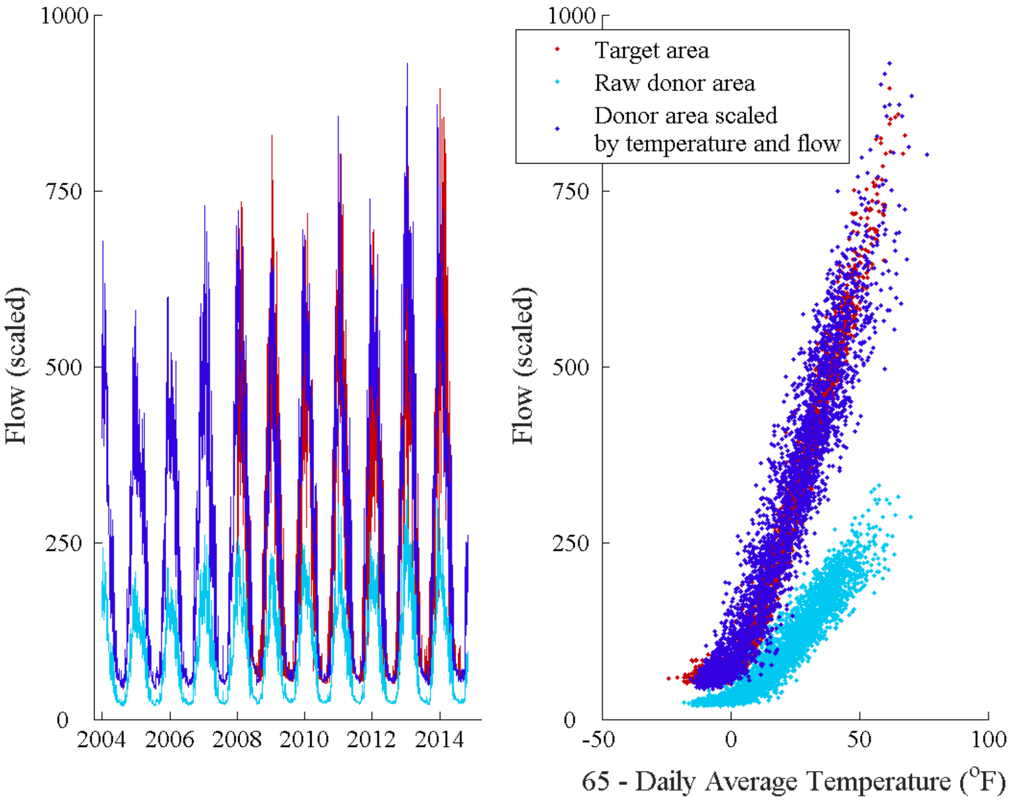

Figure 4.2: Target area, raw donor area from the Southwest, and donor area with temperature and flow scaled to match that of the target area.

and the coefficients of the 17-parameter linear regression model that was explained in Chapter 3 match that of the target area.

The next potential donor comes from the Midwestern United States. As the original data in Figure 4.4 shows, this area is much bigger and gets colder than our target area. Like the previous example, several more years of historical data are available from this area.

Figure 4.5 adds the temperature- and flow-scaled data (in dark blue). Once

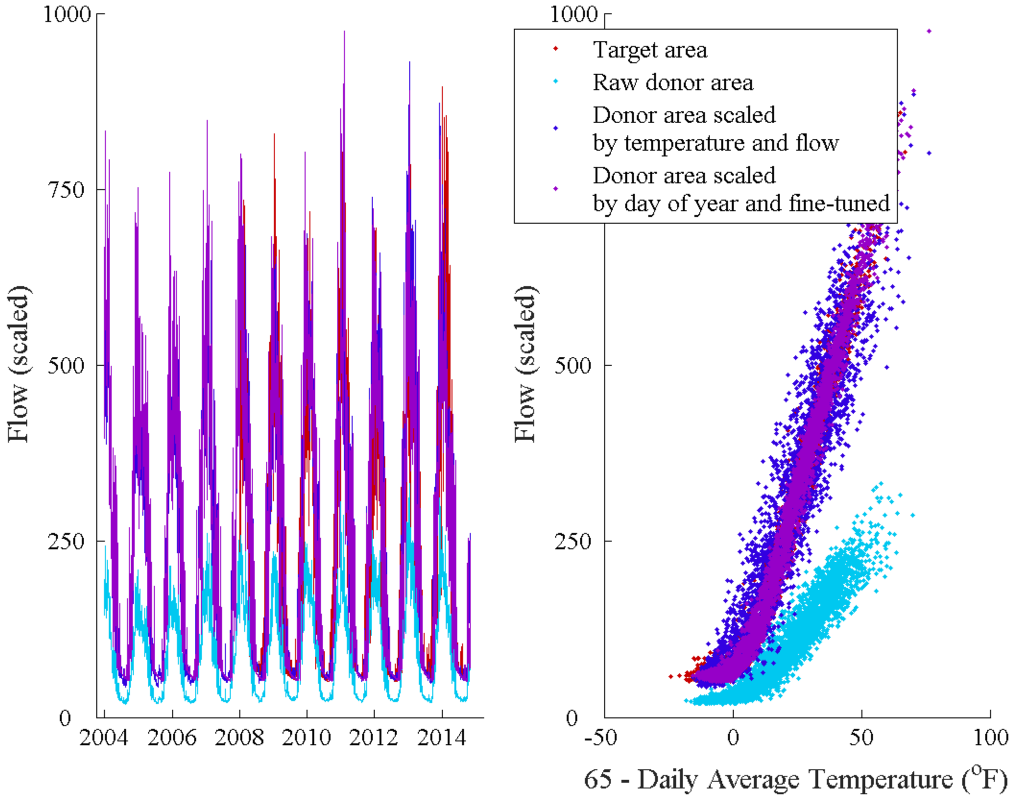

Figure 4.3: The day of year for the donor area has been scaled, and the data has also been fine-tuned with our 17-parameter model.

again, it has the same general range as the target, but still looks like it originated from a different source.

The fine-tuned donor data are added to the plot in Figure 4.6. The trend has clearly increased, bringing up the points on the coldest days (which appear at the far right). This data has been transformed to match the same 17-parameter model as the red data.

These two examples show how data from areas of the country with different

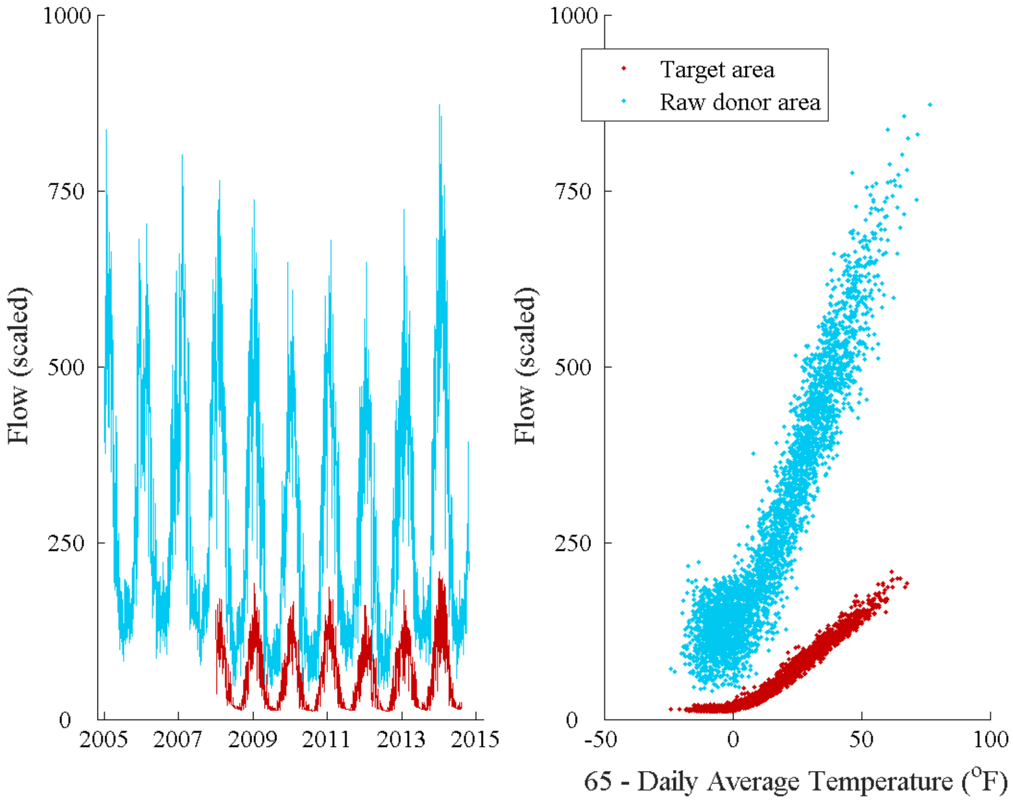

Figure 4.4: Target area from the East Coast and potential donor area from the Midwest.

climates and customer bases can be transformed successfully to match a given target area with different parameters. We use the transformation algorithm from Section 3.2, which ensures that the transformed data matches the target data perfectly with our specific model.

We now discuss how our forecasting models are used to determine how using this transformed data helps improve forecasting accuracy.

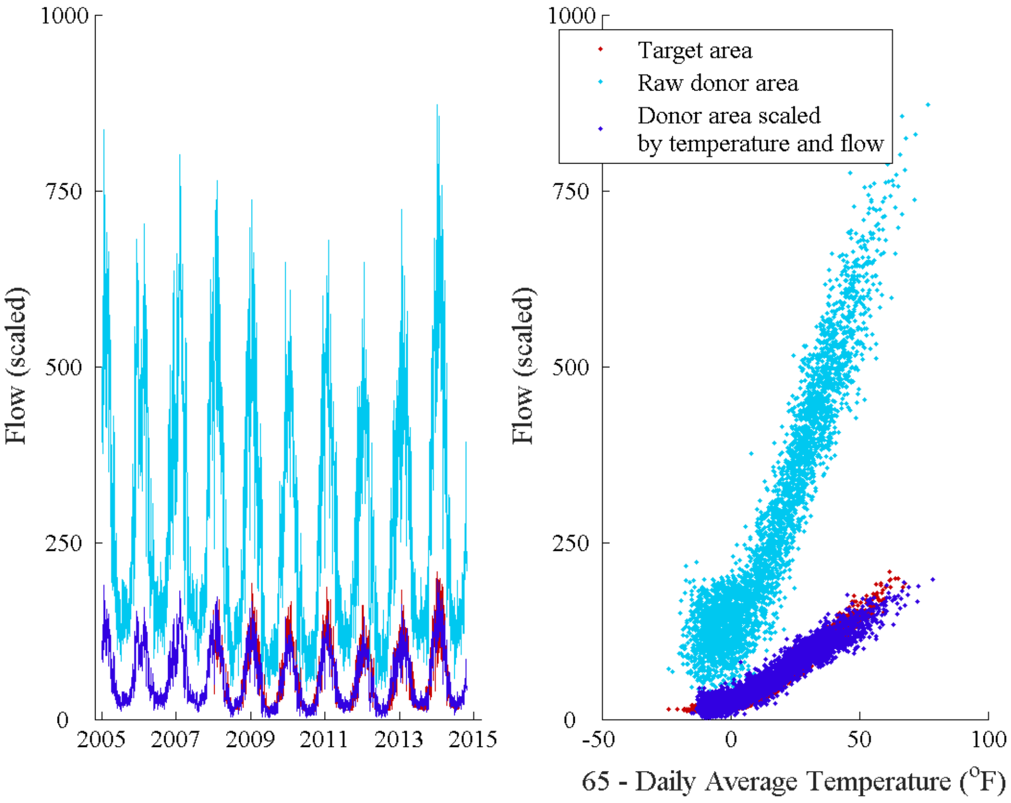

Figure 4.5: Target area, raw donor area from the Midwest, and donor area with temperature and flow scaled to match that of the target area.

## 4.2 Our Forecasting Model

We use GasDay's proprietary models to backcast natural gas demand from the winter of 2013-2014, both with and without surrogate data for training. These models consist of an ensemble of linear regression and artificial neural network models [52]; they are used at about 30 utilities around the United States to predict daily gas operations that comprise 17.7% of the country's natural gas consumed for residential, commercial, and industrial purposes. These models consist of an

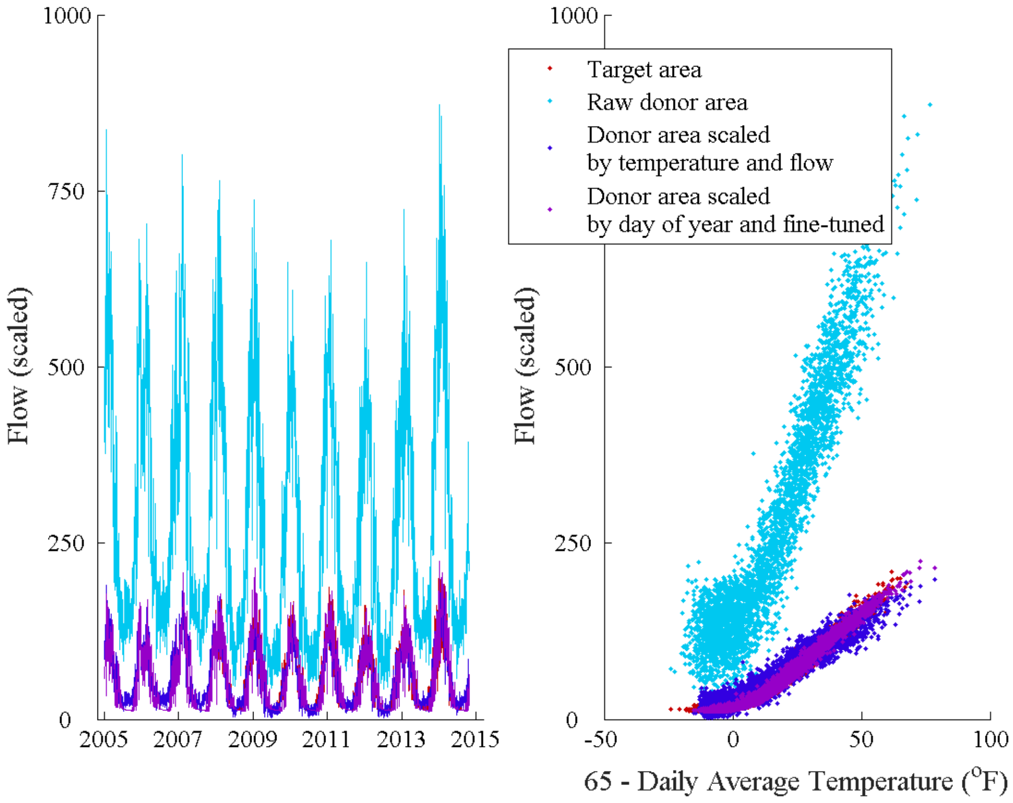

Figure 4.6: Target area from the East Coast and scaled and tuned donor area from the Midwest.

ensemble of linear regression and artificial neural network models [52]; they are used at about 30 utilities around the United States to predict daily gas operations that comprise 17.7% of the country's natural gas consumed for residential, commercial, and industrial purposes.

We present results from operating areas from different climates. Five areas from the midwest United States and one area from the East Coast were tested. We show results from one of the midwest areas, as results were consistent across the five.

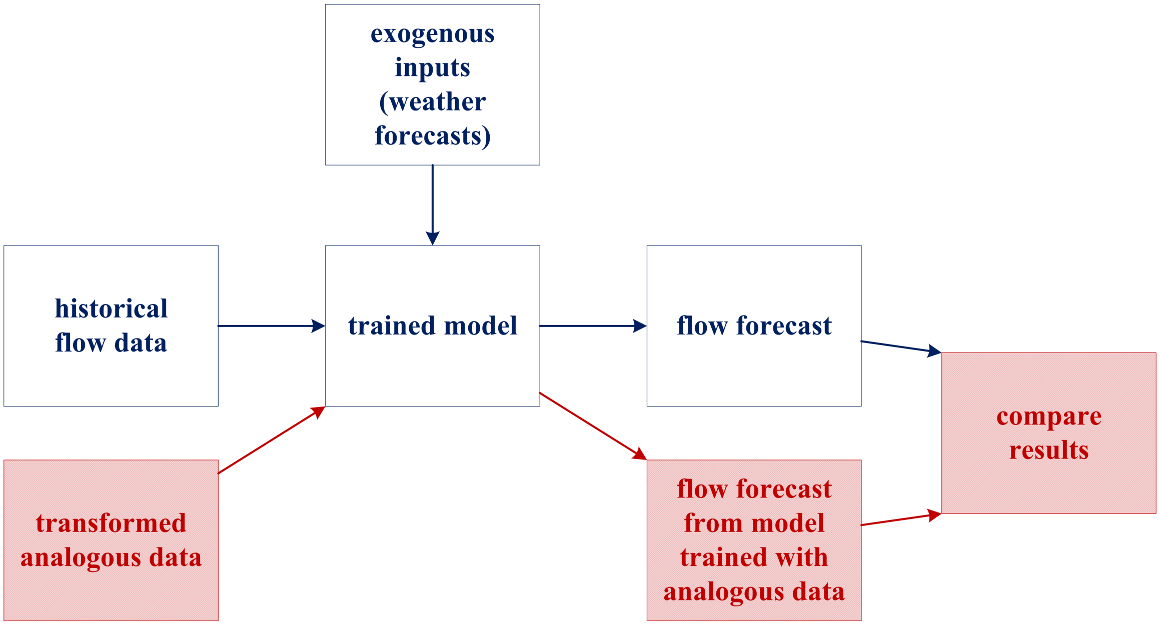
exogenous inputs (weather forecasts)
historical flow data
trained model
flow forecast
compare results
transformed analogous data
flow forecast from model trained with analogous data

Figure 4.7: Analogous data within the forecasting process. Our work adds the elements in red to the process in Figure 1.3.

We present forecasting results on the original data and show the improved results using our surrogate data algorithm to augment training data for unusual days.

## 4.3 Training and Testing Data

Ideally, we test our data on unusual weather events in the current winter. Practically, this is not a good strategy, since this would involve waiting for unusual events to occur. To mitigate this, we backtest on the past winter. This provides the benefit of enabling us to compare forecasts with the actual natural gas flow that was consumed on the range of dates. The winter of 2013–2014 was marked by several notable cold events in many areas of the United States. Since we now have forecasted and actual flow values for these events for many locations, that winter makes for ideal testing data.

### 4.4 Measuring Forecast Accuracy

As mentioned in Chapter 3, we want our forecasts to do well on unusual day types. We are good at forecasting typical days, as there are many in the database. If we show that surrogate methods result in better forecasts for unusual day types, we can use ensemble models to predict unusual events with surrogates.

We use several error metrics to gauge accuracy of linear models. Error metrics are compared for forecasts made by models trained with and without the surrogate data. The following are the error metrics used:

mean absolute percent error (MAPE):

$$\text{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{\hat{s}_k - s_k}{s_k}\right| \times 100$$

and root-mean-square-error (RMSE):

$$\text{RMSE} = \sqrt{\sum_{k=1}^{n}\frac{(\hat{s}_k - s_k)^2}{n}}$$

## 4.5 Analysis of Results

Most of GasDay's data sets have at least five years of data. While weather patterns change year-to-year, most of the data are in the normal or expected range of temperature. It is important that our forecasting model is able to model and predict accurately on these days. There is also less data for these areas. In this section, we show results and look specifically at forecasting results using surrogate data for these days with unusual weather patterns.

### 4.5.1 Results from an Operating Area in the Midwest

Figure 4.8 shows the results by month for an area in the U.S. Midwest. Results are shown for models built without surrogate data (dark blue bars), using the Brown Surrogate Algorithm that has chosen 12 surrogate donor sets via the genetic algorithm (brown bars), using the Brown Surrogate Algorithm to transform 12 surrogate donors that have been selected by our feature selection method (yellow bars), and using the Universal Surrogate Algorithm (turquoise bars). Figure 4.9 shows results for the same testing set, but plotted by the unusual day types mentioned in Section 2.6. Both surrogate data algorithms improve results across all days when compared with models trained using only the available historical data from this operating area. The Universal Surrogate Algorithm results are consistently worse than the Brown Surrogate Algorithm, however across all days and for most of the months and unusual day types, it outperforms models trained without using surrogate data.

Results from models trained using the Brown Surrogate Algorithm using surrogate donors selected using feature selection (yellow bars) validate our approach to selecting potential donor sets. Results using this method are approximately the same, and in some cases better than using the genetic algorithm of the Brown Surrogate Algorithm.

Note that the mean errors in both Figures 4.8 and 4.9 are consistently negative. (This is also true for the results in the next section.) We would expect that the mean error from the models would fluctuate between positive and negative values, and overall be close to zero. These results are average results across seven different neural network models. The consistent negative results indicate that the neural network models are under-forecasting. This may be because the 2013-2014 winter was noticeably colder than normal in the United States. GasDay's ensemble model, built into the GasDay software, is able to tune itself over time based on residual error and correct for models that consistently over- or under-forecast.

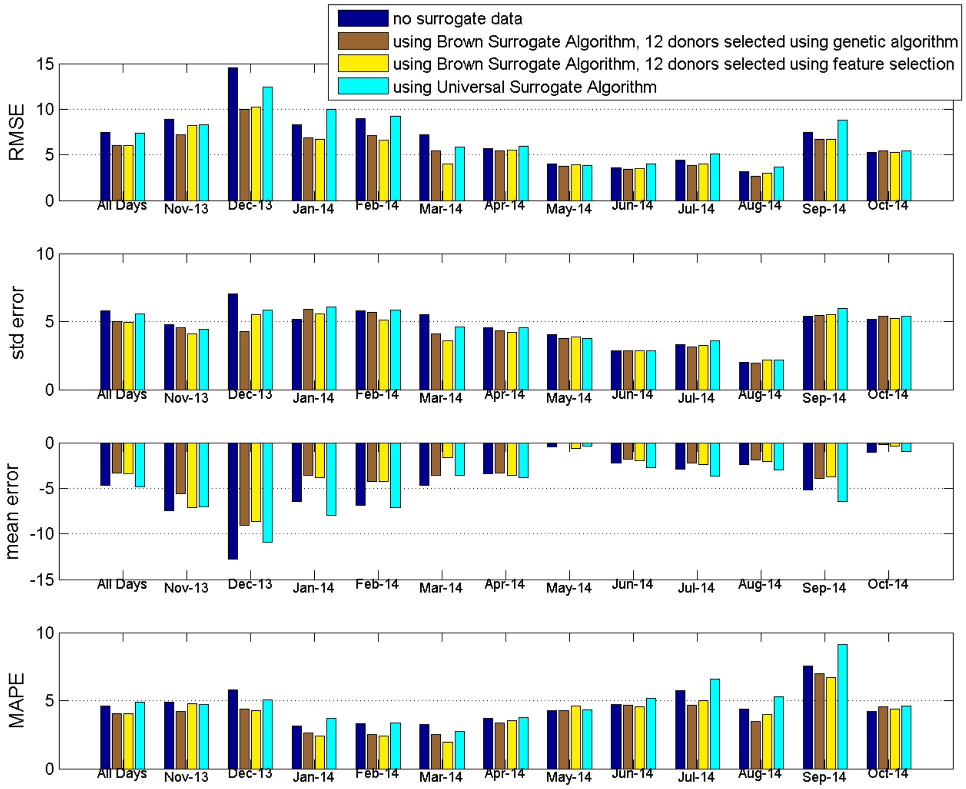

Figure 4.8: Forecast accuracy by month for an area in the Midwest.

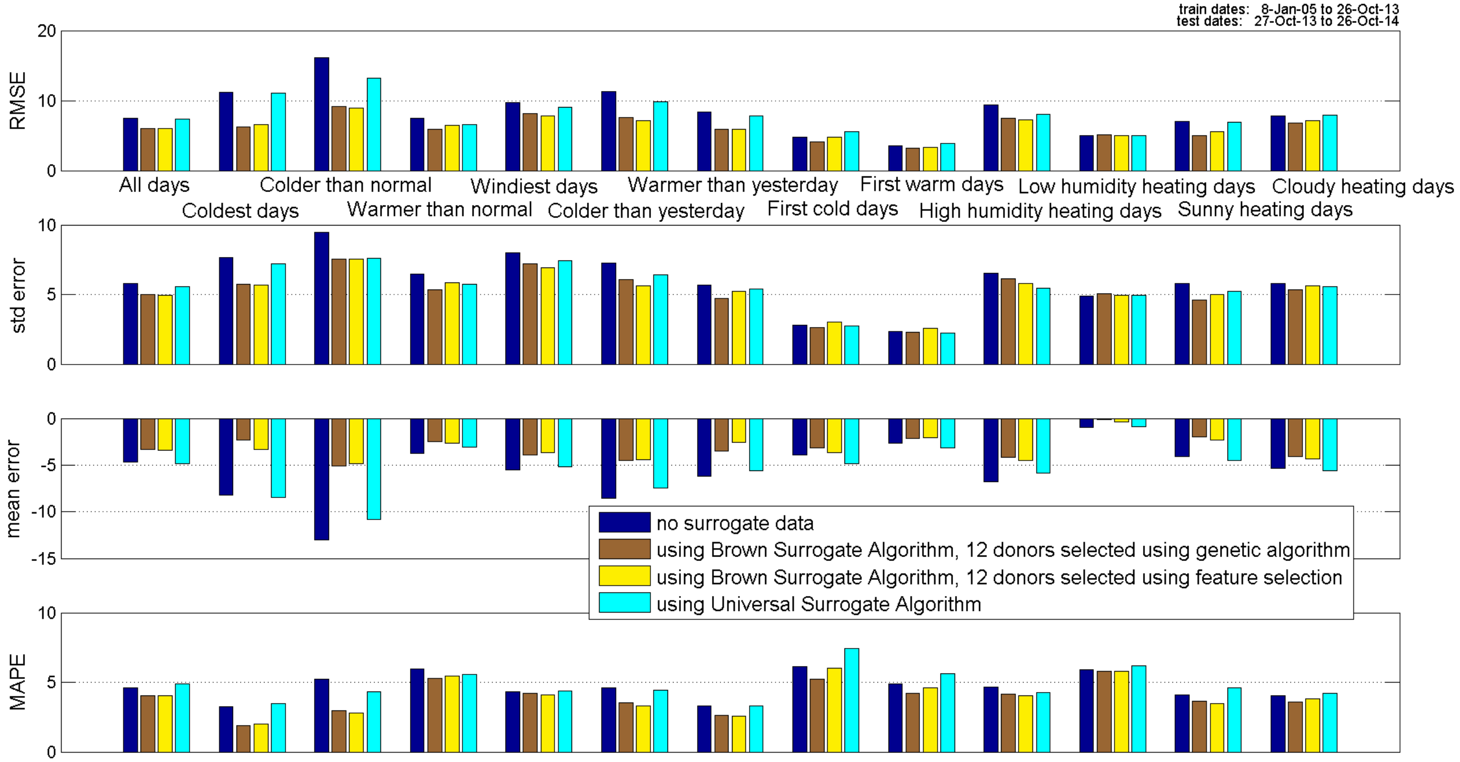

Figure 4.9: Error by Unusual Day type for an area in the Midwest.

### 4.5.2 Results from an Operating Area on the East Coast

Figures 4.10 and 4.11 show results by month and unusual day types, respectively, for an area on the East Coast. Models trained using surrogate data from Universal Surrogate Algorithm consistently perform better than models using the Brown Surrogate Algorithm, with the exception of February 2014. Models using the Universal Surrogate Algorithm also perform better on unusual day types.

The data for the baseline model in this case was produced from disaggregating [12] available data that was not at daily frequency. As a result, our transformed surrogate data has not been matched to available historical data, as this data does not exist. Rather, surrogate data has been transformed to match parameters of the operating area. While the artificial historical data appears to generate better models, these results illustrate another benefit of surrogate data: it can be used when historical data is only available in a lower frequency than desired for forecasting models.

As with the area in the Midwest, this test validates the donor selection method of the Universal Surrogate Algorithm. Selecting potential donors based on features is more efficient than a genetic algorithm, and works as well (if not better) in this case.

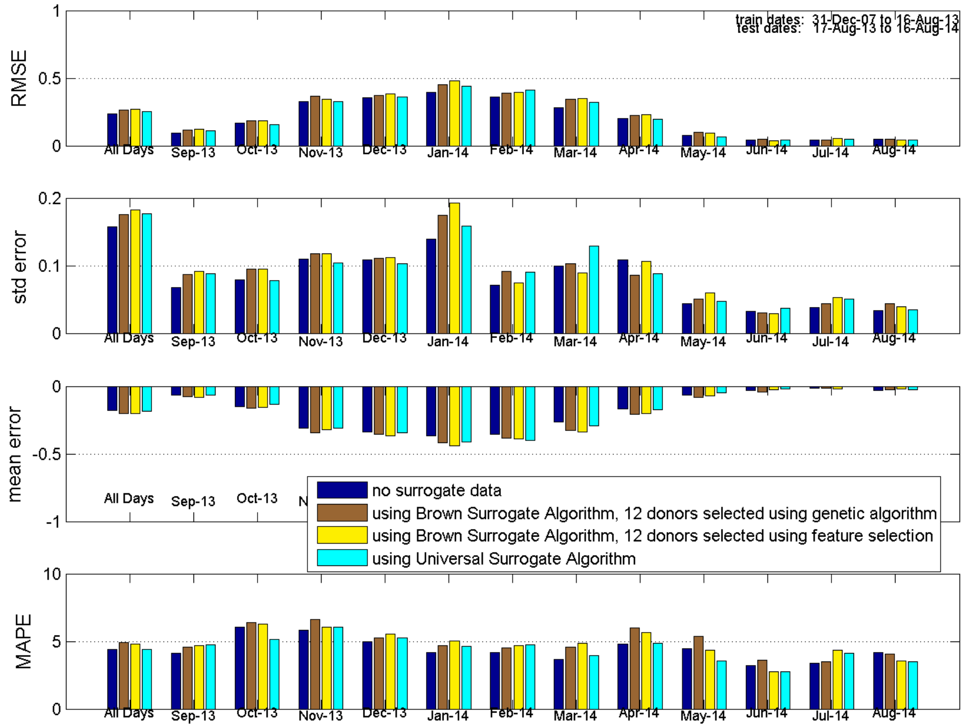

Figure 4.10: Forecast accuracy by month for an area on the East Coast.

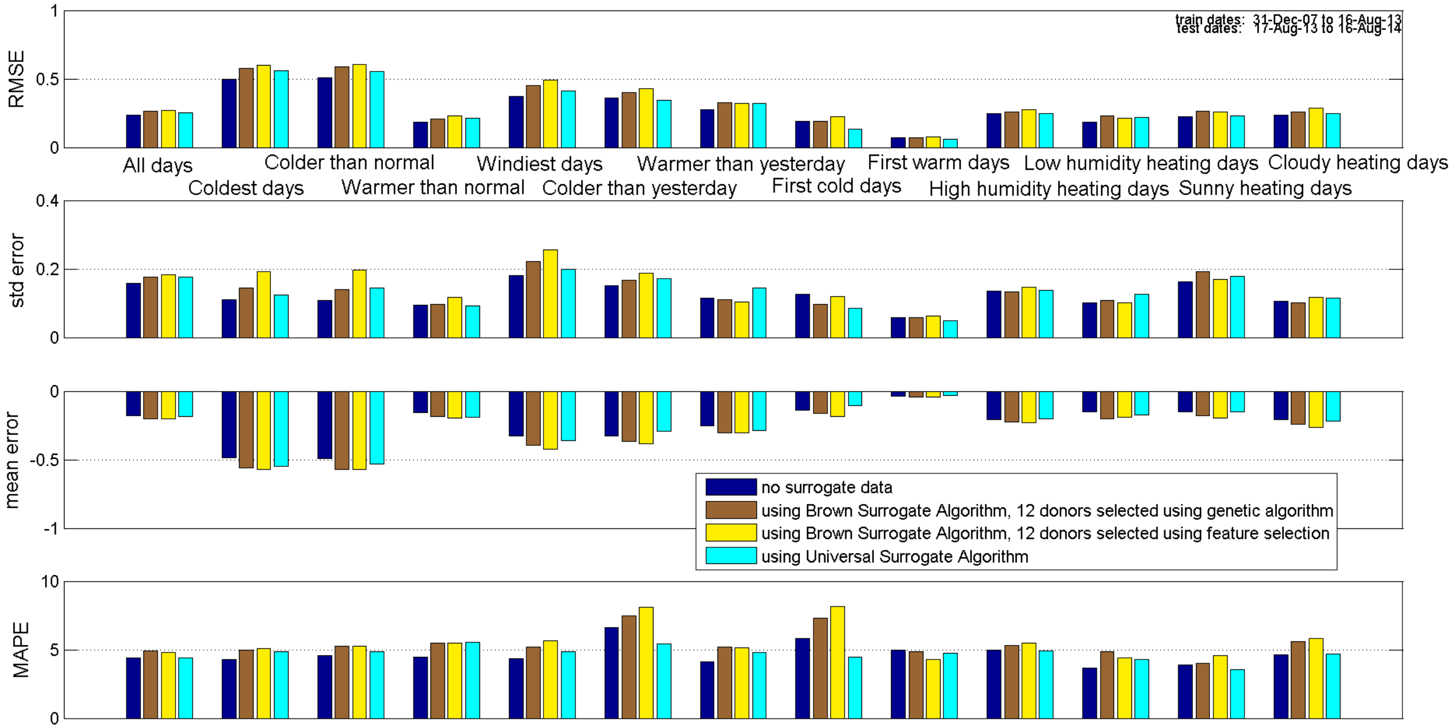

Figure 4.11: Error by Unusual Day type for an area on the East Coast.

## 4.6 Concluding Remarks

Using surrogate data improves forecast accuracy for days with unusual weather types. In some cases, the resulting forecasts are worse than models trained without surrogate data, but forecast accuracy is almost always close or better using surrogates. This affirms that surrogate data should be used to increase forecast accuracy, especially on unusual day types. Surrogate data should be used, as it is the best available method for solving the problem of insufficient data when building forecasting models.

The modifications present in the Universal Surrogate Algorithm do not consistently improve forecast accuracy across all operating areas tested. Future research should investigate why transforming day-of-year variables does not improve results. Our research does validate our improved method of donor selection. Selecting potential donors using feature selection requires less computational resources and works as well as a genetic algorithm.

# CHAPTER 5

## Benefits of Analogous Data and Future Considerations

This chapter discusses conclusions from our research. Future work possibilities and suggestions are presented.

One of the benefits of our research is a greater widespread use of analogous data in GasDay models. Rather than having to run a genetic algorithm to determine good surrogate donors for a given area, our methods enable the use of surrogate data for any area to be determined with significantly less computational time involved. Our use of feature selection and improved transformation algorithms allow for better use of surrogate data.

### 5.1 Contributions of Our Research

This research shows that surrogate data is a reasonable technique for enhancing forecasts where available data are insufficient. If there is a short historical range, or the historical data only includes mild weather conditions, the use of surrogate data may improve forecast accuracy. This is especially true for days with unusual weather patterns.

Our transformation algorithms are effective for target operating areas from places with vastly different climates. Because the algorithm transforms based on weather conditions and features of the target set, it is robust in its ability to make donor data match characteristics from vastly different places. The end product is surrogate data that is viable for supplementing available data in the target data set.

Our work also illustrates that surrogate data may be used in the case where the original data are only available at a lower frequency than desired. Available data can be disaggregated [12], and the parameters of the disaggregated data may be used to transform analogous data.

## 5.2 Improvements on the Universal Surrogate Algorithm

Our method outputs a list of potential donor sets, ranking how similar they are to the target set. When we transform $n$ surrogate donors, they are selected based on this ranking. A logical next step would be to develop a score for how well a given donor set may work when transformed to match a target area. Likewise, we should consider using other error metrics. MAPE and RMSE are commonly used in the field of energy forecasting. However as Hyndman and Koehler [28] mention, metrics such as Mean Absolute Error (MAE) are recommended in forecasting due to being less sensitive to outliers. MAE and RMSE are metrics that have the same units as the residuals, which means that the dollar value of the forecasting error is much more easy to quantify than with MAPE, for example.

In addition to these improvements, our work has revealed several directions of future research, which will be explained in the next section.

## 5.3 Future Research

Our work presents many opportunities for further research. Some domain knowledge from natural gas forecasting can be applied to electricity demand forecasting. It may be the case that surrogate gas data can help train electric demand models (or *vice-versa*). Also, similar techniques applied here may be applied to surrogate data used in electricity forecasting. For example, electricity demand also has unusual days in both the winter and the summer, since electricity is used for both air conditioning and heating.

We use the HDDW65 variable, as natural gas consumption is higher in the winters due to space heating. Including a cooling degree day term adjusted for wind may enhance models used to transform surrogate data in the electricity load forecasting domain. As both natural gas and electricity loads are highly dependent on weather, there may be some value in transforming natural gas data to enhance electricity demand forecasting models. Further research is needed.

González-Rivera writes, “The good news is that the ‘rare event’ is predictable; at the very least, it can be measured probabilistically” [25]. It may be possible to generate a canonical probabilistic model of rare events in natural gas consumption data. This model could be tailored to specific areas based on different consumption parameters. We could then use this model to generate probabilistic surrogate data, which would help predict one-in-$n$ events. Likewise, a probabilistic model of natural gas consumption could improve disaggregation algorithms.

An application of this work which could leverage the concepts of hierarchical models mentioned in the literature [19, 27] would be to develop a pooled model, perhaps for an entire state, region, or country. Lower levels of models would eventually get to the level of the operating areas focused on in our work. The benefits of the hierarchical model include filling in gaps where data may be lacking and modeling trends that may be seen in other areas of a given state, or similar states. This concept has been discussed previously by Brown [8].

Bootstrapping [21] is a technique based on the statistical concept of the jackknife. These ideas are similar to our method of surrogate data. Essentially, repeated random samples can be used to supplement available data. Rather than picking similar areas, one could perhaps bootstrap samples of areas and build a distribution of unusual events for a given utility or operating area.

GasDay trains neural network models as part of their forecasting package [52]. There may be benefit to increasing the number of epochs or neurons in these models due to the increased amount of data available with our surrogate models.

We transform data with the intent of matching a number of features that have been found to impact natural gas demand. Future research can explore transforming data to match other features, as was mentioned in Section 3.2.3. Future work may also match these features after transformations, as was mentioned in Section  3.2.3. We also scale to match the $5^{th}$- and $95^{th}$-percentile values of temperature and flow. Rather, we could match the one-in-$n$ threshold temperature and flow values [18] in our scaling. We could also select a number of transformed one-in-$n$ data to be used as surrogates, in addition to the unusual days data.

We showed that our models using surrogate data result in better forecasts on days with unusual weather patterns. We suggest that a rule-based ensemble model [1] takes this into account. Surrogate data can be a great help for multi-day events with unusual weather patterns. An ensemble model could switch off surrogate data when it might not be as much help. In this way, we could use surrogate data in such a way that it does not make our forecasts work, but is only used where it is certain to improve forecasts.

Our work implicitly leverages the fact that different winters have different characteristics. Some winters are milder, but have a few notable bitterly cold events. Contrarily, there are winters that are consistently much colder than normal but may not see cold events that one would expect to happen less frequently than once per year. It would be beneficial to have a way to characterize winters and classify them. If the upcoming winter is expected to be of a certain type, surrogate data may be drawn from similar type winters that have been seen in the past or in different areas of the country.

We scale our data sets based on temperature, wind, gas flow, and day of the year. There are other variables used in our natural gas forecasting algorithms, and there may be benefit in exploring scaling these variables as well. For example, there may be a way to scale the Prior Day Weather Sensitivity feature mentioned in Section 3.2.4.

## 5.4 Conclusion

In conclusion, the problem of insufficient data can be addressed by transforming data of a similar nature from other sources. The results shown in Chapter 4 show that natural gas demand data sets from operating areas around the country exhibit similar trends that may be successfully transformed and used to supplement available data from a target area. This work also reveals several notable possibilities for future research.

# BIBLIOGRAPHY

[1] M. Adya, F. Collopy, J. S. Armstrong, and M. Kennedy, “Automatic identification of time series features for rule-based forecasting,” *International Journal of Forecasting*, vol. 17, no. 2, pp. 143–157, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207001000796

[2] M. Akkurt, O. F. Demirel, and S. Zaim, “Forecasting Turkey’s natural gas consumption by using time series methods,” *European Journal of Economic and Political Studies*, vol. 3, no. 2, 2010. [Online]. Available: http://ejeps.fatih.edu.tr/docs/articles/108.pdf

[3] H. N. Akouemo, “Data cleaning in the energy domain,” Ph.D. dissertation, Marquette University, Department of Electrical and Computer Engineering, 2015.

[4] H. N. Akouemo and R. J. Povinelli, “Time series outlier detection and imputation,” in *IEEE Power and Energy Society 2014*, July 27-31 2014, pp. 1–5.

[5] J. S. Armstrong, “Long-range forecasting for a consumer durable in an international market,” Ph.D. dissertation, 1968. [Online]. Available: http://www.researchgate.net/publication/228120297_Long-Range_Forecasting_for_a_Consumer_Durable_in_an_International_Market

[6] P. Balestra and M. Nerlove, “Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas,” *Econometrica*, vol. 34, no. 3, pp. 585–612, July 1966. [Online]. Available: http://www.jstor.org/stable/1909771

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, ser. Holden-Day Series in Time Series Analysis. San Francisco: Holden-Day, 1970.

[8] R. H. Brown, “Domain knowledge that drove the design of our surrogate data,” presented at GasDay Seminar, July 11, 2013, Milwaukee, WI.

[9] ——, "In search of the hook equation: Modeling behavioral response during bitter cold events," in *2014 Gas Forecasters Forum*, October 16, 2014.

[10] ——, "Research results: The heck-with-it hook and other observations," in *Southern Gas Association Conference: Gas Forecasters Forum*, October 16 2007.

[11] R. H. Brown, D. Clark, G. F. Corliss, F. Nourzad, T. Quinn, and C. Twetten, "Forecasting natural gas demand: the role of physical and economic factors," in *32nd Annual International Symposium on Forecasting*, 2012. [Online]. Available: http://www.forecasters.org/proceedings12/QUINN_THOMAS_ISF2012_2012-07-16.pdf

[12] R. H. Brown, P. E. Kaefer, C. R. Jay, and S. R. Vitullo, "Forecasting natural gas design day demand from historical monthly data," in *Proceedings of the Pipeline Simulation Interest Group 2014 Conference*, May 6-9 2014.

[13] R. H. Brown, I. Matin, P. Kharouf, and L. P. Piessens, "Development of artificial neural network models to predict daily gas consumption," *American Gas Association Forecasting Review*, vol. 5, pp. 1–22, March 1996.

[14] R. H. Brown, S. R. Vitullo, G. F. Corliss, M. Adya, P. E. Kaefer, and R. J. Povinelli, "Detrending daily natural gas consumption series to improve short-term forecasts," in *IEEE Power and Energy Society 2015 Conference*, July 26-30 2015.

[15] C. Chatfield, "The initial examination of data," *Journal of the Royal Statistical Society. Series A (General)*, vol. 148, no. 3, pp. 214–253, 1985. [Online]. Available: http://www.jstor.org/stable/2981969

[16] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207006000021

[17] P. M. Dixon, A. M. Ellison, and N. J. Gotelli, "Improving the precision of estimates of the frequency of rare events," *Ecology*, vol. 86, no. 5, pp. 1114–1123, May 2005. [Online]. Available: http://www.jstor.org/stable/3450872

[18] A. D'Silva, "Estimating the extreme low-temperature event using nonparametric methods," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, 2015.

[19] G. T. Duncan, W. L. Gorr, and J. Szczypula, "Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting," *Management Science*, vol. 39, no. 3, pp. pp. 275–293, Mar. 1993. [Online]. Available: http://www.jstor.org/stable/2632644

[20] ——, *Forecasting Analogous Time Series*, ser. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, 2001, pp. 195–213.

[21] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann.Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979. [Online]. Available: http://dx.doi.org/10.1214/aos/1176344552

[22] Federal Energy Regulatory Commission staff, "Gas-electric coordination quarterly report to the commission," Federal Energy Regulatory Commission, Docket No. AD12-12-000, June 19, 2014. [Online]. Available: http://www.ferc.gov/legal/staff-reports/2014/06-19-14-gas-electric-cord-quarterly.pdf

[23] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide.* John Wiley & Sons, September 2008.

[24] A. Galka, *Topics in Nonlinear Time Series Analysis: with Implications for EEG Analysis.* River Edge, N.J.: World Scientific, 2000.

[25] G. González-Rivera, "Predicting rare events: Evaluating systemic and idiosyncratic risk," *International Journal of Forecasting*, vol. 30, no. 3, pp. 688–690, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207014000235

[26] J. Gramz, "Using evolutionary programming to increase the accuracy of an ensemble model for energy forecasting," Master's thesis, Marquette University. Department of Electrical and Computer Engineering, Milwaukee, WI, 2014.

[27] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357–363, 2014. [Online].

Available:
http://www.sciencedirect.com/science/article/pii/S0169207013000745

[28] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006. [Online]. Available:
http://www.sciencedirect.com/science/article/pii/S0169207006000239

[29] P. E. Kaefer, B. Ishola, R. H. Brown, and G. F. Corliss, “Using surrogate data to mitigate the risks of natural gas forecasting on unusual days,” in *35th Annual International Symposium on Forecasting*, 2015.

[30] A. Kani, M. Abbaspour, and Z. Abedi, “Estimation of natural gas demand in industry sector of Iran: A nonlinear approach,” *International Journal of Economics and Finance*, vol. 5, no. 9, pp. 148–155, 2013.

[31] A. Khotanzad, Elragal, and T.-L. Lu, “Combination of artificial neural-network forecasters for prediction of natural gas consumption,” *Neural Networks, IEEE Transactions on*, vol. 11, no. 2, pp. 464–473, 2000. [Online]. Available:
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=839015

[32] S. Kiware, “Detection of outliers in time series data,” Master’s thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, 2010. [Online]. Available:
http://search.proquest.com/docview/193658871

[33] S. Lv, X.-N. Zhou, Y. Zhang, H.-X. Liu, D. Zhu, W.-G. Yin, P. Steinmann, X.-H. Wang, and T.-W. Jia, “The effect of temperature on the development of *Angiostrongylus cantonensis* (Chen 1935) in *Pomacea canaliculata* (Lamarck 1822),” *Parasitology research*, vol. 99, no. 5, pp. 583–587, October 2006. [Online]. Available: http://dx.doi.org/10.1007/s00436-006-0198-8

[34] B. Mak, J. T. Kwok, and S. Ho, “Kernel eigenvoice speaker adaptation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 984–992, 2005.

[35] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, “The accuracy of extrapolation (time series) methods: Results of a forecasting competition,” *Journal of Forecasting*, vol. 1, no. 2, pp. 111–153, 1982. [Online]. Available:
http://onlinelibrary.wiley.com/doi/10.1002/for.3980010202/abstract

[36] T. H. McCurdy and I. G. Morgan, "Tests of the martingale hypothesis for foreign currency futures with time-varying volatility," *International Journal of Forecasting*, vol. 3, p. 131, 1986. [Online]. Available: http://ssrn.com/abstract=2012681

[37] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning.* Cambridge, MA: MIT Press, 2012.

[38] R. L. Ott and M. T. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, sixth edition ed. Belmont, CA: Brooks/Cole, Cengage Learning, 2010.

[39] B. Pang, "The impact of additional weather inputs on gas load forecasting," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, 2012. [Online]. Available: http://search.proquest.com/docview/1034448491

[40] J. L. Piggott, "The METGAS system: A computerised system for receiving, storing and using weather forecasts to predict short-term gas demand," *Gas Engineering and Management*, vol. 22, no. 7/8, pp. 303–312, July/August 1982.

[41] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D*, vol. 142, pp. 346–382, 1999. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.3999

[42] L. Silva, "A feature engineering approach to wind power forecasting: GEFCom 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 395–401, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207013000836

[43] V. I. Smirnov, *Linear Algebra and Group Theory*, R. A. Silverman, Ed. New York: Dover Publications, 1961.

[44] B. Soldo, "Forecasting natural gas consumption," *Applied Energy*, vol. 92, pp. 26–37, April 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306261911006982

[45] R. Swope and W. S. Howell, *Trading by Numbers: Scoring Strategies for Every Market.* Hoboken, N.J.: John Wiley & Sons, 2012.

[46] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Boston: Pearson Education, Inc., 2006.

[47] S. Tenneti, “Identification of nontemperature-sensitive natural gas customers and forecasting their demand,” Master’s thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, May 2009.

[48] J. Theiler and D. Prichard, “Using “surrogate surrogate data” to calibrate the actual rate of false positives in tests for nonlinearity in time series,” *Fields Institute Communications*, vol. 11, pp. 99–112, 1997. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.3434

[49] R. J. Thomas, “Estimating market growth for new products: An analogical diffusion model approach,” *Journal of Product Innovation Management*, vol. 2, no. 1, pp. 45–55, March 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0737678285900153

[50] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR ’91., IEEE Computer Society Conference on*, 1991, pp. 586–591. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=139758

[51] U.S. Energy Information Administration, “Global natural gas consumption doubled from 1980 to 2010,” April 12, 2012, accessed 6/2/2014. [Online]. Available: http://www.eia.gov/todayinenergy/detail.cfm?id=5810

[52] S. R. Vitullo, R. H. Brown, G. F. Corliss, and B. M. Marx, “Mathematical models for natural gas forecasting,” *Canadian Applied Mathematics Quarterly*, vol. 17, no. 4, pp. 807–827, Jan. 2009.

[53] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008. [Online]. Available: http://dx.doi.org/10.1007/s10115-007-0114-2