Using Surrogate Data to Mitigate the Risks of Natural Gas Forecasting on Unusual Days

Paul E. Kaefer, Babatunde Ishola, George F. Corliss, and Ronald H. Brown
GasDay™ Project, Marquette University, Milwaukee, WI

***Abstract*—Energy utilities see higher risk when forecasting for their operating areas (service territories) on days that are high-demand or difficult to forecast. These days often have unusual weather patterns (e.g., colder than normal or significant temperature fluctuation from previous days). Due to their unusual nature, data describing these days are scarce. We present a method that successfully transforms natural gas consumption data from operating areas in vastly different geographic regions and climates, with different customer bases, to make better forecasts for areas that have insufficient historical data. Our surrogate data transformation algorithm results in higher forecast accuracy, thereby reducing the risk to energy utilities.**

***Index Terms*—surrogate data, analogous time series, data transformation, design day forecasts**

I. Introduction

Forecasting natural gas demand on days with unusual weather patterns is inherently risky [1], in the sense of having higher than average uncertainty. On the historically coldest days, natural gas demand often reaches its peak. However, it is difficult to forecast whether this peak will occur on the coldest day, the day with the largest temperature drop, or perhaps a day after a very cold day [2]. There is great risk in forecasting these events, as energy utilities strive to avoid being unable to provide sufficient gas for all of their customers or purchasing too much gas.

Days with unusual weather patterns are by nature infrequent. An event with an expected frequency of once in 30 years may only show up once in a 30-year data set. If those 30 years of data are all from years that had mild winters, there may be no such event in the historical data. This makes the task of developing good models very difficult, as mathematical models require data. To combat this problem, available data from other service territories may be transformed as surrogate data to improve a model for a territory where there is insufficient historical data. Transformed surrogate data also benefits areas that have many years of available data, but when most of this data has normal or mild weather conditions.

Marquette University GasDay is a research laboratory that works with 30 energy utilities and forecasts about 18% of the natural gas consumed for residential, commercial, and industrial purposes in the United States. This makes GasDay uniquely qualified to work on this problem. GasDay maintains daily and hourly data for these utilities and for more than 1,000 weather stations. This data can be used as surrogate data to supplement available data for a target area. Our work shows that the use of surrogate data mitigates the risks of natural gas forecasts on days with unusual weather patterns by reducing forecast error.

Prepared for the 35^{th} International Symposium on Forecasting
Riverside, CA, 2015
*Corresponding co-author: Paul Kaefer – (414) 288-4418 – paul.kaefer@marquette.edu

II. Background

The use of analogous time series have been shown to improve forecasts. Duncan, Gorr, and Szczypula [3] pool data from analogous time series and find that it helps with volatile time series. Their methods work across data sets that are analogous despite having unrelated origins [4]. In our work, natural gas demand and the weather variables that impact it [5] are both highly volatile. Natural gas demand in service territories from different climates seems to be unrelated, yet show similar patterns due to weather-responsive customer behavior. Thus, we expect to find that using analogous data improves forecasting models.

Dixon, et al. [6] looks at the rare event of a *Darlingtonia californica* plant consuming an insect for food. Data collection is tedious and not very precise. They explore several methods of solving the problem of accurately determining the frequency of these rare events. They find that auxiliary data, which can come from pooling or aggregating data from different sources, improves precision when it can be obtained.

Thomas looks at using data from similar products when developing marketing models for a new product [7]. He finds that using parameters from existing products improves models for new products for which limited market data is yet available. Similar is the work of Lyness [8], who used data from nearby regions to create artificial data to supplement available weather data for areas that did not have the desired historical range. Likewise, our work aims to use available historical data from different service territories that show similar trends.

Another related technique is the bootstrap or jackknife [9]. These techniques involve repeated random sampling and could be used to supplement available data. Likewise, we use sampling, but rather than synthesizing random data, we have data from other sources (i.e., different service territories) that can be transformed and then sampled to supplement available historical data.

Brown [2] introduces the idea of using surrogate data in the natural gas demand forecasting domain. We describe the methods used at GasDay to improve forecasts for gas utilities around the United States.

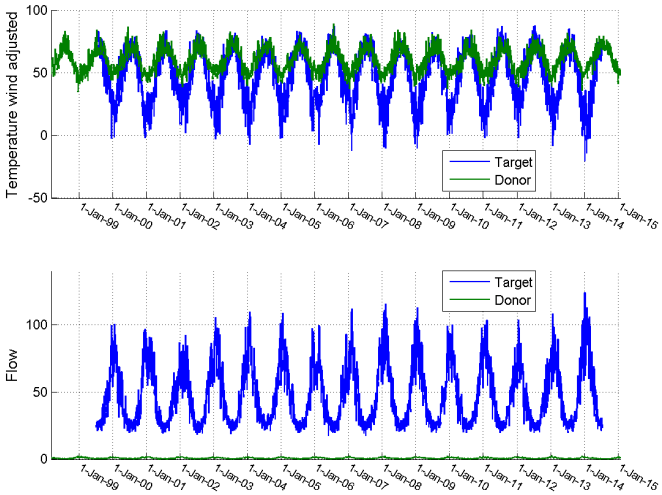

Figure 1: Time series of donor and target data sets.

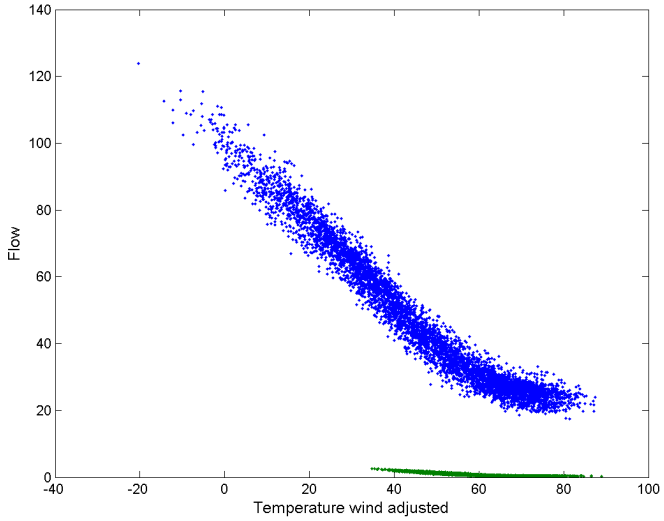

Figure 2: Scatter plot of donor and target data sets. Gas flow is plotted against wind-adjusted temperature.

III. Techniques

In this section, we discuss the procedure for transforming data from a donor service territory to match a target territory. Donor areas are selected using domain knowledge; the selection process is not the scope of this paper. As donor sets come from different climates and customer bases, they are not suited to work as donor data before transformation.

Figure 1 shows the time series for both wind-adjusted temperature and gas flow (scaled to protect proprietary data) for both the donor and target sets. (We adjust for wind as increased wind speeds causes heat loss in buildings.) While some similarities exist, such as higher gas flow when the weather is colder, using raw data from the donor area will not make a good model for the chosen target. Figure 2 shows a scatter plot for both raw data sets for gas flow plotted against wind-adjusted temperature. While both sets show the same “hockey stick" shape (due to the baseload and heatload characteristics mentioned in [5]), the donor has a much lower flow than the target area.

To transform this donor data to make good surrogate data, we first warp temperature values of the donor to match the target. Figure 3 illustrates this transformation. As the donor area typically has much warmer temperatures, the warping transforms the 5^{th}-percentile temperature of $43°\text{F}$ for the donor area to be $7°\text{F}$ to match that of the target area. Likewise, the optimal heating degree day (HDD) reference temperature [5] of about $60°\text{F}$ is slightly modified to be about $59°\text{F}$. While we use temperatures in Fahrenheit, the same methods could be applied to Celsius.

Once these values have been calculated, temperature values for all days for the donor set are scaled. The scatter plot with warped temperature is shown in Figure 4. The data still will not help our models; however, it matches temperature characteristics of our target. Next, we perform gross scaling of the flow.

Illustrated in Figure 5 are the scaling values of the natural gas flow from our selected donor area. A flow of about 0.44 units in the donor area is scaled to be about 27.1 units for the

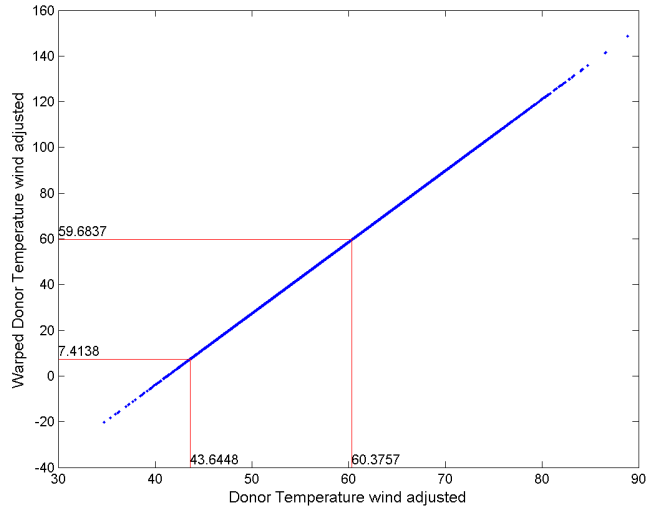

Figure 3: Temperature warping of the donor to match the target.

target area. Likewise, a flow of about 1.8 units will be scaled to match about 92 units for the target set. Figure 6 shows the scatter plot with this gross scaling applied. While the trend is clearly not the same with this scaling, it looks much closer to the target data than the raw donor data does.

Figure 7 shows both the temperature warping and the gross scaling of flow of the donor to match the target area. The transformed donor data (in green) is a much better match for the target area (in blue). It looks like it could be used as surrogate data.

IV. Results and Discussion

A. Data

We apply our surrogate data transformation techniques to about 160 operating areas from energy utilities around the

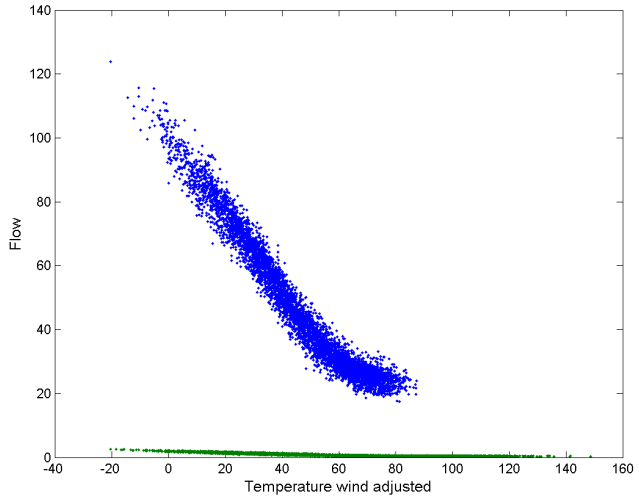

Figure 4: Donor area scatter plot with warped temperatures.

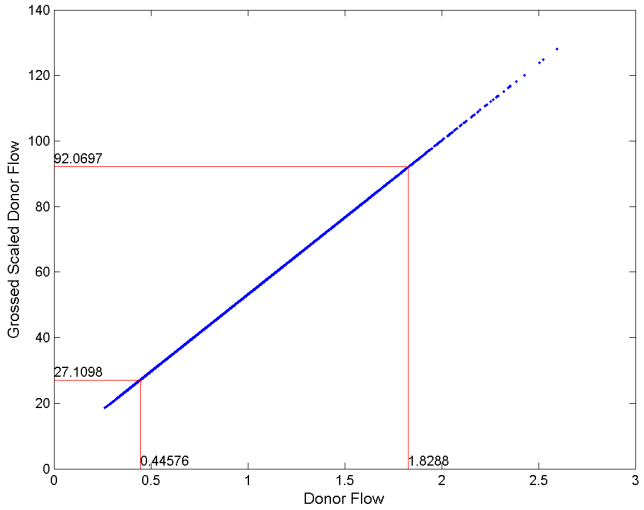

Figure 5: Gross scaling of the donor flow to match the target.

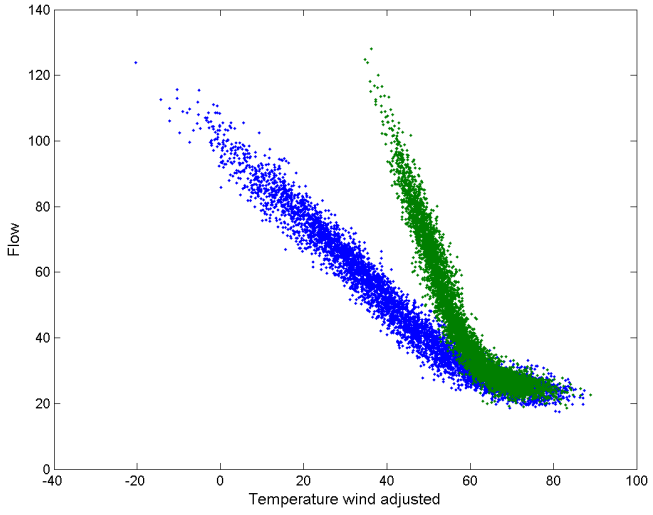

Figure 6: Scatter plot with gross scaling of flow applied to the donor set.

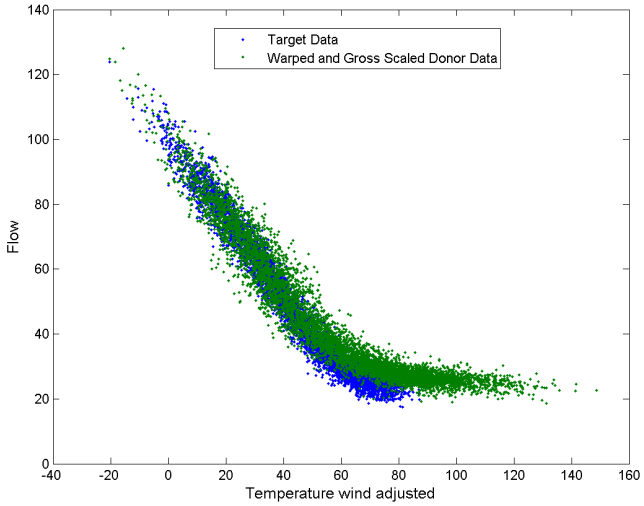

Figure 7: Scatter plot with temperature warping and gross scaling of flow applied to the donor set.

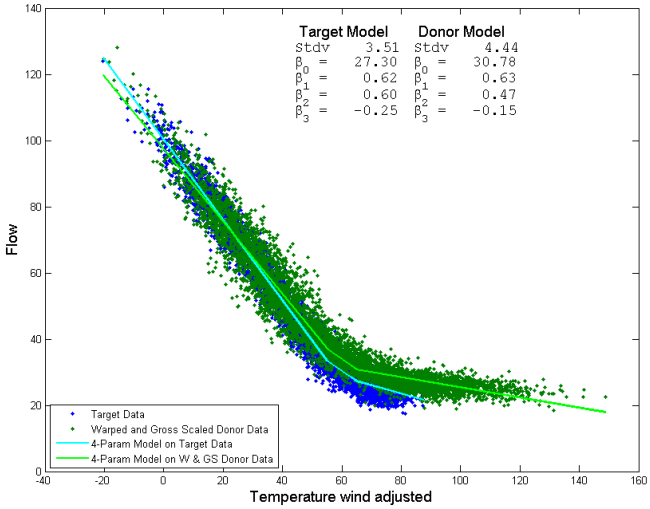

Figure 8: Evaluating transformed data using a four parameter model.

United States. Using domain knowledge, a subset of these potential donors is used as surrogate data for a given target.

B. Results

Figure 8 fits a four-parameter model to both the target data and one donor data set. The coefficients on the model for the transformed donor data are very close to those of the model fit to the target data. We have matched temperature and flow percentiles. The new data looks like it originated in the target data set, and can thus be used as surrogate data.

To evaluate our methods, we consider an event from the Southwestern United States in 2011 that is shown in Figure 9. The event is a “hook" event, as defined by [2], and is characterized by extreme cold and peak gas flow. The observed actual data is shown in teal, with circles marking individual days. Our forecasting model underforecasts (red trace). When

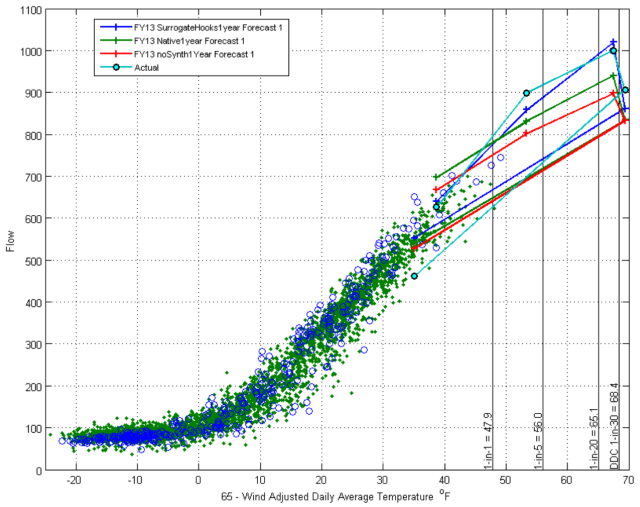

Figure 9: Using surrogate data to improve forecasts of an extreme event.

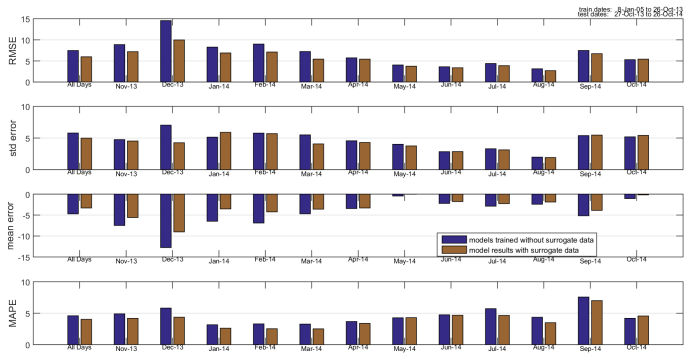

Figure 10: Forecasting error for all days and by month for models trained with and without surrogate data.

using synthetic data, the forecast is much better (green trace). When applying the methods described in this paper, we are able to produce forecasts represented by the blue trace. These forecasts are very close to the actual, illustrating how much surrogate data helps reduce risk on high flow days.

Our surrogate data methods do not just improve forecasts for individual multi-day cold events. Figure 10 shows error for all days and by month for an operating area in the Midwest. Models were trained without surrogate data (blue bars) and with our transformed surrogate data (brown bars). Across all days, and for almost every month, RMSE is smaller for models trained using our surrogate data. These results are consistent across operating areas in different parts of the country.

V. Conclusion

We find that our algorithm of transforming surrogate data improves forecasting models for days with unusual weather. By scaling temperature and flow values and fine tuning using a domain-specific model, we are able to improve accuracy through the use of surrogate data.

Future work includes developing a better metric for measuring error. As there is more risk in forecasts for unusual days, they should be weighted more heavily. MAPE and RMSE are commonly used in this field, but do not incorporate the increased risk on days that are more difficult to forecast.

As many variables impact natural gas demand for residential, commercial, and industrial purposes, there are other ways in which we may transform data to build surrogate data sets. We transform temperature and wind variables, for example, but by transforming additional weather inputs [10], we may be able to develop better surrogate data.

Finally, there are opportunities for improving how we select which of the available data sets should be used as surrogate sets. Experts can use their domain knowledge to select areas that will work well when their data has been transformed to match a target. This could also be done programmatically by looking at data features that may suggest they will make good surrogates.

References

[1] T. Quinn, "Risks and mitigations for near design day forecasting in the absence of recent historical events," in *Southern Gas Association Conference: Gas Forecasters Forum*, September 18 2012.

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

[3] G. T. Duncan, W. L. Gorr, and J. Szczypula, *Forecasting Analogous Time Series*, ser. Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, 2001, pp. 195–213.

[4] ——, "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

[5] 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. 7, pp. 807–827, Jan. 2009.

[6] 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

[7] 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, 3 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0737678285900153

[8] F. K. Lyness, "Consistent forecasting of severe winter gas demand," *The Journal of the Operational Research Society*, vol. 32, no. 5, pp. 347–459, 1981.

[9] 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

[10] 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