Detrending Daily Natural Gas Consumption Series to Improve Short-Term Forecasts

Ronald H. Brown[1], Steven R. Vitullo[1], George F. Corliss[1], Monica Adya[2], Paul E. Kaefer[1], *Student Member*, *IEEE*,
Richard J. Povinelli[1], *Senior Member*, *IEEE*

[1]Department of Electrical and Computer Engineering, College of Engineering, Marquette University, Milwaukee, Wisconsin USA
[2]Department of Management, College of Business, Marquette University, Milwaukee, Wisconsin USA

Abstract—**This paper presents a novel detrending algorithm that allows long-term natural gas demand signals to be used effectively to generate high quality short-term natural gas demand forecasting models. Short data sets in natural gas forecasting inadequately represent the range of consumption patterns necessary for accurate short-term forecasting. In contrast, longer data sets present a wide range of customer characteristics, but their long-term historical trends must be adjusted to resemble recent data before models can be developed. Our approach detrends historical natural gas data using domain knowledge. Forecasting models trained on data detrended using our algorithm are more accurate than models trained using non-detrended data or data detrended by benchmark methods. Forecasting accuracy improves using detrended longer-term signals, while forecast accuracy decreases using non-detrended long-term signals.**

Index Terms – **Energy forecasting, Detrending, Natural gas industry, Industrial applications, Multiple linear regression, Short-term demand forecasting**

I. INTRODUCTION

This paper presents a novel detrending algorithm for natural gas demand signals. We show that detrended long-term natural demand signals yield better short-term forecasting models than either models trained on non-detrended long term natural gas demand signals or models trained on benchmark detrended signals.

Natural gas use is increasing rapidly in the field of electric power generation, space heating, and transportation because of relatively low cost and environmental benefits from its low emissions. According to the American Gas Association [1] and the U.S. Energy Information Administration [2], about 58 million American homes use natural gas for its comfort, cleanliness, and reliability. To furnish uninterrupted services, Local Distribution Companies (LDCs) that deliver natural gas to end consumers must provide demand forecasts to pipeline companies so they, in turn, can have adequate gas available to meet LDC needs [3], [4]. Cases of under- and over-forecasting can have significant economic implications for LDCs and for their customers through increased billings [3], [5].

To run operations economically and safely, LDCs need to forecast with high accuracy customer consumption for the next day and several days beyond. Often, long-term daily historical data are preferred for building daily gas consumption forecasting models, as these longer data sets capture a greater range of consumption patterns. For instance, natural gas consumption is expected to be high on cold days and to deviate from normal patterns on unusual days such as holidays and days on which cold and warm fronts pass. When training gas forecasting models, a rich set of unusual days in the training set is crucial for calibrating the forecasting model to such days. However, such long-run data also run the risk of misrepresenting consumption patterns of the current customer base as customer count, customer equipment, consumption behaviors, efficiency improvements, economic impacts, price, and other factors may vary with time. Paradoxically, building forecasting models using long historical data often leads to biased forecasts.

Two methods to account for the non-stationarity of the historical data when building a forecast model are 1) include variables in the model that represent the trends (detrend the model) and 2) detrend the historical data. In this paper, we show short-term natural gas demand forecasting models built on detrended historical data using the our algorithm are superior to model detrending methods.

Data detrending techniques adjust a historical series so that it can approximate a stationary customer base whose consumption patterns reflect current behavior, but whose responses to unusual weather and holiday patterns reflect several years of history. Numerous studies have demonstrated that such detrending techniques improve forecast accuracy [4], [6]-[8]. To illustrate, suppose a forecasting model is developed using daily natural gas consumption data from the most recent five years for a gas territory with substantial, approximately linear growth. If all days in this model's training period are equally weighted, the resulting model best predicts the load for the average customer base in the training period. Residual errors of the model are smallest for the middle year, positive (forecasts greater than actual consumption) over the first two years of the training period, and negative (forecasts lower than

actual consumption) over the last two years. Contrary to the goal of building a model to predict consumption for the forthcoming heating season, such a model best fits data for the heating season three years prior. By detrending consumption data, the resulting model forecasts the forthcoming heating season with reduced error as the residual for the first, middle, and last years will be closer to uniform.

Standard detrending models, however, work poorly in the natural gas domain, as these models overlook critical domain knowledge that captures nonlinear dynamics of gas consumption. To address this limitation, this study describes a five-parameter detrending algorithm based on natural gas domain knowledge. For benchmark comparison, we also use two well-accepted detrending techniques derived from best practices. Gas consumption series are detrended using these three techniques. These detrended training series then are used to forecast a one year test series using multiple linear regression. The forecast accuracy of these detrended series is compared with those for non-detrended series. After an overview of the natural gas forecasting domain, the paper proceeds to describe the five-parameter detrending algorithm and its validation.

II. BACKGROUND AND MOTIVATION

Natural gas consumption is influenced by factors such as temperature, wind, day of the week, and holidays [4], [9]-[13]. Temperature is the most important variable, since natural gas primarily is used for space heating by residential and commercial consumers [9], [12]. Gas consumption, often measured using decatherms (Dth), increases with declining average temperature. However, as temperature increases beyond about 65°F, consumption levels off near some constant value called base load. Base load captures consumption driven by non-heating needs. Figure 1 juxtaposes typical daily average temperature and the daily gas consumption of a region for a U.S. utility.
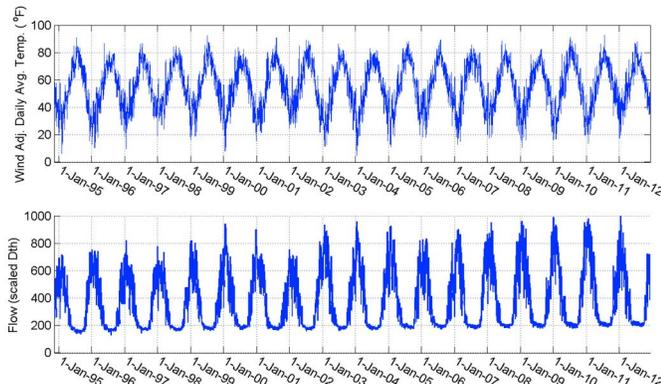

Figure 1. Wind adjusted daily average temperature and daily gas load.

This nonlinear relationship between consumption and temperature has been useful in identifying the most important variable in gas consumption forecasting, the heating degree day (HDD) [4], [9], [15]-[17].

$$HDD_k = \max(T_{ref} - T_k, 0) \,, \qquad (1)$$

where T_k is the average temperature for the kth day, and T_{ref} is the reference temperature, historically set to 65°F. HDD is a simple metric for quantifying the amount of heating needed for a particular customer during a certain period, in this study, for one day. We also adjust temperatures to account for wind effects. Wind creates the "wind chill effect" and also causes buildings to lose heat faster [4].

A detrending adjustment is a widely used technique in time series analysis of non-stationary data sets [18], [19]. In their respective domains, these studies demonstrated that the problem of short series can be alleviated by detrending long historical data sets, driving gains in forecast accuracy. These long series benefit models by deriving trends in customers' characteristics, yielding unbiased forecasts.

A range of time series detrending methods has appeared in the forecasting literature. The simplest approach is to use a linear detrending regression factor as done by Gujarati [20], Nelson and Kang [21], South [22], Kaun [23], and Raffalovich [24]. Despite its wide use, there is consensus that it is not a particularly effective method for detrending, as it fails to capture cross-term variables and trend [21], [25]. Others such as Harvey and Jaeger [19], Haida and Muto [6], and Gaunholt [18] have used filters or discrete Fourier transforms to find a smooth trend through time series in economics, road profile measurements, and electric power forecasting, respectively.

Unfortunately, in natural gas consumption, the trend over time is not smooth. Rather, gas consumption is shaped by both base load and heat load, which vary with factors we mentioned earlier. Considering this, two simple detrending methods are primarily used in natural gas forecasting. These two models serve as effective benchmarks for our proposed model.

Let $l(.)$ represent a typical linear regression model for forecasting daily natural gas demand [4] and be Benchmark Model 1. Let Benchmark Model 2 be Benchmark Model 1 with an additional linear trend term.

$$\hat{S}_k = l(.) + \beta_1 k \,, \qquad (2)$$

where $\hat{S}_k$ is the consumption for the kth day, $l(.)$ is Benchmark Model 1, and $\beta_1 k$ is the trend on day k.

Domain knowledge informs us that the load can be broken into base load and heat load components, and these component loads can change independently with time. In Benchmark Model 3, we introduce an estimate for consumption,

$$\hat{S}_k = l(.) + \beta_1 k + \beta_2 k \cdot HDD_k \,, \qquad (3)$$

where HDD_k is the heating degree day on the kth day. The advantage of this technique is that it captures both trend in baseload with $\beta_1 k$ and trend in heat load with $\beta_2 k \cdot HDD_k$.

III. DETRENDING ALGORITHM FOR NATURAL GAS DATA

Benchmark Models 2 and 3 have terms to account partially for the trends in the historical data. In this section, we present an approach to detrend the historical data before estimating the parameters of the forecasting model $l(.)$. The problem of a non-stationary customer base can be overcome partially by detrending older historical data. A simple way to adjust historical data is to transform its characteristics to match the most

recent heating season by calculating a linear regression model on each heating season and considering its coefficients as time varying. Consumption from heating seasons prior to the most recent one can be adjusted by adding a base load factor to each day to equate the base load to the most recent season and additional consumption proportional to the daily HDD to adjust the use per HDD factor to be the same as the use per HDD for the most recent season. For example, consider the two-parameter model $\hat{S}_k = \beta_0 k + \beta_1 HDD_k$. Using 2009–2010 heating season data (July 2009 through June 2010), we build a two-parameter model

$$\hat{S}_k = \beta_0^{2010} + \beta_1^{2010} HDD_k. \tag{4}$$

The two-parameter model for 2008–2009 data is

$$\hat{S}_k = \beta_0^{2009} + \beta_1^{2009} HDD_k. \tag{5}$$

The difference between β_0^{2009} and β_0^{2010} is the trend in the base load between the two heating seasons. The difference between β_1^{2009} and β_1^{2010} is the trend in the consumption per HDD. If conditions from the 2008–2009 heating season had occurred on the 2009–2010 customer base, we would expect the base load to change by $\beta_0^{2010} - \beta_0^{2009}$, i.e., the change in the base load between the heating seasons. The expected heat load is the change in consumption per heating degree day times the heating degree day, which is $\left(\beta_1^{2010} - \beta_1^{2009}\right) HDD_k$.

All the data in the 2008–2009 heating season is detrended by adding this adjustment to the daily consumption. For each day k in the 2008–2009 heating season, we replace the original consumption $s_k^{orig\,2009}$ with

$$s_k^{detrended\,2009} = s_k^{orig\,2009} + \left(\beta_0^{2010} - \beta_0^{2009}\right) \\ + \left(\beta_1^{2010} - \beta_1^{2009}\right) HDD_k. \tag{6}$$

If the two-parameter linear regression model is fit to the detrended 2008–2009 consumption data, its base load and heat load coefficients are the same as the base load and heat load coefficients of the 2009–2010 model.

This process repeats for historical heating season data and learns another two-parameter model.

$$\hat{S}_k = \beta_0^{2008} + \beta_1^{2008} HDD_k. \tag{7}$$

Using these model parameters, we similarly detrended this year of data,

$$s_k^{detrended\,2008} = s_k^{orig\,2008} + \left(\beta_0^{2010} - \beta_0^{2008}\right) \\ + \left(\beta_1^{2010} - \beta_1^{2008}\right) HDD_k. \tag{8}$$

The detrended data has the same base load and heat load characteristics as the 2009–2010 data. Historical load is adjusted using the two-parameter approach described above. The time series for original consumption, detrended consumption, and growth in consumption (the amount of adjustment that is added to the original consumption to get the detrended consumption) are shown in Figure 2, where we observe a gradual increase in consumption during the training period. The consumption versus temperature scatter plots of the

original data (before detrending) and adjusted data (after detrending) are proximally shown in Figure 3. The scatter plot of adjusted data is much tighter. For example, the consumption at 40 HDD in the original plot is between 650 and 950 Dth. After detrending, the consumption at 40 HDD is between 800 and 950 Dth.
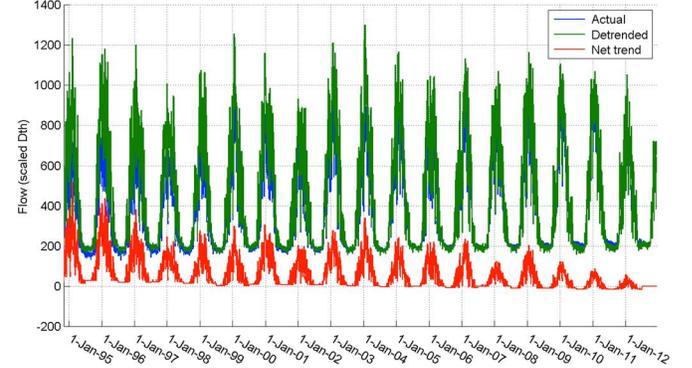

Figure 2. Actual consumption, detrended consumption, and net trend in load.

The two-parameter detrending algorithm introduced above produced a more tightly clustered series than the original series. However, it introduced opportunities for further improvement. First, the two-parameter model has only two degrees of freedom. Including more domain-relevant factors certainly leads to a better model [4]. Second, adjusting the trend for each heating season yields level discontinuities. Building the detrending models with higher temporal resolution smooths out the discontinuities. These two improvements lead to the five-parameter algorithm described next.
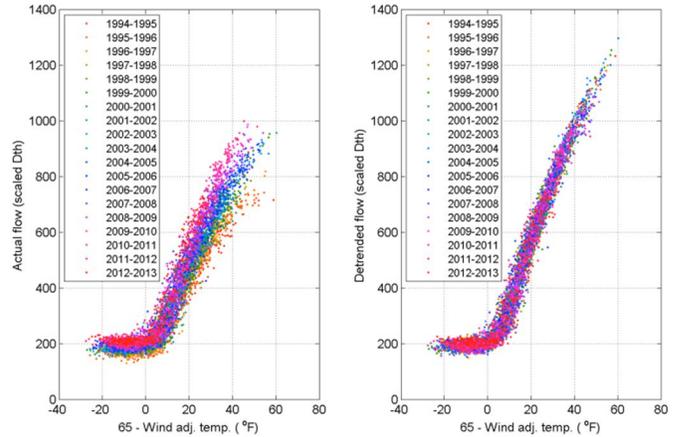

Figure 3. Scatter plot of actual load (left) and detrended load (right)

The HDD reference temperature that best captures customer behavior varies with time. Using a single fixed reference temperature in the model is insufficient, as it does not capture the optimal reference temperature for the utility service region. One way to adjust this change is to give the model an extra degree of freedom by including a second HDD factor with a reference temperature of 55°F for an improved fit when the optimal HDD reference temperature is between 65°F and 55°F.

Domain expertise and thermodynamics tell us that heat loss in homes and buildings is a dynamic process [4], [7]. As such, the heat load may vary, even at the same outside temperature. Consider heating a typical house. Suppose two days have the same temperature, say 30°F. If the first 30°F day follows a day that is much colder, and the second 30°F follows a day that is much warmer, the first 30°F day usually requires more gas than the second. To account for this effect, we introduce a new term for the change of HDD from the previous day,

$$\Delta HDD_k = HDD_k - HDD_{k-1}, \qquad (9)$$

where ΔHDD_k is the change in heating degree day from day $k-1$ to day k.

Additionally, to account for variation in consumption of natural gas for temperatures above 65°F, a fifth factor, a cooling degree day

$$CDD_k^{65} = \max\left(T_k - 65, 0\right) \qquad (10)$$

is added [4]. These additional variables, well established and used in the energy forecasting domain [4], [11], [16], were added to the two-parameter model to produce the five-parameter model.

$$\begin{aligned} \hat{S}_k = \beta_0 + \beta_1 HDD_k^{65} + \beta_2 HDD_k^{55} + \beta_3 \Delta HDD_k \\ + \beta_4 CDD_k^{65} \end{aligned}, \qquad (11)$$

where HDD_k^{65} is the heating degree day using a reference of 65°F for day k, HDD_k^{55} is heating degree day with a reference of 55°F for day k.

The method described in (4) – (8) is built with one trend model per year, yielding discontinuities in trends between years. A model with higher temporal resolution eliminates this effect. To achieve this, instead of building one model per year, we build models on a sliding one year window of data. We slide the one-year window of data through all the historical data with one month increments to get time-varying detrending model coefficients. For instance, if the most recent model was built on data from July 2009 through June 2010, the second model is built on data from June 2009 through May 2010; the third from May 2009 through April 2010, etc. These coefficients are then smoothed to get continuous-valued time-varying detrending model coefficients. Then, these curves are used to adjust the historical data.

IV. BENCHMARK TESTING OF THE DETRENDING MODELS

The validity and effectiveness of the five-parameter data detrending method presented here can best be tested by building models and evaluating forecasts made with the models.

The same natural gas consumption series from a U.S.-based LDC are forecast using the four multiple linear regression models discussed previously. The original series is used to parameterize the coefficients of Benchmark Model 1 (typical model without trend factors), Benchmark Model 2 (typical model with linear trend), Benchmark Model 3 (typical model with linear trend and linear trend crossed with HDD).

The fourth model is the same typical model without trend factors as Benchmark Model 1 but parameterized on the five-parameter model detrended data (it is not the five-parameter model).

The original series contains 15 years of gas consumption data. The last year of this series is held back for ex-ante testing to evaluate accuracy, while the forecasting models are built using the most recent three, five, seven, and all 14 years of the remaining data as a training set. The magnitude of gas load data has been scaled in this paper to protect the proprietary LDC data. To assess forecast accuracy, Root Mean Square Error (RMSE) and Weighted Mean Average Percentage Error (WMAPE) are used. RMSE penalizes larger errors much more than smaller ones, and WMAPE yields percentage error but weights the error relative to the amount of consumption. Hence, the highest weighting was given to winter days, followed by spring and fall, with summer being weighted the least. These metrics reflect the penalties that the LDC's customers pay for inaccurate forecasts better than the commonly used Mean Absolute Percentage Error (MAPE) or Mean Absolute Error (MAE).
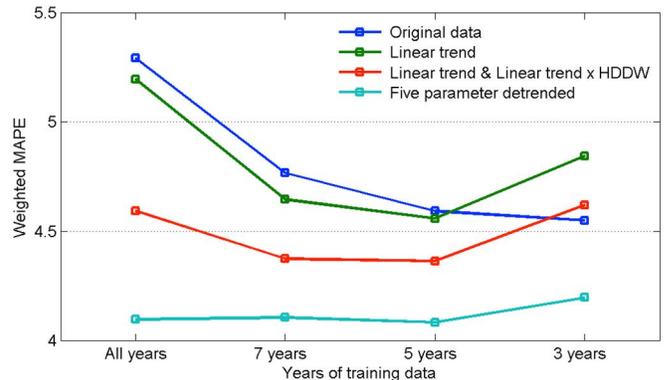

Figure 4. Test set WMAPE for models by number of years of training data

Figure 4 shows the WMAPE for the test sets. Figure 4 shows that, on forecasting models without historical trend adjustment to the training data, forecast error increases as more years of training data are used for model development. Benchmark Models 2 and 3 performed better than the typical model, but also suffer when the training data gets long as the non-stationarity is not modeled well. The forecast error of the typical model built on the detrended historical data decreases with more training data.

When comparing the model built on domain-based five-parameter detrended data with the traditional benchmark detrending models, we see that nonlinear features are important in modeling the non-stationarity of the data. These nonlinear features capture the natural gas consumption trends more accurately than the traditional benchmark models. Even though the traditional model can take linearly increasing load into account, it does not capture seasonal variance. Further, Benchmark Model 3 with a product of linear trend and HDD term shows large improvements over the simple trend model (Benchmark Model 2).

V. Conclusions

In this study, we have shown that better short-term natural gas demand forecast models can be built using long detrended historical data sets than models with detrending factors built on the original data. We have presented the detrending algorithm.

A short-term natural gas forecasting model without detrending was compared to three detrending techniques. The study provides evidence that domains constrained by non-stationarity can leverage larger training sets by detrending older data. Specifically, it proposes that certain domains, such as natural gas consumption, that pose a challenge for traditional detrending approaches may benefit from examining alternative detrending techniques that build upon parameters unique to their domain. The methods described in this paper for detrending and forecasting are currently implemented at 30 LDCs in the U.S. and have been improving LDC's forecasts for several years.

The idea of detrending time series discussed in this paper can be applied to a wide range of series in different domains, such as engineering, business, and economic studies, to generate improved forecasts. For instance, to our knowledge these techniques have not been applied to electric load demand forecasting, but are easily transferrable to this domain. For instance, the electric power forecasting domain could benefit from a similar algorithm by using Cooling Degree Days (CDD) instead of HDD to model summer air conditioning load.

Furthermore, interesting variations might emerge from applying these approaches to other national data. Natural gas consumption patterns in Europe and Asia, for example, are distinctive from those in the U.S. because of differences in environmental factors, technologies, and consumption behaveiors, among others. Calibrating the detrending models to these diverse settings could enhance the opportunities for domain-based application of detrending variables. Additional research opportunities may emerge from applications to other renewable energy domains such as nuclear power, and economic and financial phenomena. For instance, the test set in this study ranged from the 2009 to 2010 heating season. Natural gas consumption during this heating season was affected by changes in economic conditions strong enough to alter consumption characteristics. Incorporating economic variables in these models could yield further opportunities to refine the proposed models, and work is in progress to do so. Finally, researchers also may find beneficial extensions of this study through investigation of additional or alternative variables and functional forms of the model for detrending LDC consumption data.

References

[1] (2010, May 1). American Gas Association (AGA), [Online]. Available: http://www.aga.org.

[2] (2010, May 1). Energy Information Administration (EIA), [Online]. Available: http://www.eia.gov.

[3] T. W. Knowles and J. P. Wirick, Jr. "The Peoples Gas Light and Coke Company Plans Gas Supply," Interfaces, vol. 28, pp. 1–12, 1998.

[4] 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, pp. 807–827, 2009.

[5] W. Avery, G. G. Brown, J. A. Rosenfranz, and R. K. Wood, "Optimization of Purchase, Storage and Transmission Contracts for Natural Gas Utilities," Operations Research, vol. 40, pp. 446–462, 1992.

[6] T. Haida and S. Muto, "Regression Based Peak Load Forecasting Using a Transformation Technique," IEEE Transactions on Power System, vol. 9, pp. 1788–1794, 1994.

[7] A. Taware, "Forecasting and Identification Methods Applied To Gas Load Estimation Problems," M. S. Thesis, Marquette University, 1998.

[8] D. Hughes and R. H. Brown, "Analysis of Customer Base Growth over Time," presented at Gas Forecasters Forum, Greensboro, North Carolina, USA, 2002.

[9] H. Aras and N. Aras, "Forecasting Residential Natural Gas Demand," Energy Sources, vol. 26, pp. 463–472, 2004.

[10] R. H. Brown, P. Kharouf, I. Matin, 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, 1996.

[11] R. H. Brown, "OK, who is playing with the thermostat?" presented at Gas Forecasters Forum, Albuquerque, New Mexico, USA, 2003.

[12] F. B. Gorucuand and F. Gumrah, "Evaluation and Forecasting of Gas Consumption by Statistical Analysis," Energy Sources, vol. 26, pp. 267–276, 2004.

[13] H. L. Lim and R. H. Brown, "Hourly Gas Load Forecasting Model Input Factor Identification Using a Genetic Algorithm," in Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems, 2001, vol. 2, pp. 670-673.

[14] J. G. Asbury, C. Maslowski, and R. O. Mueller, "Solar Availability for Winter Space Heating: An Analysis of SOLMET Data, 1953 to 1975," Science, vol. 206, pp. 679–681, 1979.

[15] D. G. Baker, "Effect of Observation Time on Mean Temperature Estimation," Journal of Applied Meteorology, vol. 14, pp. 471–476, 1975.

[16] H. G. Berrisford, "Relation between Gas Demand and Temperature: A Study in Statistical Demand," Operations Research Quarterly, vol. 16, pp. 229–246, 1965.

[17] P. M. Dare, "A Study of the Severity of the Midwestern Winters of 1977 and 1978 Using Heating Degree Days Determined from Both Measured and Wind Chill Temperatures," Bulletin of the American Meteorological Society, vol. 62, 974–982, 1981.

[18] H. Gaunholt, "A Subharmonic Detrending or Data-smoothing Approach for Longitudinal Road Profile Measurements," in Proceedings of the IEEE International Symposium on Circuits and Systems, 2001, vol. 2, pp. 413-416.

[19] A. C. Harvey and A. Jaeger, "Detrending, Stylized Facts and the Business Cycle," Journal of Applied Econometrics, vol. 8, pp. 231–247, 1993.

[20] D. N. Gujarati, Basic Econometrics, Boston, Massachusetts, USA: McGraw Hill, 2003.

[21] C. R. Nelson and H. Kang, "Pitfalls in the Use of Time Series as an Explanatory Variable in Regression," Journal of Business and Economic Statistics, vol. 2, pp. 73–82, 1984.

[22] S. J. South, "Economic Conditions and the Divorce Rate: A Time-series Analysis of the Postwar United States," Journal of Marriage and the Family, vol. 47, pp. 31–41, 1985.

[23] D. E. Kaun, "War and Wall Street: The Impact of Military Conflict on Investor Attitudes," Cambridge Journal of Macroeconomics, vol. 10, pp. 492–519, 1990.

[24] L. E. Raffalovich, "Detrending Time Series: A Cautionary Note," Sociological Methods Research, vol. 22, pp. 492-519, 1994.

[25] R. McCleary, and R. A. Hay, Jr. Applied Times Series Analysis for the Social Sciences. Beverly Hills, California, USA: Sage, 1980.