Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 695084, 9 pages
http://dx.doi.org/10.1155/2014/695084

Research Article

US Natural Gas Market Classification Using Pooled Regression

Vyacheslav V. Kalashnikov,[1,2,3] Gerardo A. Pérez-Valdés,[4] Timothy I. Matis,[5] and Nataliya I. Kalashnykova[3,6]

[1] *Departamento de Ingeniería Industrial y de Sistemas, Tecnológico de Monterrey, Avenida Eugenio Garza Sada 2501 Sur, 64849 Monterrey, NL, Mexico*
[2] *Department of Experimental Economics, Central Economics and Mathematics Institute (CEMI), The Russian Academy of Sciences (RAS), Nakhimovsky Pr. 17, Moscow 117418, Russia*
[3] *Department of Electronics and Computing, Sumy State University, Rimsky-Korsakov Street 2, Sumy 40007, Ukraine*
[4] *Institutt for Industriell Økonomi og Teknologiledelse, Norges Teknisk Naturvitenskapelige Universitet, Alfred Getz Vie 3, 7491 Trondheim, Norway*
[5] *Department of Industrial Engineering, Texas Tech University (TTU), P.O. Box 43061, Lubbock, TX 79409, USA*
[6] *Departamento de Ciencias Físico-Matemáticas (FICA), Universidad Autónoma de Nuevo León (UANL), Avenida Universidad S/N, 66450 San Nicolás de los Garza, NL, Mexico*

Correspondence should be addressed to Vyacheslav V. Kalashnikov; kalash@itesm.mx

Received 25 December 2013; Accepted 18 February 2014; Published 23 March 2014

Academic Editor: Sergii V. Kavun

Natural gas marketing has considerably evolved since the early 1990s, when a set of liberalizing rules were passed in both the United States and the European Union that eliminated state-driven regulations in favor of open energy markets. These new rules changed many things in the business of energetics, and therefore new research opportunities arose. Econometric studies about natural gas emerged as an important area of study since natural gas may now be sold and traded in a number of stock markets, each one responding to potentially different behavioral drives. In this work, we present a method to differentiate sets of time series based on a regression model relating price, consumption, supply, and other factors. Our objective is to develop a method to classify different areas, regions, or states into groups or classes that share similar regression parameters. Once obtained, these groups may be used to make assumptions about corresponding natural gas prices in further studies.

1. Introduction

In the early 1990s, several regulations were passed in the US and the European Union [1–3] changing the way natural gas was marketed and traded. Particularly, this liberalization [4] effectively ended a period in which natural gas was a state-driven industry. The liberalization has also created the emergent natural gas markets, as well as a strong demand for models to better tackle the new problems and profit from this new setting [5, 6].

Owing not only to this liberalization, but also to the new local conditions that are more open to competition, new small players entered the natural gas industry, especially at the local scale. Indeed, the US has over 80 interstate, long-distance pipelines [7], serving different regions with various climatic, demographic, economic, and political circumstances. Natural gas usage in Alabama, for example, intuitively is not the same as in Oregon; thus the market dynamics of the fuel are also different, and this, we presume, should be reflected in some way in the econometric data of the states.

Not only macroeconomic trends, however, are affected by this setting. When doing cross-regions studies of various aspects of the supply chain, such as the forecasting of demand [8, 9], the balancing of the pipelines after imbalances have been created by the natural gas shippers [10–12], or the dynamics of interstate-intrastate systems [13], one has to take into account the existence of different markets. The existence of a common relationship between price and consumption of natural gas across several zones allows for strong claims

of uniformity, which are useful when, for example, we are building scenarios for a stochastic problem. Indeed, if we manage to group the regions in clusters with similar price and consumption functions, we can reduce the number of variables needed in a scenario tree formulation [6, 14].

As such, we specify a regression function that relates many of the most relevant econometric figures for each of the 48 contiguous states of the American Union, modeling price as a function of explicative variables such as natural gas consumption, supply, and storage levels, as well as population (number of costumers), oil prices, temperatures, and production. The regression coefficients are then used to divide the set of states into several subsets or groups, obtaining a partition in which all the states in a group share the same regression parameters, and thus can be classified as an (implicit) market. The partition is made considering both statistical and nonstatistical characteristics of the obtained regression coefficients. The resulting partitions are next compared with others in their similitude and statistical significance, which would validate the goodness of the combination of the dendrogram and GRASP grouping methods.

This paper is organized as follows. The motivation and literature review on natural gas econometric regression is given in Section 2. Section 3 describes the way the regression function is derived, while Section 4 details the method for using the said function to perform the classification. Section 5 presents and discusses the results of the study, and conclusions are given in Section 6.

2. Natural Gas Price-Consumption Model

This work was motivated by our previous research in the natural gas supply chain, specifically developing an optimization model that addresses issues in interstate pipelines. The data used in this model, however, came from different regions, and therefore the time series involved did not necessarily behave in the same way.

As an example, suppose we are trying to model a certain problem that involves forecasting the residential consumption and price of natural gas in the states of Washington and Oregon, that is, four time series. If the robustness of the model is also a concern, then we should additionally consider different forecasting scenarios. Even with only two possible forecasting scenarios for each series (high/low consumption or prices) this translates into 2^4 possible behaviors of the econometric parameters. If consumption is expressed as a function of price, however, then the scenario tree has only 2^2 branches. Furthermore, if the regression function for both states is the same, then the number of scenarios can be reduced to just two. As the number of states being modeled scales up, that is, there are more than two parameters of interest, common assumptions like those mentioned above help reduce greatly the amount of scenarios in a stochastic model, optimization, or otherwise.

As we studied particular sets of data, it was noted that historical data of consumption and price showed conspicuous properties that could be used for the sake of our aims. Even though these data collections were taken from different states, all pairs of time series showed elastic consumption/demand [15, 16]; exponentially growing price averages [15, 17]; and both series in every pair seemed to be highly correlated to each other.

Indeed, the possibility of characterizing one set of series as a (regression) function of the other was interesting, as it would reduce the amount of data we needed to consider when modeling optimization problems. It is, of course, a common practice in economic and managerial sciences to do that since, for example, demand data is simpler to work with than price data [18]. The latter is mainly because the demand is usually easier to predict, and its behavior is less chaotic than that of prices. Such historical relationships between price and consumption is a common topic of study in time series economic analysis [19], which is mostly performed with the inclusion of other explicative variables, such as the price of substitutes (electricity, coal) and weather conditions.

This is the case of several models where the calculation of elasticities is the primary goal of the study [20]. Log-linear models [21–24] are generally favored because of the ease they provide when computing elasticity figures. However, linear models also have applications in the natural gas industry, like the Short-Term Integrated Forecasting system (STIFS) used by the United States Energy Information Agency in order to estimate natural gas demand as a function of several types of important variables related to the energy industry [25].

2.1. Former and Current Approaches. As explained in our previous work [26], a carefully designed regression function can help achieve such strong assumptions. Nevertheless, the study of such relationships and the possibility of forming state clusters based merely upon time series data analysis turned out to be interesting by itself, and we developed two different approaches to partition the collection of states. As we observed, neighboring states showed a large amount of diversity, yet different methods of grouping seemed to place certain states consistently together.

Two major areas of opportunity discovered were the design of the regression function, and the trade-off that each partition algorithm made use of.

Our previous paper [26] aimed at a very definite objective regarding the qualities of the regression model: it had to correlate consumption and price of residential natural gas series, using the former as the explicative variable because of the ease in its forecasting. The expression thus obtained served its purpose well, as demonstrated in its application to the optimization models in [27]; nevertheless, a more inclusive approach would involve series that comprise more information. Following the examples found in the literature and our own experience, we revealed that including more explicative series provided very good results in terms of regression fit. This has led to the model presented in the next section.

Coming back to the partitioning method, the two approaches presented before were as follows.

(i) The first one is the Dendrogram Grouping Method, which "cuts" a binary tree (whose nodes represent regression parameters) based on how close to each other the parameters are with respect to a given metric

function and a weight scheme for the entries. This method proved replicative and fast, but it does not provide statistical significance to the grouped states' parameters (i.e., one state might find that temperature is a significant regressor, whereas some other state in the same group may not).

(ii) Another one is a greedy heuristic that starts with a number of states called "group leaders," and iteratively selects for each remaining state the group that suits the state best, based on its regression coefficient R^2. Because of the large amount of regressions performed, this method was reported to be slower and subject to accidental fluctuations, but the final results always guaranteed that all states in one group shared the same significance in their parameters.

In the following sections, we explain how we have improved over our latest approach, adding explicative power and robustness to the partitioning method and, ultimately, creating a better technique to identify similar regions based on their econometric data.

3. Regression Analysis

3.1. Individual Multiple Linear Regression (IMLR). Let n be the total number of states, m the number of observations per time series (months, in this case), $I = \{1, 2, \ldots, n\}$ the set of the 48 contiguous states of the American Union, $t \in \mathbf{T} = \{1, 2, \ldots, m\}$ the (discrete) time parameter, $\{P'_{i,t}\}$ the differenced residential natural gas price in state $i \in I$ at time $t \in \mathbf{T}$, $\{T'_{i,t}\}$ the differenced temperature, in Kelvin, shifted so that the minimum figure is e, $\{O'_t\}$ the differenced average spot price of oil in the US at time $t \in \mathbf{T}$, $\{N'_{i,t}\}$ the differenced number of residential consumers of natural gas in state $i \in I$ at time $t \in \mathbf{T}$, and $\{C'_{i,t}\}$ the differenced consumption of natural gas in state i at time t.

Notice that all these series are *differenced*, or more precisely, lag-(-1) differenced from the original values. This is because the said original values all tested positive for unit roots in the advanced Dickey-Fuller test. In contrast to the original series, the differenced series prove to be stationary; hence we make use of the latter.

This is the linear model we devised to relate the above-described series:

$$\widehat{C'_{i,t}} = \alpha_{0,i} + \alpha_{1,i} P'_{i,t} + \alpha_{2,i} C'_{i,t-12} + \alpha_{3,i} T'_{i,t} + \alpha_{4,i} O'_t + \alpha_{5,i} N'_{i,t}; \quad t \in \mathbf{T}; \quad i \in I. \tag{1}$$

We choose a Robust Regression Analysis using Huber weights to fit the series over traditional least-squares method due to nonnormality of the residuals experienced with the latter. Furthermore, due to the steps described in the next sections, heteroskedasticity would likely appear in the residuals once the pooling regression is carried on.

While most of the series were reasonably fit by (1), a couple of them showed very erratic behavior in either their natural gas price or consumption series. This is expected insofar economic forecasting is commonly subject to the large instability at time t. As the driving force behind short-term fluctuations in natural gas pricing is consumer demand rather than production supply, price was shown to be a significant factor when describing market consumption.

The selection of the descriptive variables was made considering other consumption models in the literature, the available data, and the significance found in the preliminary regression analysis. In particular, electricity prices and the natural gas supply and production, as well as a time index, were tested but found not to be significant in most of the states. This was especially interesting in the case of electricity prices, which certain sources cite as usual descriptors for the natural gas demand, but which were found to be 0.05 significant in only 12 of the 48 cases and thus dropped from the model.

The consumption and price of natural gas are endogenous variables as both are correlated to system shocks, such as unstable governments or weather-related events. As an alternative to the use of least squares regression to fit the model given in (1), a two-stage least squares approach could be employed with such instrumental variables as the number of gas producing wells, reserve estimates, and underground storage, to name only a few. However, this approach is not considered here, because the response (reaction) time of consumers' consumption habits to the shocks is much longer than that to the spot prices set by the market every day.

3.2. Pooled Multiple Linear Regression (PMLR). Now we address the issue of how one can use the same regression formula for more than one state, which would create several classes of states where demand responds to changes in the descriptors in a similar mode.

Assume that we have split n collections of state time series into several classes, with the members of each class sharing a common set of regression parameters. Then the pooled data from the groups would be regressed at the same time, creating *pooled regressions*.

Let $\mathbf{I} = \{I_1, I_2, \ldots, I_K\}$ be a partition of I, and consider the model:

$$\widehat{C'_{i,t}} = \beta_{0,i} + \beta_{1,k} P'_{i,t} + \beta_{2,k} C'_{i,t-12} + \beta_{3,k} T'_{i,t} + \beta_{4,k} O'_t + \beta_{5,k} N'_{i,t}; \quad t \in T, \quad \forall i \in I_k, \quad k = 1, 2, \ldots, K. \tag{2}$$

Note that this model—called the Pooled Multiple Linear Regression (PMLR) model—has K sets of parameters for each regressor variable, except for the intercepts a_0^i, which we allow to be different for each state. In comparison, model (1) has n sets of parameters.

How should one define the partition $\mathbf{I}$ of the set of states? A good partition is expected to deliver groups of more or less congruent sizes, while maintaining a high individual R^2 value for each state. A good partition method should also be replicative (i.e., the same partition is obtained for the same

group of states), be fast enough, and support the statistical significance.

4. Dendrogram-GRASP Grouping Method (DGGM)

In this section, a combination of both grouping methods mentioned in [26] into a GRASP heuristic is proposed. The resulting technique inherits the replicative property of the dendrogram method, while retaining the statistical significance of the heuristic algorithm.

4.1. Dendrograms. Dendrograms are binary trees in which two observation vectors a and b form the (sub-)branches of a higher branch c, so that

(i) these two observation vectors are "closer" to each other than to any other observation d,

(ii) c is not an observation *per se*, but a new, artificial vector formed by some linear combination of a and b.

The term "closer" is interpreted with respect to some metric (e.g., the Euclidean metric), while the artificial observations are produced by the weighted combination method. Once the dendrogram is formed, it is cut down from the root and thus generating (sub-) dendrograms with the branches resulting from the cut. The height of the cut is determined according to one of several criteria (the number of subdendrograms produced, the maximum allowed membership for the subdendrogram, etc.) The leaves pertaining to a given subdendrogram will pool their regression data together and form one group for the PMLR.

Previous experiments [26] have shown that what is called the "average euclidean" metric [28] delivers satisfactorily high R^2 levels with a better homogeneity in the resulting groups than other linkage function options.

4.2. GRASP Heuristics. GRASP stands for Greedy Randomized Adaptive Search Process; it is a metaheuristic, that is, a general method designed to provide good—but not necessarily optimal—results in problems otherwise too complicated to find an optimal solution, especially combinatorial problems [29].

Summarily, our GRASP approach will start with a seed formed by several one-state groups; then, for each state, it will identify those groups that deliver higher R^2 values once the data for the current state is pooled with that of the group. This is called the Restricted List of Candidates or RLC. A group I_k from the RLC is chosen at random, and the current state is added to I_k, pooling its data with those already in the group. A number of swaps and movements are performed once the states are all in place, in order to try to improve the values of the resulting statistics R^2.

It is important to note that setting the values for the GRASP routine is rather subjective, since there is no definite objective to be achieved. Indeed, one cannot determine what number of groups is optimal, or which way is the best to define the greedy function. For example, one could prefer to increase the grouped R^2 value in each group rather than the average of the individual R^2s in that group, or vice versa. This is exemplified by the function

$$F_w\left(I_k\right) = \omega R^2_{I_k} + \frac{1-\omega}{\left|I_k\right|}\sum_{i\in I_k} R^2_i, \tag{3}$$

where

$$R^2_{I_k} = 1 - \frac{\sum_{t\in\mathbf{T}, i\in I_k}\left(y_{i,t} - \widehat{y}_{i,t}\right)^2}{\sum_{t\in\mathbf{T}, i\in I_k}\left(y_{i,t} - \overline{y}_{I_k}\right)},$$
$$R^2_i = 1 - \frac{\sum_{t\in\mathbf{T}}\left(y_{i,t} - \widehat{y}_{i,t}\right)^2}{\sum_{t\in\mathbf{T}}\left(y_{i,t} - \overline{y}_i\right)}. \tag{4}$$

Here, $y_{i,t} = \ln C'_{i,t}$, and $\overline{y}_\iota$ is understood as the average of all of the observations belonging to ι if the latter is a state (e.g., $\iota = i$) or as the average of the observations of the states in ι, if the latter is a set of states (e.g., $\iota = I_k$).

For the local search, we handle the improvement function $G_\tau(I_k, I_\ell, I_i)$, which is used when deciding if it is convenient to move state i from group k to group ℓ. It is parametrized by the improvement weight τ:

$$G_\tau\left(I_k, I_l, I_i\right) = (1-\tau)\,\frac{R^2_{I_k} + R^2_{I_\ell}}{2} + \tau R^2_i. \tag{5}$$

4.3. Dendrogram-GRASP Algorithm. The following algorithm is used to classify the set of 48 contiguous states of the United States into groups that share a common regression function.

(1) Initialize the values for each of the time series in each of the 48 states. Set a seed size s_{Seed}, a maximum number of groups s_{Max}, a RLC size s_{RLC}, an individual/grouped R^2 weight $\omega \in [0,1]$, an individual/grouped threshold $\varphi \in (0,1)$, an improvement weight $\tau \in (0,1)$, a relative improvement threshold $\psi \in [0,1]$, and a maximum number of local search steps, s_{ls}.

Seed Selection

(2) Perform an IMLR on each of the 48 sets of time series, obtaining $\alpha_{j,i}$, $j = 0, \ldots, 6$, $i \in I$.

(3) Form a dendrogram of 48 leaves with the vectors α, using the average euclidean mean as the linkage function, and cut it so that there are exactly s_{Seed} subtrees.

(4) Select the state with the highest R^2_i from each of the obtained groups and call it the kth group's leader. Define the one-state groups obtained as the partition $\mathbf{I}_k$. All the nonselected (spare) states form the set *Active*.

Greedy Process

(5) For each state x in the set *Active*,

(a) pool the data of x with the data of each of the formed groups and perform a pooled

regression; select a number of s_{RLC} groups with the highest value of the greedy function F_w and form the RLC;

(b) choose randomly one of the groups from the RLC, for example, I_a.

(i) If none of the candidate groups in the RLC delivers $F_w(I_k) > \varphi$ and we have not yet reached the maximum number of groups s_{Max}, create a new group $I_x = \{x\}$ containing only x, remove x from the active set, and update all the parameters.

(ii) Otherwise, assign x to I_a, remove x from the active set, and update all the parameters.

(6) All of the states are now partitioned into the groups, and we can begin the local search.

Local Search

(7) For $l = i$ to $l = s_{\mathrm{ls}}$,

(a) randomly select one of the formed groups, I_a, and one state in that group, x; select another group, I_b; compute $g_1 = G_\tau(I_a, I_b, x)$;

(b) remove x's data from I_a and pool the same data of x with I_b; compute $g_2 = G_\tau(I_a, I_b, x)$;

(c) if $g_1 \geq (1 + \psi)g_2$, remove x from I_b and return it to I_a; otherwise, continue.

(8) Report the obtained groups as the desired partition.

(9) End.

4.4. Partition Similarity. To determine the similitude of two partitions, we will use an expression that, roughly speaking, counts the number of coincidences found in two partitions and divides it by the number of total possible coincidences, given the sizes of the groups in each partition. While there are many disputable ways to measure the similitude between partitions with a different number of elements, this method was chosen because of its normality. Indeed, it will always return 1 when both partitions are identical and will always return 0 when there are no coincidences between two partitions, that is, when no two states share a group in both partitions and no state is single-grouped in both partitions.

Let $\mathbf{I} = \{I_1, I_2, \ldots, I_K\}, \mathbf{J} = \{J_1, J_2, \ldots, J_L\}$ be two arbitrary partitions of the set of states, with $I_i = \{I_1^i, I_2^i, \ldots, I_{k^i}^i\}, i = 1, \ldots, K$, and $J_j = \{J_1^j, J_2^j, \ldots, J_{l^j}^j\}, j = 1, \ldots, L$.

The function $\mathbf{a}_{\mathbf{I},\mathbf{J}}$ defined by

$$\mathbf{a}_{\mathbf{I},\mathbf{J}}(I_i) = \begin{cases} 1, & \text{if } I_i = \{m\} = J \text{ for any } J \in \mathbf{J}, \\ 0, & \text{otherwise,} \end{cases} \quad I_i \in \mathbf{I}, \tag{6}$$

assumes the value 1 if group I_i contains a single state in partition $\mathbf{I}$ and this state also forms a group-singleton in partition $\mathbf{J}$.

For every pair of states, we will assess if they share a group in a given partition using the following function $\mathbf{b}_{\mathbf{J}}$:

$$\mathbf{b}_{\mathbf{J}}(m, n) = \begin{cases} 1, & \text{if } m, n \in J_j, \text{ for any } j; \\ 0, & \text{otherwise,} \end{cases} \quad m, n \in \mathbf{I}. \tag{7}$$

In case the function $\mathbf{a}_{\mathbf{I},\mathbf{J}}$ has the value of 1, we say that we have a (one-state) coincidence, which means that the state has been found incompatible with other states twice, no matter which method formed partitions $\mathbf{I}, \mathbf{J}$.

Similarly, if the function $\mathbf{b}_{\mathbf{J}}$ returns 1 for two states *in a group from the partition* $\mathbf{I}$, we say that we have a (two-state) coincidence; that is, in both partitions, the two states are members of the same group.

To measure the number of coincidences between two partitions, we use the function:

$$\begin{aligned} &C_q(I_i, \mathbf{I}, \mathbf{J}) \\ &\quad = \mathbf{a}_{\mathbf{I},\mathbf{J}}(I_i) + (1 - \mathbf{a}_{\mathbf{I},\mathbf{J}}(I_i)) \\ &\qquad \times \left(\sum_{m \in I_i} \sum_{n \in I_i, n \neq m} \frac{q\mathbf{b}_{\mathbf{J}}(m, n) + (1 - q)}{2} \right), \end{aligned} \tag{8}$$

for $I_i \in \mathbf{I}, q = \{0, 1\}$.

If the parameter q equals 1, then the function C_q counts the number of either type of coincidences that couples of states reveal in the group I_i in comparison to the groups they belong to in the partition $\mathbf{J}$. Conversely, if $q = 0$, then we simply count the total number of possible coincidences for the states in the group $I_i \in \mathbf{I}$. Note that the function C_q is not necessarily symmetric with respect to the pairs of partitions: $C_q(I_i, \mathbf{I}, \mathbf{J})$ need not have the same value as $C_q(I_i, \mathbf{J}, \mathbf{I})$.

The similitude function used Sim is defined as follows:

$$\mathrm{Sim}(\mathbf{I}, \mathbf{J}) = \frac{\sum_{I_i \in \mathbf{I}} C_1(I_i, \mathbf{I}, \mathbf{J}) + \sum_{J_j \in \mathbf{J}} C_1(J_j, \mathbf{J}, \mathbf{I})}{\sum_{I_i \in \mathbf{I}} C_0(I_i, \mathbf{I}, \mathbf{J}) + \sum_{J_j \in \mathbf{J}} C_0(J_j, \mathbf{J}, \mathbf{I})}. \tag{9}$$

Notice that if there is at least one group in either partition containing more than one element, then C_0 for that group is at least 1, whereas if there exists no such group in either partition, then $\mathbf{a}_{\mathbf{I},\mathbf{J}}(a) = 1$ and consequently $C_0(a, \mathbf{I}, \mathbf{J}) = 1$ for any $a \in \mathbf{I} \cap \mathbf{J}$. Therefore, the denominator is never 0, which makes this function well defined.

Lemma 1. *Let* $\mathbf{I}$ *and* $\mathbf{J}$ *be two partitions of the set* $I = \{1, 2, \ldots, n\}$*, and let function* Sim *be defined by* (9)*. The following statements are true:*

(1) $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) = \mathrm{Sim}(\mathbf{J}, \mathbf{I})$*;*

(2) $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) = 1$ *if and only if* $\mathbf{I} = \mathbf{J}$*;*

(3) $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) = 0$ *if and only if there are neither one-state nor two-state coincidences between* $\mathbf{I}$ *and* $\mathbf{J}$*;*

(4) $0 \leq \mathrm{Sim}(\mathbf{I}, \mathbf{J}) \leq 1$.

Proof. (1) This is easy to see from the structure of the function.

(2) Let $\mathbf{I} = \mathbf{J}$. If $I_i = \{m\} = J_k$ for some i and k, then $C_1(I_i, \mathbf{I}, \mathbf{J}) = C_0(I_i, \mathbf{I}, \mathbf{J})$. Otherwise, if the order of I_i is greater than one, then the second term in (8) (the definition of C_q) assumes the same value no matter whether $q = 1$ or $q = 0$. Therefore, the numerator and denominator in Sim are equal.

Conversely, if there exists one I_i such that $I_i \neq J$ for all $J \in \mathbf{J}$, then $C_1(I_i, \mathbf{I}, \mathbf{J})$ is strictly less than $C_0(I_i, \mathbf{I}, \mathbf{J})$. Since $C_1(J_i, \mathbf{I}, \mathbf{J}) \leq C_0(J_i, \mathbf{I}, \mathbf{J})$, it follows that the numerator in (9) (defining Sim) is strictly smaller than the denominator, and therefore $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) < 1$.

(3) If there is at least one one-state coincidence, or a two-state coincidence, then the numerator in Sim is larger than 0, and therefore $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) > 0$.

Conversely, since C_q is nonnegative for every value of q, $\mathrm{Sim}(\mathbf{I}, \mathbf{J}) = 0$ means that both terms in the numerator are zero, which is only possible if $\mathbf{a}_{\mathbf{I},\mathbf{J}}(I_i) = \mathbf{a}_{\mathbf{I},\mathbf{J}}(J_j) = 0$ for every member of $\mathbf{I}$ and $\mathbf{J}$, *and* $\mathbf{b}_{\mathbf{J}}(m, n) = 0$ for every $m, n \in I$, which means that there is no coincidence of any type between these two partitions.

(4) The first inequality follows from the fact that both the numerator and denominator in (9) are positive. The second inequality comes from the same argument as in item (2); that is, the numerator is either equal or strictly less than the denominator. □

5. Experimental Results

This section presents the results of the numerical experimentation performed on a number of times series pertaining to each of the 48 data sets. The values for the historical natural gas prices, consumption, and number of consumers, as well as the oil spot prices were taken from the US Energy Information Agency, whereas the temperature figures for each state were obtained from the US Department of Commerce National Oceanographic and Atmospheric Agency [30].

5.1. IMLR Results. The first step was to perform the IMLR for the 48 sets of time series; this provided the regression parameters for the dendrogram formation. The five time series corresponding to every state had 240 monthly observations each.

Individual regression models showed regression R^2 coefficients with the average of 0.77 and the minimum of 0.61. The normality and heteroskedasticity were not tested due to the use of Robust Regression with Huber weights. Randomness of the residuals was tested, and high P values were found for many states.

5.2. Dendrogram-GRASP Grouping Results. There are two main aspects we wanted to consider when evaluating the effectiveness of the Dendrogram-GRASP approach: how replicative it is, and how good a partition is produced. The first issue is evaluated by examining how good and how similar the partitions are that come from the same seed (as opposed to those that come from randomly generated seeds). The goodness of one partition is measured with the average group [state] coefficient of determination, $R^2_{I_k}$ $[R^2_i]$, calculated across all the groups [states] of the partitions.

There are, however, a number of different design parameters that should be included in the experimentation. Each experimental observation consists of the generation of 10 partitions, using the following parameters.

(i) A seed choice: the dendrogram seed (DDR), a random seed common to all 20 partitions (FIX), and a random seed for each partition (RND).

(ii) The individual versus grouped R^2 weight, ω, which determines what is more important when adding a new state to an existing group in the GRASP routine: values considered in the experimentation are $\omega = 0$ (only the single states' R^2s are considered), 0.5, and 1 (only the groups' R^2s are important).

(iii) The new group threshold, φ: the closer the value of φ to 1, the more likely new single-state groups will be created in the GRASP routine. The tested values are $\varphi \in \{0.90, 0.95\}$.

(iv) The length of the restricted candidate list, s_{RCL}: the values considered are $s_{\mathrm{RLC}} \in \{1, 5\}$.

(v) The number of local search moves: $s_{\mathrm{ls}} \in \{0, 100\}$.

(vi) The local search individual/grouped R^2 weight, τ: considered values are $\tau \in \{0, 0.66, 1\}$.

The starting number of groups was fixed at 10, and the maximum number of groups allowed was set at 15. Each combination of levels was replicated 20 times. This resulted in 5760 experimental observations.

In each observation, we calculated the average similitude between the various partitions involved, as well as their similitude with a randomly created partition. The compared similitudes were as follows:

(i) the average similitude of the dendrogram partition to each of the 20 GRASP partitions (DG);

(ii) the average similitude of a random partition and each of the 20 GRASP partitions (GR);

(iii) the average similitude of the 20 GRASP partitions among themselves (GG).

The first part of the analysis consisted in testing all the experimental observations. After that, only the most convenient levels were kept.

Tables 1 and 2 present a summary of the results of the experimental runs. The first three data columns show the average similarities for each of the three comparisons of interest, whereas the last two columns show the average of the individual and grouped coefficients of determination.

A quick look at this table suggests that the similitude figures are characteristically low: the average similarity of an arbitrary partition to a randomly formed one, calculated using all the observations, is 0.0947. This will be called the partitions' randomness. If columns 3 and (particularly) 5 approach the average randomness for this experiment, the partition method is not very efficient. This especially concerns the cases $s_{\mathrm{ls}} = 5$, $\omega = 0$, and $\tau = 1$, whose similarity measures are fairly low. Luckily enough, in all these cases the average GG similarities were found to be statistically different

Table 1: Experimental Results I.

Factor	Level	Av. similitude			Av. R^2 values	
		DG	GR	GG	Av. $R^2_{I_i}$	Av. $R^2_{I_k}$
φ	0.90	0.145	0.079	0.178	0.503	0.535
	0.95	0.149	0.079	0.177	0.499	0.537
Seed	DDR	0.182	0.083	0.194	0.513	0.568
	FIX	0.130	0.077	0.154	0.489	0.521
	RND	0.128	0.077	0.184	0.501	0.520
ω	0	0.146	0.082	0.136	0.427	0.564
	0.5	0.147	0.079	0.183	0.534	0.535
	1	0.148	0.077	0.213	0.542	0.511
τ	0	0.160	0.080	0.232	0.699	0.502
	0.66	0.141	0.079	0.150	0.455	0.554
	1	0.140	0.079	0.149	0.349	0.553

Table 2: Experimental Results II.

Factor	Level	Av. similitude			Av. R^2 values	
		DG	GR	GG	Max. $R^2_{I_i}$	Max. $R^2_{I_k}$
s_{RLC}	1	0.160	0.082	0.247	0.879	0.862
	5	0.134	0.076	0.107	0.879	0.871
s_{ls}	0	0.167	0.084	0.236	0.876	0.857
	100	0.126	0.075	0.119	0.882	0.876

(higher) than their respective GR similarities by making use of the Wilcoxon signed-rank (WSR) $\alpha = 0.95$ test.

The average R^2 values in columns 6 and 7 do not deviate much from the averages across all the observations, 0.602 and 0.624, respectively, with the exception of the grouped individual parameter $R^2_{I_i}$ for $\tau = 1$. It is clear that certain similarity values for some levels are consistently lower than others. There is, for example, a very large difference between the average DG similitude obtained using a DDR seed than using a RND or FIX seed and so on. Based on this, we decided to discard some of the levels whose averages are not only considerably lower, but also the observations for each level are determined to be different by a WSR test.

Now let us look at each of the level values we should consider to drop. The first level, the GRASP new group threshold φ, shows a very similar GG figure, and equally similar R^2 values. We decide to keep the factor levels intact, in case these figures change once other levels are removed.

Seeds are more difficult to assess. The FIX seed shows lower values than the DDG one, but still higher than the RND. Weight τ shows much better numbers in all but the grouped R^2 entry. Because of this, we pick it as the only label for the later study. On the contrary, ω is better at value 1, except again in the grouped R^2 column. This result for ω is very counter intuitive! However, the two values serve a similar purpose at different parts of the process, so this behavior might indeed be justified.

The factors s_{RLC} and s_{ls} were introduced to add variation in the GRASP routine, and their results appear separated in Table 2. This is because, while their similitude values work in the same way as the other factors, the R^2 measurements per observation are not the average across all 10 partitions in the observation, but rather the maximum obtained. In a common GRASP routine, the process will be repeated several times and the best solution will be adopted. For our case, this means that we should choose the best of the 20 partitions in each observation, and this decision becomes the result for that observation. Arguably, both the individual and grouped average maximum coefficients of determination seem to show little difference. In particular, the differences are deemed not large enough to justify the trade-off with similarity in all cases. While this was expected from the extended RLC size, the poor results obtained by the local search suggest that we should rethink our local search procedure in the future.

Based on similarity alone, we decided to eliminate the poorest levels and kept only a single-group state list and a zero-swaps local search for the second part of the analysis. After deciding to drop several levels, we will rewrite the results table including only the accepted levels, to see how the figures change once the poorest results are winnowed.

The much smaller Table 3 is the consequence of fixing $\omega = 1$, $\tau = 0$, $s_{\text{RLC}} = 1$, and $s_{\text{ls}} = 0$ and eliminating the RND seed choice, which results in 100 observations. Now the similitudes look much better: we have the sample average of 0.438 and the maximum of 0.477, which means that, for the parameters chosen, the similitudes obtained are remarkably higher than the average randomness.

For the first factor, φ, the similitudes are of little difference, the same as the determination coefficients in all accounts. However, for the seed levels, the DDR seed clearly favors similitude between the seed and the resulting partition.

Table 3: Experimental Results III.

Factor	Level	Av. similitude			Av. R^2 values	
		DG	GR	GG	Av. R_i^2	Av. $R_{I_k}^2$
φ	0.90	0.171	0.077	0.432	0.760	0.340
	0.95	0.178	0.085	0.432	0.759	0.349
Seed	DDR	0.238	0.090	0.454	0.757	0.432
	FIX	0.143	0.074	0.365	0.760	0.299
	RND	0.143	0.079	0.477	0.761	0.302

Similitude among resulting partitions is also good at the RND partition, which could indicate the particular FIX seed was initially a bad choice when compared to either an average partition seed or one selected in a methodical way.

The coefficients of determination R^2 present a rather interesting development. The individual coefficients R_i^2 are decent enough when compared to the ones from the dropped levels, but there is a dramatic drop in the group figures $R_{I_k}^2$, which decreased from an average of around 0.53 to as low as 0.299. This happens because, while focusing on similitude, we chose in favor of $s_{\text{ls}} = 0$, which yields the mean $R_{I_k}^2$ of only 0.366, as opposed to the 0.706 value obtained after fixing $s_{\text{ls}} = 100$. In Table 2, however, we see the greater max $R_{I_k}^2$ because it was relevant to that table. If we were to remake Table 3 using the value of $s_{\text{ls}} = 100$ for this level, similitudes would fall around 10%, but the average group determination coefficients $R_{I_k}^2$ would increase to roughly 0.43, which is much better than that with $s_{\text{ls}} = 0$. Maximum values for the different R^2s, correspondent to those in Table 3, remain mostly unchanged.

6. Concluding Remarks

In this paper, we propose and justify a heuristic method to group several zones based on a regression function that estimates several factors related to the natural gas demand. The groups thus obtained share key information regarding the behavior of natural gas-related historic econometric data.

We start by developing a linear regression model that correlates natural gas historic residential consumption and several explicative variables, such as the residential price, number of consumers, and temperature. This model, inspired by several examples in the literature, fits well the time series employed and has good predictive power, but it is by no means the only one that can be used nor necessarily the best.

The results of each of the 48 regressions performed are then used to create dendrogram-based partitions, which are in turn used as the starting point in a GRASP routine. The latter, while tending to form rather dissimilar partitions (compared to the dendrogram grouping), has the advantage of adding statistical significance to all the regressions in all the groups formed.

We tested several parameters in an experimental design consisting of more than 4300 observations, six factors, and two or three levels per factor. Using ad hoc and nonparametric selections, we tried to obtain a good combination of parameters, namely, one that delivers high similitude between partitions obtained from the same seed and a satisfactory goodness of the pooled regressions.

Similitude is measured by a standardized function which equals 0 if there are no common groups between two partitions of a fixed set and 1 if both partitions are identical. We were able to obtain experimental conditions with similitudes (mostly) above 0.43, which are deemed good considering that the average randomness of a partition in the study is around 0.09.

It is encouraging that, using the regression function herein proposed, the GRASP routine worked well by itself and also when combined with the dendrogram partitioning method. Unfortunately, the inclusion of randomness did not provide for good results, as it offered no increase in goodness of the partitions but a considerable decrease in similitude when a long RLC was used. The proposed local search approach was found to have a negative impact on the similitude values, though not overly so. However, at the same time it did affect heavily the values of the grouped coefficients of determination when the maximum values were considered in the selection but the averaged values were looked into in the end results. The "goodness" of the regressions, as discussed, must then be judged with a more nuanced approach.

The entire work frame summarized here is intended to provide a way to identify individuals (states, in this case) with common econometric behavior among themselves by means of statistically significant information. Such results used to help us in the past in the context of optimization theory (by greatly decreasing the number of variables in stochastic problems), and we believe this technique has other applications in economic analysis.

The planned future work includes enhancing the robustness of the method by designing better GRASP RLC and local search procedures, trying sampled regressions when forming large groups to gain on time and studying how different data sets and regression models would work in combination with the Dendrogram-GRASP approach proposed in the paper.

Conflict of Interests

The authors declare that there is no conflict of interests for any of the authors of the paper.

Acknowledgments

The research activity of the first author was financially supported by the R&D Department (Cátedra de Investigación)

CAT-174 of the Tecnológico de Monterrey (ITESM), Campus Monterrey and by the SEP-CONACYT Project CB-2008-01-106664, Mexico. Also, the work of the fourth author was supported by the National Council of Science and Technology (CONACyT) of Mexico as part of the Project CB-2011-01-169765, PROMEP 103.5/11/4330, and PAICYT 464-10.

References

[1] Energy Information Administration, "FERC Order 636: The Restructuring Rule," 2005, http://www.eia.doe.gov/emeu/steo/pub/document/textng.html.

[2] Energy Information Administration, "FERC Policy on System Ownership Since 1992," 2005, http://www.eia.doe.gov/oil_gas/natural_gas/analysis_publications/ngmajorleg/fercpolicy.html.

[3] A. Soto, "FERC Order 636 & 637," 2008, http://www.aga.org/Legislative/issuesummaries/FERCOrder636637.html.

[4] IHS Engineering, "EC Proposes New Legislation for European Energy Policy," 2007, http://engineers.ihs.com/news/eu-en-energy-policy-9-07.html.

[5] Environmental Protection Agency, "The Impacts of FERC Order 636 on coal mine gas project development," 2008, http://www.epa.gov/cmop/docs/pol004.pdf.

[6] K. T. Midthun, *Optimization Models for Liberalized Natural Gas Markets*, Norwegian University of Science and Technology. Faculty of Social Science and Technology Management Department of Industrial Economics and Technology Management, Trondheim, Norway, 2009.

[7] M. J. Doane and D. F. Spulber, "Open access and the evolution of the US spot market for natural gas," *Journal of Law and Economics*, vol. 34, no. 2, pp. 447–517, 1994.

[8] R. Gutiérrez, A. Nafidi, and R. G. Sánchez, "Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model," *Applied Energy*, vol. 80, no. 2, pp. 115–124, 2005.

[9] F. K. Lyness, "Gas demand forecasting," *Journal of the Royal Statistical Society—Series D*, vol. 33, no. 1, pp. 9–12, 1984.

[10] S. Dempe, V. Kalashnikov, and R. Z. Ríos-Mercado, "Discrete bilevel programming: application to a natural gas cash-out problem," *European Journal of Operational Research*, vol. 166, no. 2, pp. 469–488, 2005.

[11] V. V. Kalashnikov and R. Z. Ríos-Mercado, "A natural gas cash-out problem: a bilevel programming framework and a penalty function method," *Optimization and Engineering*, vol. 7, no. 4, pp. 403–420, 2006.

[12] N. Keyaerts, L. Meeus, and W. D'haeseleer, "Analysis of balancing-system design and contracting behavior in the natural gas markets," in *Proceedings of the European Doctoral Seminar on Natural Gas Research*, Delft, The Netherlands, 2008.

[13] H. G. Huntington, "Federal price regulation and the supply of natural gas in a segmented field market," *Land Economics*, vol. 54, no. 3, pp. 337–347, 1978.

[14] A. Tomasgard, F. Romo, M. Fodstad, and K. T. Midthun, "Optimization models for the natural gas value chain," in *Geometric Modeling, Numerical Simulation, and Optimization: Applied Mathematics at SINTEF*, optimization models for the natural gas value chain, Springer, 2007.

[15] C. Nelder, "Natural Gas Price ForecastThe future of Natural gas: It's time to invest," 2009, http://www.energyandcapital.com/articles/natural-gas-price-forecast/916.

[16] EuroGas, "EuroGaslong term outlook to 2030," 2009, http://www.eurogas.org/uploads/media/Statistics_Eurogas_LT_Outlook_2007-2030_Final_25.11.10.pdf.

[17] P. W. MacAvoy, *The Natural Gas Market: Sixty Years of Regulation and Deregulation*, Yale University Press, New Haven, Conn, USA, 2000.

[18] K. T. Talluri and G. J. van Ryzin, *The Theory and Practice of Revenue Management*, Springer, New York, NY, USA, 2004.

[19] P. W. Keat and P. K. Y. Young, *Managerial Economics: Economic Tools for Today's Decision Makers*, Prentice Hall, Englewood Cliffs, NJ, USA, 2006.

[20] J. M. Gowdy, "Industrial demand for natural gas. Inter-industry variation in New York state," *Energy Economics*, vol. 5, no. 3, pp. 171–177, 1983.

[21] J. G. Beierlein, J. W. Dunn, and J. G. McConnon Jr., "The demand for electricity and natural gas in the northeastern United States," *The Review of Economics and Statistics*, vol. 63, no. 3, pp. 403–408, 1981.

[22] W. T. Lin, Y. H. Chen, and R. Chatov, "The demand for natural gas, electricity and heating oil in the United States," *Resources and Energy*, vol. 9, no. 3, pp. 233–258, 1987.

[23] N. Krichene, "World crude oil and natural gas: a demand and supply model," *Energy Economics*, vol. 24, no. 6, pp. 557–576, 2002.

[24] S.-H. Yoo, H.-J. Lim, and S.-J. Kwak, "Estimating the residential demand function for natural gas in Seoul with correction for sample selection bias," *Applied Energy*, vol. 86, no. 4, pp. 460–465, 2009.

[25] Energy Information Administration, "Natural Gas Model Description," 1999, http://www.eia.doe.gov/oil_gas/natural_gas/analysis_publications/ngmajorleg/ferc636.html.

[26] V. V. Kalashnikov, T. I. Matis, and G. A. Pérez-Valdés, "Time series analysis applied to construct US natural gas price functions for groups of states," *Energy Economics*, vol. 32, no. 4, pp. 887–900, 2010.

[27] V. V. Kalashnikov, G. A. Pérez-Valdés, A. Tomasgard, and N. I. Kalashnykova, "Natural gas cash-out problem: bilevel stochastic optimization approach," *European Journal of Operational Research*, vol. 206, no. 1, pp. 18–33, 2010.

[28] Mathworks Inc, "Statistics Toolbox: Linkage," 2008, http://www.mathworks.com.

[29] P. Festa and M. G. C. Resende, "GRASP: an annotated bibliography," in *Essays and Surveys in Metaheuristics*, pp. 325–367, Citeseer, 2002.

[30] "NOAA Satellite and Information Service," 2010, ftp://ftp.ncdc.noaa.gov/pub/data/cirs/.