Available online at www.sciencedirect.com

SciVerse ScienceDirect

Transportation Research Procedia 00 (2016) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

20th EURO Working Group on Transportation Meeting, EWGT2017, 4-6 September 2017, Budapest, Hungary

The Cost of Transportation : Spatial Analysis of US Fuel Prices

Juste Raimbault [a,b,*], Antonin Bergeaud [c]

[a] *UMR CNRS 8504 Géographie-cités, 13 rue du Four, Paris, France*
[b] *UMR-T IFSTTAR 9403 LVMT, Cité Descartes, Champs-sur-Marne, France*
[c] *Paris School of Economics - EHESS, Paris, France*

Abstract

The geography of fuel prices has many various implications, from its significant impact on accessibility to being an indicator of territorial equity and transportation policy. In this paper, we study the spatio-temporal patterns of fuel price in the US at a very high resolution using a newly constructed dataset collecting daily oil prices for two months, on a significant proportion of US gas facilities. These data have been collected using a specifically-designed large scale data crawling technology that we describe. The behavior of spatial autocorrelation suggests the use of specific spatial econometric methods to study the role of explanatory variables that are either geographical or temporal. We study the influence of socio-economic variables, by using complementary methods: Geographically Weighted Regression to take into account spatial non-stationarity, and Multi-level modeling to condition both at the state and county level. The former yields an optimal spatial range roughly corresponding to stationarity scale, and significant influence of variables such as median income or wage per job, with a non-simple spatial behavior that confirms the importance of geographical particularities. On the other hand, multi-level modeling reveals a strong state fixed effect, and also a non-negligible county effect. Through the combination of such methods, we unveil the superposition of a governance process with a local socio-economical spatial process. Results are furthermore consistent across the different dates. We discuss one important application that is the elaboration of locally parametrized car-regulation policies.

Peer-review under responsibility of the Scientific Committee of EWGT2017.

Keywords: Fuel Price, Data Crawling, Spatial Analysis, Geographically Weighted Regression, Multi-level Modeling

1. Introduction

What drives the price of fuel? Using a new database on oil price at a gas station level collected during two months, we explore its variability across time and space. Variation in the cost of fuel can have many causes, from the crude oil price to local tax policy and geographical features, all having heterogeneous effect in space and time. If the evolution of the average fuel price in time is an indicator that is carefully followed and analyzed by many financial institution, its variability across space remain a rather unexplored topic in the literature. Yet, such differences can reflect variation

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: juste.raimbault@polytechnique.edu

2214-241X
Peer-review under responsibility of the Scientific Committee of EWGT2017.

in more indirect socio-economic indicators such as territorial inequalities and geographical singularities or consumer preferences.

There exists to our knowledge no systematic mapping in space and time of retail fuel prices for a country. The main reason is probably that the availability of data have been a significant obstacle. It is also likely that the nature of the problem may also have influence, as it lies at the crossroad of several disciplines. While economists study price elasticity and measurement in different markets, transportation geography with method such as transportation prices in spatialized models, puts more emphasis on spatial distribution than on precise market mechanisms. Nevertheless, examples of somehow related works can be found. For example, Rietveld et al. (2001) studies the impact of cross-border differences in fuel price and the implications for gradual spatial taxation in Netherlands. At the country-level, Rietveld and van Woudenberg (2005) provides statistical models to explain fuel price variability across European countries. Macharis et al. (2010) models the impact of spatial fuel price variation on patterns of inter-modality, implying that the spatial heterogeneity of fuel prices has a strong impact on user behavior. With a similar view on the geography of transportation, Gregg et al. (2009) studies spatial distribution of gas emission at the US-state level. The geography of fuel prices also have important implications on effective costs, as shows Combes and Lafourcade (2005) by determining accurate transportation costs across urban areas for France. More closely related to our work, and using very similar daily open data for France, Gautier and Saout (2015) investigate dynamics of transmission from crude oil prices to fuel retail prices. However, they do not introduce an explicit spatial model of prices diffusion and do not study spatio-temporal dynamics.

In this paper we take the analysis one step further by proceeding to exploratory spatial analysis on US fuel prices. to unveil possible spatial patterns and how they can link to socio-economic properties.

The rest of the paper is organized as follows: in the next section, we describe a generic procedure and the tool used for a systematic data collection. We also present our dataset. In section 3 we conduct statistical analysis in order to study the spatio-temporal variation of fuel price and test the potential correlation with some covariates. Finally, in section 4 we discuss our results and conclude.

2. Dataset

Our dataset contain daily information on fuel price at the gas station level for the whole US mainland territory. These information have been constructed from self-reported fuel price and span almost the entire universe of gas station in the US. We start by describing data collection and then give some statistics about this new dataset.

2.1. Collecting large scale heterogeneous data

The availability of new type of data has induced consequent changes in various disciplines from social science (e.g. online social network analysis (Tan et al. (2013))) to geography (e.g. new insights into urban mobility or perspectives on “smarter” cities (Batty (2013))) including economics where the availability of exhaustive individual or firm level data is seen as a revolution of the field. Most studies involving these new data are at the interface of implied disciplines, what is both an advantage but also difficulties. For example misunderstandings between physics and urban sciences described in Dupuy and Benguigui (2015) are in particular caused by different attitudes towards unconventional data or divergent interpretations and ontologies of it. Collection and use of new data has therefore become a crucial stack in social-science. The construction of such datasets is however far from straightforward because of the incomplete and noisy nature of data. Specific technical tools have to be implemented but have often been designed to overcome one specific problem and are difficult to generalize. We develop such a tool that fills the following constraints that are typical of large scale data collection: (i) reasonable level of flexibility and generality; (ii) optimized performance through parallel collection jobs; (iii) anonymity of collection jobs to avoid any possible bias in the behavior of the data source. The architecture, at a high level, has the following structure:

- An independent pool of tasks runs continuously socket proxies to pipe requests through `tor`.
- A manager monitors current collection tasks, split collection between subtasks and launches new ones when necessary.

Table 1. Descriptive statistics on Fuel Price ($ per gallon)

Mean	Std. Dev.	p10	p25	p50	p75	p90
2.28	0.27	2.02	2.09	2.21	2.39	2.65

- Subtasks can be any callable application taken as argument destination urls, they proceed to the crawling, parsing and storage of collected data.

The application is open and its modules are reusable: source code is available on the repository of the project.[1] We constructed our dataset by using the tool continuously in time during two months to collect crowdsourced data available from various online sources.

2.2. Dataset

Our dataset comprises around $41 \cdot 10^6$ unique observations of retail fuel prices at the station level, spanning the period starting the 10^{th} of January 2017 and ending the 19^{th} of March 2017 and corresponds to 118,573 unique retail stations. For each of these stations, we associate a precise geographical location (city resolution). On average we have 377 price information by station. Prices correspond to a unique purchase mode (credit card, other modes such as cash being less than 10% in test datasets, they were discarded in the final dataset) and four possible fuel types: Diesel (18% of observations), Regular (34%), Midgrade (24%) and Premium (24%). The best coverage of stations is for Regular fuel type with on average 4,629 price information by county. We therefore choose to focus the study to this type of fuel, keeping in mind that further developments with the dataset may include comparative analysis on fuel types. Our final dataset thus contains 14,192,352 observations from 117,155 gas station, followed during 68 days. We further aggregate these data by day, taking the average of the observed price per gallon, to obtain a panel of 5,204,398 gas station - day observations.[2] Table 1 gives some basic descriptive statistics of on price data showing that the distribution of oil price is highly concentrated with a small skewness (the ratio of the 99^{th} to the 1^{st} percentile is 1.6). Finally, in the spatial analysis, we will also use socio-economic data at the county level, available from the US Census Bureau. We shall use the latest available, which most of the time implies relying to the 2010 Census).

3. Results

3.1. Spatio-temporal Patterns of Prices

Before moving to a more systematic study of the variation of fuel price, we propose a first exploratory introduction to give insight about its spatio-temporal structure. This exercise is a crucial stage to guide further analyses, but also to understand their implications in a geographical context. To explore the data, we built a simple web application which allow to map the data in space and time. This application is available on this page. We also show one example of mapping the data at the county level in Figure 1 where we used average price over the whole period. We clearly see regional patterns with the Southcentral and Southeast regions having the lowest prices and the Pacific cost and Northeast the highest prices. Of course, plotting aggregated data over the whole period does not bring much information about the time variation of the data. As we will show more in detail below most of the variation of fuel price occurs across space. A variance decomposition of fuel price yields only 11% of the total variance is explained by within gas station variations. Similarly, the Spearman's rank correlation coefficient between the gas station price of regular fuel in the first day of dataset and in the last day is 0.867, and the null hypothesis that these two information are independent is strongly rejected.

[1] at `https://github.com/JusteRaimbault/EnergyPrice`

[2] The panel is not balanced as prices are not reported every day in every station. The average gas station has information on price for 44 days (over 68).

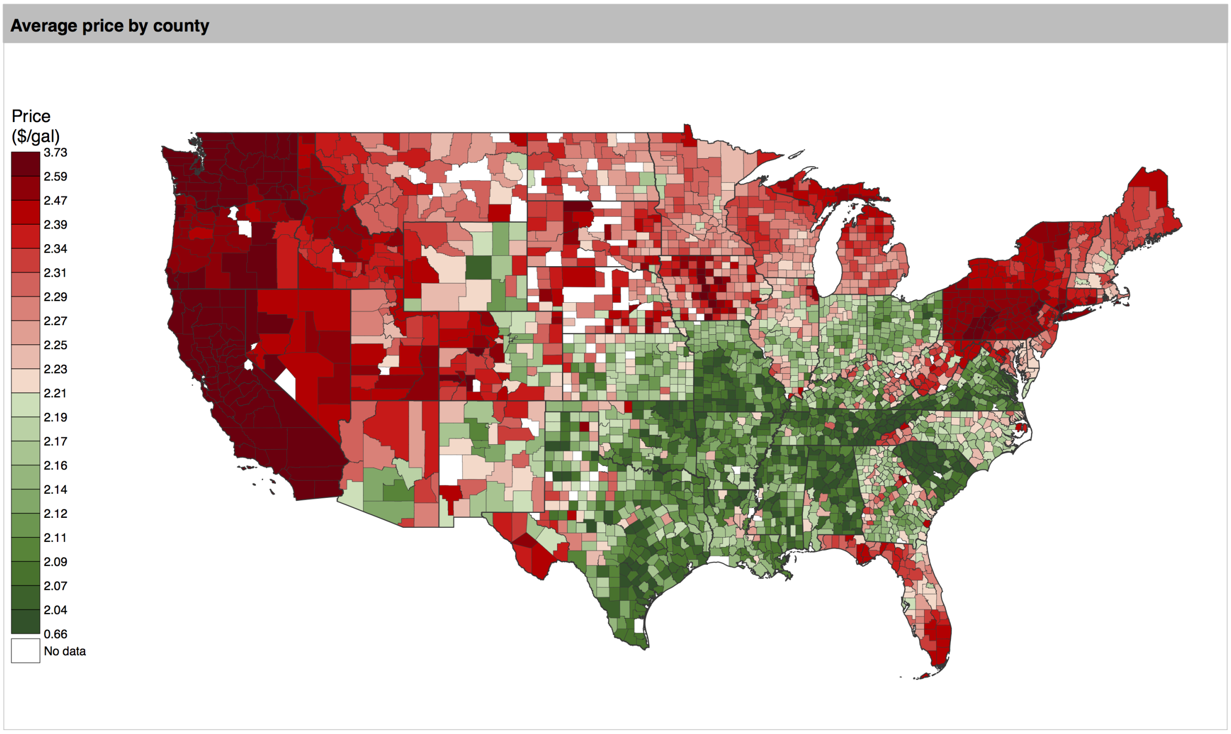

Fig. 1. Map of mean price for counties, regular fuel, averaged over the whole period.

3.2. Spatial Autocorrelation

We now turn to the study of spatial autocorrelation of prices, which can be seen as an indicator of spatial heterogeneity. We use the Moran index (Tsai (2005)), with spatial weights of the form $\exp\left(-d_{ij}/d_0\right)$ with d_{ij} being the distance between spatial entities i and j, and d_0 a decay parameter giving the spatial range of interactions accounted for in the computation. We show in Fig. 2 its variations in time and as a function of decay parameter. The fluctuations in time of the daily Moran index for low and medium spatial range, confirms geographical specificities in the sense of locally changing correlation regimes. These are logically smoothed for long ranges, as price correlations drop down with distance. The behavior of spatial autocorrelation with decay distance is particularly interesting: we observe a first regime change around 10km (from constant to piecewise linear regime), and a second important one around 1000km, both consistent across weekly time windows. We postulate that these correspond to typical spatial scales of the involved processes: the low regime would be local specificities and the middle one the state level processes. This behavior confirms that prices are non-stationary in space, and that therefore appropriate statistical techniques must be used to study potential drivers at different level. The two next subsections follow this idea and investigate potential explicative variables of local fuel prices, using two different techniques corresponding to two complementary paradigms: geographically weighted regression that puts the emphasis on neighborhood effects, and multi-level regression taking into account administrative boundaries.

3.3. Geographically Weighted Regression

The issue of spatial non-stationarity of geographical processes has always been a source of biased aggregated analyses or misinterpretations when applying general conclusions to local cases. To take it into account into statistical models, numerous techniques have been developed, among which the simple but very elegant Geographically Weighted Regression, that estimates non-stationary regressions by weighting observations in space similarly to kernel estimation methods. It was introduced in a seminal paper by Brunsdon et al. (1996) and has been consequently used and matured

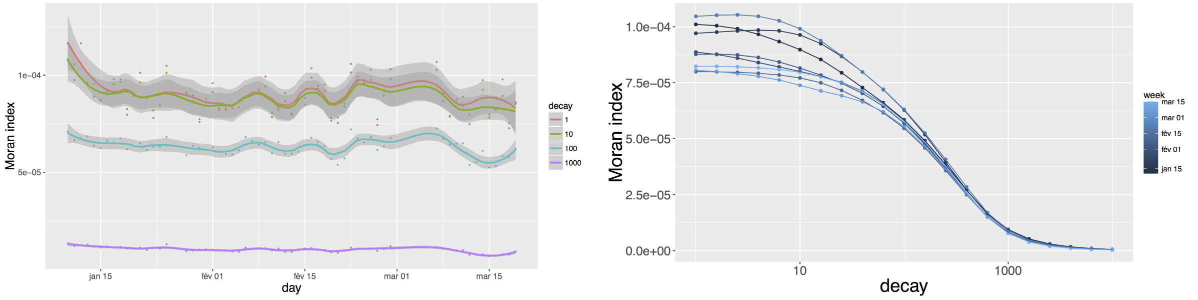

Fig. 2. **Behavior of Moran spatial-autocorrelation index.** (Left) Evolution in time of Moran index computed on daily time windows, for different decay parameter values. (Right) Moran index as a function of decay parameter, computed on weekly time windows.

since. The significant advantage of this technique is that an optimal spatial range in the sense of model performance can be inferred, and that the corresponding model gives effect of variables varying in space, revealing thus local effects that can occur at different spatial scales or across boundaries. We proceed to multi-modeling to find the best model and associated kernel and spatial range. The workflow is the following: (i) we generate all possible linear models from the five potential variables (income, population, wage per job, jobs per capita, jobs); (ii) for each model and each candidate kernel shape (exponential, gaussian, bisquare, step), we determine the optimal bandwidth in the sense of both cross-validation and corrected Akaike Information Criterion (AICc) which quantifies information included in the model, taking into account both model fit and number of parameters to avoid overfitting; (iii) models are fitted with this bandwidth. We choose the model with the best overall AICc, namely $price = \beta \cdot (income, wage, percapjobs)$ for a bandwidth of 22 neighbors and a gaussian kernel[3], with an AICc of 2900. The median AICc difference with all other models tested is 122. The global r-squared is 0.27, what is relatively good also compared to the best r-squared of 0.29 (obtained for the model with all variables, which clearly overfits with an AICc of 3010; furthermore, effective dimension is less than 5 as 90% of variance is explained by the three first principal components for the normalized variables).

The coefficients and local r-squared for the best model are shown in Fig. 3. The spatial distribution of residuals (not shown here) seems globally random, which confirms in a way the consistency of the approach: if a distinguishable geographical structure is found in residuals, it means that the geographical model or the variable taken into account fail to translate spatial structure. Let turn now to an interpretation of the spatial structures we obtain. First of all, the spatial distribution of model performance (local r-squared) reveals that regions where these simple socio-economic factors explain well prices cluster on the west coast, the south border, a north-east region from lakes to the east coast, and a stripe from Chicago to the south of Texas. The corresponding coefficients have different behaviors across the areas, suggesting different regimes[4]. For example, the influence of income in each region seems to be inverted when the distance to the coast increases (from north to south-east in the west, south to north in Texas, east to west in the east), what may be a fingerprint of different economic specializations. On the contrary, the regime shifts for wage show a clear cut between west (except around Seattle) and middle/east, that does not correspond to state-policies only as Texas splits in two. The same way, jobs per capita show an opposition between east and west, what could be due for example to cultural differences. These results are difficult to interpret directly, and must be understood as a confirmation that geographical particularities matters, as regions differ in regimes of role for each of the simple socio-economic-variables. Further precise knowledge could be obtained through targeted geographical studies including qualitative field studies and quantitative analyses, that are beyond the scope of this paper. An alternative that corresponds more to an econometric approach, is to refine explicative variables and their respective effects, without focusing on geographical structures: this will be done in the next subsection.

[3] note that the kernel shape does not have much influence as soon as gradually decaying functions are used

[4] we comment their behavior in areas where the model has a minimal performance, that we fix arbitrarily as a local r-squared of 0.5

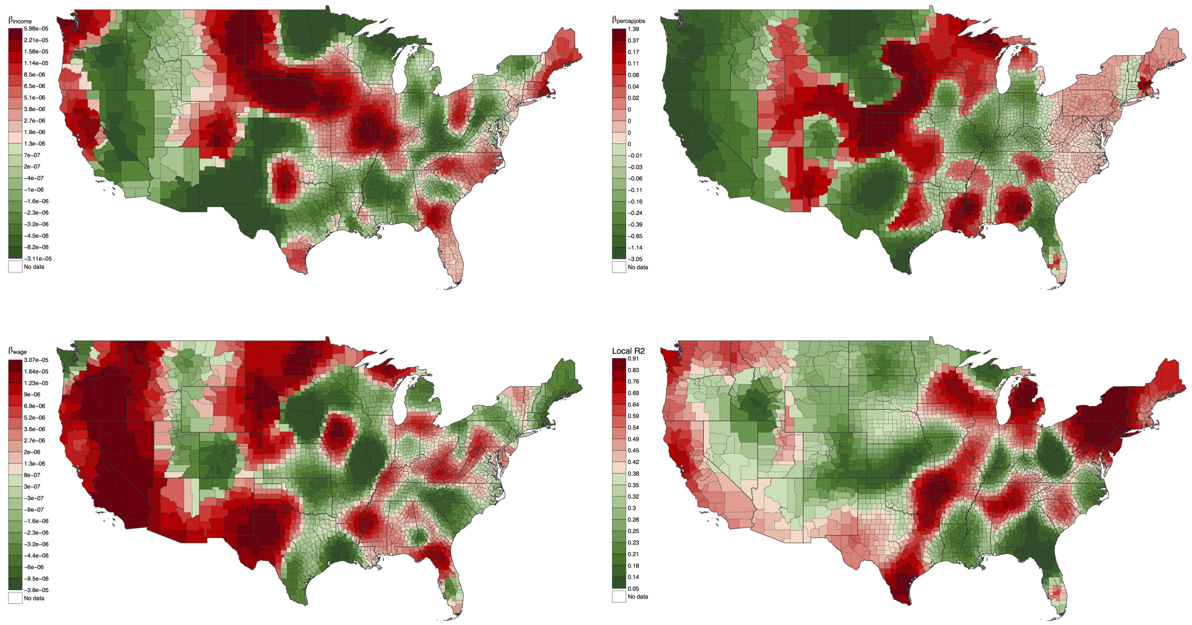

Fig. 3. **Results of GWR analyses.** For the best model in the sense of AICc, we map the spatial distribution of fitted coefficient, in order from left to right and top to bottom, β_{income}, $\beta_{percapjobs}$, β_{wage}, and finally the local r-squared values.

Finally, we extract the spatial scale of processes studied, we compute the distribution of nearest neighbors distance with the optimal bandwidth. It yields roughly a log-normal distribution, of median 77km and interquartile 30km. We interpret this scale as the spatial stationarity scale of price processes.

3.4. Multi-level Regression

An other approach to take into account geographical heterogeneities without explicitly including spatial relations between objects is to proceed to regressions with random effects depending on geographical aggregations at different levels. The method of multi-level regression implements this idea, and has been shown to be particularly adapted for geographical research Jones (1991). We take here into account two administrative levels, the state and the county. Explorations of section 3.1 suggest a strong state-effect that GWR cannot capture around boundaries, what confirms the relevance and complementarity of using Multi-level Regression.

4. Discussion

C : hesite pas si tu as des ides pour la discussion/conclusion, j'avais mis des ides plutot geo et qui me venaient en tete vite fait

4.1. On the complementarity of Econometric and Spatial Analysis methods

One important aspect of our contribution is methodological e.g. Chen and Truong (2012) integrates both in a single approach.

4.2. Towards localized car-regulation policies

C : knowing local price dynamics and their drivers should help designing specific regulations/taxing policies ?

5. Conclusion

References

Batty, M., 2013. Big data, smart cities and city planning. Dialogues in Human Geography 3, 274–279.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 28, 281–298.

Chen, D.R., Truong, K., 2012. Using multilevel modeling and geographically weighted regression to identify spatial variations in the relationship between place-level disadvantages and obesity in taiwan. Applied Geography 32, 737–745.

Combes, P.P., Lafourcade, M., 2005. Transport costs: measures, determinants, and regional policy implications for france. Journal of Economic Geography 5, 319–349.

Dupuy, G., Benguigui, L.G., 2015. Sciences urbaines: interdisciplinarités passive, naïve, transitive, offensive. Métropoles .

Gautier, E., Saout, R.L., 2015. The dynamics of gasoline prices: Evidence from daily french micro data. Journal of Money, Credit and Banking 47, 1063–1089.

Gregg, J.S., Losey, L.M., Andres, R.J., Blasing, T., Marland, G., 2009. The temporal and spatial distribution of carbon dioxide emissions from fossil-fuel use in north america. Journal of Applied Meteorology and Climatology 48, 2528–2542.

Jones, K., 1991. Specifying and estimating multi-level models for geographical research. Transactions of the institute of British geographers , 148–159.

Macharis, C., Van Hoeck, E., Pekin, E., Van Lier, T., 2010. A decision analysis framework for intermodal transport: Comparing fuel price increases and the internalisation of external costs. Transportation Research Part A: Policy and Practice 44, 550–561.

Rietveld, P., Bruinsma, F., Van Vuuren, D., 2001. Spatial graduation of fuel taxes; consequences for cross-border and domestic fuelling. Transportation Research Part A: Policy and Practice 35, 433–457.

Rietveld, P., van Woudenberg, S., 2005. Why fuel prices differ. Energy Economics 27, 79–92.

Tan, W., Blake, M.B., Saleh, I., Dustdar, S., 2013. Social-network-sourced big data analytics. IEEE Internet Computing 17, 62–69.

Tsai, Y.H., 2005. Quantifying urban form: compactness versus' sprawl'. Urban studies 42, 141–161.