Predicting Hourly Residential Electricity Demand in Ontario Using Signal Separation Approaches and Autoregressive Models

Barinder Thind, Lucas Wu, Nikola Surjanovic, & Zubia Mansoor

Simon Fraser University

Abstract

We predict the hourly residential electricity demand in Ontario by augmenting the provided dataset with publicly available data (such as zonal electric demand and population data). An Autoregressive (AR) model is used to provide accurate hourly predictions aggregated across sectors, while various signal separation techniques are used to provide insight into the proportion of electricity usage at each hour that is associated with the residential sector. Our model produces results that are in line with our expectations of the residential sector.

Keywords: independent component analysis, non-negative matrix factorization, autoregressive models, electricity demand

Introduction / Objectives

The ultimate objective is to predict hourly electricity demand in the residential sector in Ontario. We divide this task into two sub-tasks:

- Separating the hourly residential sector signal from the sector-aggregated hourly data.
- Building a model that does a proficient job in predicting sector-aggregated energy usage.

Due to the high resolution of the *sector-aggregated* data, the second problem is easier to address. We implement an AR model with regressors for this task, but a recurrent neural network (RNN) can also be used. However, we direct the most attention to the first task: separating the residential sector hourly signal. We do this by augmenting the provided dataset using publicly available data and making use of algorithms such as independent component analysis (ICA) and non-negative matrix factorization (NMF).

Exploratory Data Analysis

We hypothesize that the residential energy usage is highly correlated to the number of residents in each Independent Electricity System Operator (IESO) zone. The correlation between the zonal population and annual total energy is around 0.753.

Zone	Total Pop.	Total Dwellings	Total Energy Used
Southwest	4986667	2064784	29978201
Toronto	2937462	971895	50578053
East	1712995	738992	9334885
West	1145668	500394	17375629
Ottawa	883391	370217	11633400
Essa	563298	295136	8305927
Northeast	508982	247422	12465401
Niagara	431346	188877	5387975
Northwest	224034	111633	8026642
Bruce	66102	40033	328522

Table 1: This table is a companion to Figure 1. The total populations from Statistics Canada are calculated using the geographic areas in each zone.

Before diving into the methods, we explored the dataset, and some observations were:

- There is an outlier day in August 2003 due to a power outage. Fixing this was a matter of removing the outlier days.
- There is a noticeable difference between summer and winter electricity usage
- Some years had leap days, requiring appropriate indexing

Exploratory Data Analysis (cont.)

- The yearly sums of the hourly electricity data differ from the annual data (the hourly data measures additional information). We accounted for this difference in our predictions by taking the proportions of the two yearly totals (for each year) and multiplying the residential annual total by this factor.
- The following are maps of the geographic and zonal layouts which are used in the ICA

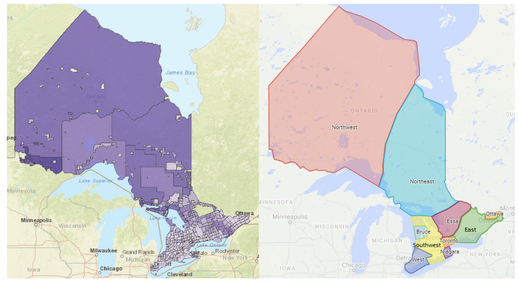

Figure 1: Census subdivisions from Statistics Canada (left). Information on the number of residences in various regions of Ontario was used in conjunction with IESO zonal (right) electricity consumption data to select zones for ICA decomposition. (See the licenses file for more information.)

Methods

Independent Component Analysis (ICA)

ICA can be used when individual components of several mixed signals are to be retrieved.

- The ICA model is defined as $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{x}$ is the vector of mixed signals, $\mathbf{A}$ is the mixing matrix, and $\mathbf{s}$ is the vector of independent components.
- Two basic assumptions: components of $\mathbf{s}$ are independent and have non-Gaussian distributions (see Discussion/Conclusions).
- We want $\mathbf{s} = \mathbf{W}\mathbf{x}$. This quantity is estimated by estimating $\mathbf{W}$.

We use ICA to estimate the proportion of electricity usage that arises from the residential sector. This is achieved by obtaining hourly zonal demand from the IESO. The key idea is that hourly demand in each of the 10 zones comes from a slightly different mixture of signals across the different sectors. To best separate the five sector signals, we identify five zones that should have highly different mixtures as suggested by Canadian census data. The ICA is run as follows:

1. Identify the geographical regions, defined by Statistics Canada, that belong to each IESO zone.
2. Calculate dwelling sizes for each of the IESO zones.
3. Using external resources, obtain the daily residential electricity pricing curves, which indicate off-peak and on-peak times.
4. Identify 5 IESO zones with the highest variability in their signal patterns (maximizing independence).
5. ICA Implementation:
   1. For each day, recover 5 (unmixed) signals and formulate as curves
   2. Find the unmixed signal that is closest in (normalized) distance to the pricing curve from (3). This becomes our residential signal.
   3. Repeat for each day and select the signal that is closest to the curve from (3) across all days to form the basis for our "weights" (i.e., the proportion of residential electricity demand relative to other sectors). Weights are obtained by comparing to other unmixed signals from ICA.

Methods (cont.)

Non-negative Matrix Factorization (NMF)

We also use NMF to decompose a matrix $\mathbf{Y}$ (daily energy usage curves) into two matrices, $\mathbf{W}$ and $\mathbf{B}$. Each row of $\mathbf{Y}$ (daily curves) is approximated by a weighted sum of the rows of $\mathbf{B}$ (signals), with each row of $\mathbf{W}$ containing the appropriate weights. That is,

$$\underset{(N\times 24)}{\mathbf{Y}} \simeq \underset{(N\times 5)}{\mathbf{W}} \cdot \underset{(5\times 24)}{\mathbf{B}}.$$

We find $\boldsymbol{W}^*, \boldsymbol{B}^* = \arg\min_{\boldsymbol{W},\boldsymbol{B}} \|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{B}\|^2$ s.t. $\boldsymbol{W} \geq \mathbf{0}, \quad \boldsymbol{B} \geq \mathbf{0}$. We separate the annual data into summer and winter and perform NMF with 5 bases on each one. An approach similar to the five-step process for ICA can be used.

Other Weighting Approaches

We also used the following approaches:

- Flat Weight: A uniform weight (approximately 0.34) applied to every hour of the sector-aggregated hourly data. This is the "naive" way of estimating hourly residential electricity usage from aggregated usage.
- Empirical Weighting: A flat weight that differs for different parts (morning, evening, etc.) of the day (based on residential pricing data).

Sector-Aggregated Predictions: Autoregressive Model

To form predictions for hourly data aggregated across sectors, our underlying model is an AR model with added regressors. Accounting for the scenario in which only energy data from the previous year is available, we include energy demand from 365 and 366 days earlier (a "shifted" order 2 AR model). Additional covariates, such as irradiance toa, irradiance surface, and temperature (all from the past 2 hours) are chosen based on distance correlation. We also introduce the hour and month as additional covariates into our prediction model. For predictions in the earliest year of the data, the last available year is added to the beginning of the data.

Results and Findings

Recovered Weights

Below are the recovered hourly electricity demand weights (proportions) in the residential sector using various methods:

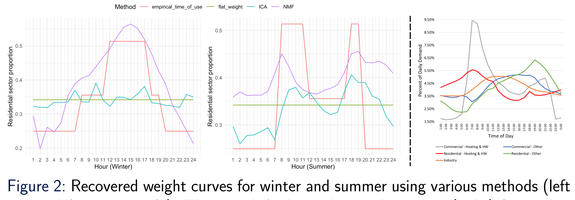

Figure 2: Recovered weight curves for winter and summer using various methods (left and middle, respectively). The true daily demand curve by sectors (right) from other research (see "Licenses") shows the peaks of the residential sector in the morning and evening. The flat weight and empirical weighting curves are defined above. The recovered curves from NMF are quite close to the true demand curve in shape.

Variable Importance

The distance correlation of Székely et al. (2007) is able to assess nonlinear and linear relationships between random variables. We used this metric to identify variables that can serve as strong predictors of electricity demand. The top five variables were found to be: hour (0.50), irradiance toa (0.40), irradiance surface (0.35), snow depth (0.27), and temperature (0.26).

Results and Findings (cont.)

Modelling Results

The suggested leave-one-year-out cross-validation method for computing MAE was used. However, it should be noted that this metric makes use of future observations to make predictions, which can be unrealistic in practice. Also, the predicted hourly residential electricity demand is not included in this version of the MAE (the predicted sum for each year is calculated), so the metric does not capture very well how the *hourly* model performs, as seen in the surprising results of Table 2, in which the flat weight approach (multiplying aggregated predictions by 0.34) works best.

Model	MAE Score (PJ)
Flat Weight + AR	6.735
Empirical Weight + AR	8.192
ICA Weight + AR	6.950
NMF Weight + AR	8.286

Table 2: MAE scores for our selected models.

Remark: Even though the ICA/NMF approach has worse MAE on yearly data, these models provide insight into the daily residential energy use trends. Since we are summing to get predictions, using the annual residential proportion (flat weight) will result in an excellent MAE. A better MAE criterion for measuring correct residential separation would be to calculate residuals at an hourly resolution. Also, the ICA/NMF-based weights were all calculated from 2011 IESO zonal data, even for the 2011 CV year. Flat weights were also not CV'd.

Discussion / Conclusions

Using our ICA/NMF/AR models, we obtained accurate estimates of the residential sector annual data. Due to the residential sector hourly data being unavailable, we are not able to verify any model's accuracy at each *hour*. We now provide some strengths and limitations.

Strengths:

- A signal separation approach is used to obtain sector-specific hourly information with very little provided residential data.
- Since an AR model is used (as opposed to an RNN), we have more capacity to interpret our results.

Limitations:

- ICA/NMF might be selecting imperfect signals/components (components not independent), affecting residential electricity demand predictions.
- Model could be further improved by taking into account differences in *sector-relative* residential demand across years

Future Work

ICA and NMF are unsupervised methods. An alternative approach is to provide additional constraints on the output signals via a Bayesian or supervised ICA or NMF method. Additionally, although we have augmented our data set, a better data associated with appliances can likely be obtained. Obtaining hourly electricity usage for even a few residents (with their consent) could greatly improve accuracy.

Acknowledgements / References

We would like to thank Dr. Lloyd Elliott for his help in this project as a faculty mentor. Additional data that was used in the project was obtained from the IESO and Statistics Canada websites. Further information is available in the attached "Licenses" document.