Electricity Demand Forecasting using Multi-Task Learning

Jean-Baptiste Fiot, Francesco Dinuzzo

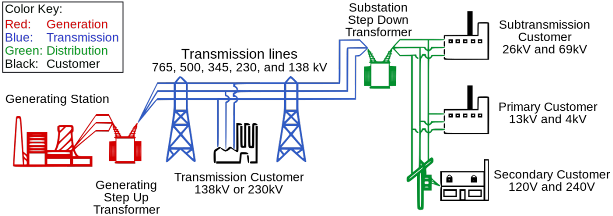
Dublin Machine Learning Meetup - July 2017

Outline

Outline

Electricity Demand Forecasting

- Electricity is a special commodity
  - It **cannot be stored** efficiently (in large quantities)
  - It **looses value** when being moved (line losses)
- Demand forecasting is critical
  - Operations, bidding, demand response, maintenance, planning, etc.
- The game is changing
  - **Distributed** renewable generation
  - Higher **volatility** on markets
  - Increased number of participants



Demand Forecasting Methods

- (Non-)linear variants of least-squares, ARMAX, fuzzy logic, etc.
- Black-box models based on neural networks [Hippert et al., 2001]
- Generalized **Additive** Models (GAM)
  - Great performance [Fan and Hyndman, 2012, Ba et al., 2012]
  - Efficient and scalable training algorithms
  - Interpretability of the model

Hippert, HS, et al.
Neural networks for short-term load forecasting: A review and evaluation.
Power Systems, IEEE Transactions on, 16(1):44–55, 2001.

Fan, S and Hyndman, R.
Short-term load forecasting based on a semi-parametric additive model.
Power Systems, IEEE Transactions on, 27(1):134–141, 2012.

Ba, A, et al.
Adaptive learning of smoothing functions: application to electricity load forecasting.
In **Advances in Neural Information Processing Systems 25 (NIPS 2012)**, pages 2519–2527. 2012.

Demand Forecasting using Kernel Methods

- In 2001, **kernel-based support vector regression** won EUNITE (European Network on Intelligent Technologies for Smart Adaptive Systems) demand forecasting competition [Chen et al., 2004]
- Later, **kernel-based regularizations** and **support vector techniques** were successfully used [Espinoza et al., 2007, Hong, 2009, Elattar et al., 2010]

Chen, B, et al.
Load forecasting using support vector machines: A study on EUNITE competition 2001.
Power Systems, IEEE Transactions on, 19(4):1821–1830, 2004.

Espinoza, M, et al.
Electric load forecasting.
Control Systems, IEEE, 27(5):43–57, 2007.

Hong, WC.
Electric load forecasting by support vector model.
Applied Mathematical Modelling, 33(5):2444–2454, 2009.

Elattar, E, et al.
Electric load forecasting based on locally weighted support vector regression.
Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(4):438–447, 2010.

Outline

Electric Demand Forecasting

$$\hat{y} = f\left(t, d, c, \quad y_l, u_l, \quad j, s_j\right),$$

- Time/Calendar features
  - $t \in [0, 24)$ is the **time of day** expressed in hours,
  - $d \in \{1, 2, \ldots, 365, 366\}$ is the **day of the year**,
  - c is the **type of day**, e.g. Monday to Sunday,
- Dynamic features
  - y_l is a real vector containing lagged values of the electric demand,
  - u_l is a real vector containing measurements of lagged values of exogenous variables other than the demand (such as temperature),
- Meter features
  - j is the meter ID in the electricity network,
  - s_j is a vector of features describing the demande measured at j.

Electric Demand Forecasting

$$\hat{y} = f\left(t, d, c, \quad y_l, u_l, \quad j, s_j\right),$$

- Time/Calendar features
  - $t \in [0, 24)$ is the time of day expressed in hours,
  - $d \in \{1, 2, \ldots, 365, 366\}$ is the day of the year,
  - c is the type of day, e.g. Monday to Sunday,
- Dynamic features
  - y_l is a real vector containing **lagged** values of the electric **demand**,
  - u_l is a real vector containing measurements of **lagged** values of **exogenous variables** other than the demand (such as temperature),
- Meter features
  - j is the meter ID in the electricity network,
  - s_j is a vector of features describing the demande measured at j.

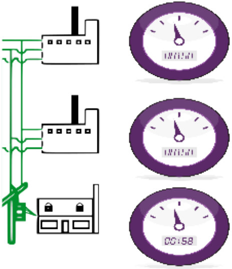
Electric Demand Forecasting

$$\hat{y} = f\left(t, d, c, \quad y_l, u_l, \quad j, s_j\right),$$

- Time/Calendar features
  - $t \in [0, 24)$ is the time of day expressed in hours,
  - $d \in \{1, 2, \ldots, 365, 366\}$ is the day of the year,
  - c is the type of day, e.g. Monday to Sunday,
- Dynamic features
  - y_l is a real vector containing lagged values of the electric demand,
  - u_l is a real vector containing measurements of lagged values of exogenous variables other than the demand (such as temperature),
- Meter features
  - j is the **meter ID** in the electricity network,
  - s_j is a vector of features describing the **demande measured at j**.

Solving Multiple Demand Forecasting Problems

- Consider m smart meters, indexed by j



- Goal: learn $\{f_j : \mathcal{X} \to \mathbb{R}\}_{1 \leq j \leq m}$ from datasets $(x_{ij}, y_{ij}) \in \mathcal{X} \times \mathbb{R}$.

Optimisation Problem

- Letting $f : \mathcal{X} \to \mathbb{R}^m$ the function with components f_j, we minimize

$$R(f, \mathbf{L}) = \sum_{j=1}^{m} \sum_{i=1}^{\ell_j} (y_{ij} - f_j(x_{ij})))^2 + \lambda \|f\|^2_{\mathcal{H}_{\mathbf{L}}}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter, and $\mathcal{H}_{\mathbf{L}}$ is a Reproducing Kernel Hilbert Space (RKHS) of vector-valued functions with (matrix-valued) kernel

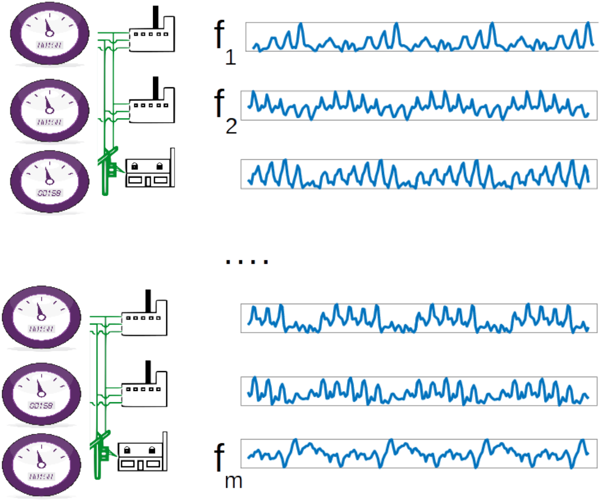
$$H(x_i, x_j) = K(x_i, x_j) \cdot \mathbf{L}, \tag{2}$$

$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the **input kernel**, and $\mathbf{L} \in \mathbb{R}^{m \times m}$ is the **output kernel**.

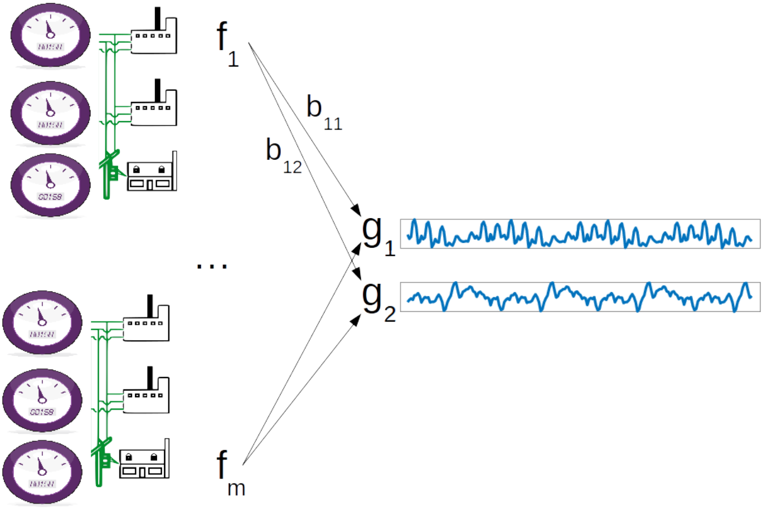
- **Representer theorem**: there exist functions $\hat{f}_j$ minimizing $R(f, \mathbf{L})$ in the form:

$$\hat{f}_j(x) = \sum_{k=1}^{m} \mathbf{L}_{jk} \sum_{i=1}^{\ell_k} c_{ik} K(x_{ik}, x). \tag{3}$$

Fixing L = I: Independent Kernel Ridge Regression



Learning $\mathbf{L} = \mathbf{I}$: Output Kernel Learning



Remark: $\mathbf{B} = (b_{ij})$ is a Cholesky factor of $\mathbf{L}$

Output Kernel Learning

- Joint optimization problem

$$\min_{\mathbf{L}\in\mathbb{S}_+^{m,p}} \quad \min_{f\in\mathcal{H}_\mathbf{L}} \quad R(f,\mathbf{L}) + \lambda \mathrm{tr}(\mathbf{L})\,,$$

  where $\mathbb{S}_+^{m,p}$ is the cone of p.s.d. matrices with rank $\leq p$.

- Re-indexing the observations $\{x_i\}_{i=1,\ldots,\ell}$, the solution becomes

$$\hat{f}_j(x) = \sum_{k=1}^{p} b_{jk} g_k(x), \quad g_k(x) = \sum_{i=1}^{\ell} a_{ik} K(x_i, x)\,,$$

  where $\begin{cases} b_{jk} \text{coefficients form a low-rank factor of } \mathbf{L}\,, \\ g_k \text{functions can be seen as } \textbf{modes} \text{ or } \textbf{typical profiles}\,. \end{cases}$

- It is sufficient to store $(\ell + m)p$ parameters, which can be much smaller than $\sum_{j=1}^{m} \ell_j$.

Outline

Multiple Seasonalities in Electricity Demand

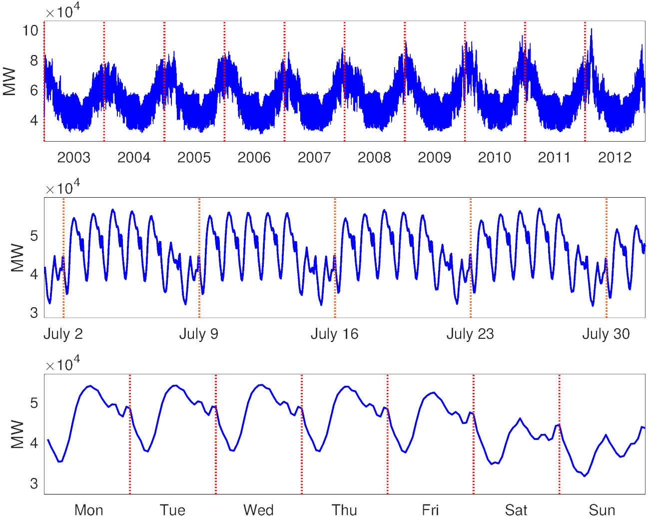


Figure: French National Demand (Réseau de Transport d'Électricité data)

Capturing Demand Seasonalities with Kernels

- **Time-of-day** kernel

$$K^t(t_1, t_2) = \exp\left(-h_T(|t_1 - t_2|)/\sigma_t\right) , \tag{4}$$

- **Day-of-year** kernel

$$K^d(d_1, d_2) = \exp\left(-h_D(|d_1 - d_2|)/\sigma_d\right) , \tag{5}$$

where $h_P(x) = \min\{x, P - x\}$ is a change of variable that yields P-periodic kernels over the square $[0, P]^2$. In our experiment, σ_t and σ_d were respectively set to 4 hours and 120 days.

- **Day-type** kernel

$$K^c(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 = c_2 \\ 0 & \text{if } c_1 \neq c_2. \end{cases} . \tag{6}$$

Kernels for Electric Demand Forecasting

To define $K((t_1, d_1, c_1), (t_2, d_2, c_2))$, we combine the basis kernels

- **Additive** Models

$$K^t(t_1, t_2) + K^d(d_1, d_2)\,, \tag{7}$$

$$K^t(t_1, t_2) + K^d(d_1, d_2) + K^c(c_1, c_2)\,, \tag{8}$$

- **Semi-Additive** Models

$$K^d(d_1, d_2) + K^t(t_1, t_2) \cdot K^c(c_1, c_2)\,, \tag{9}$$

$$\left(K^t(t_1, t_2) + K^d(d_1, d_2)\right) \cdot K^c(c_1, c_2)\,, \tag{10}$$

- **Multiplicative** Models

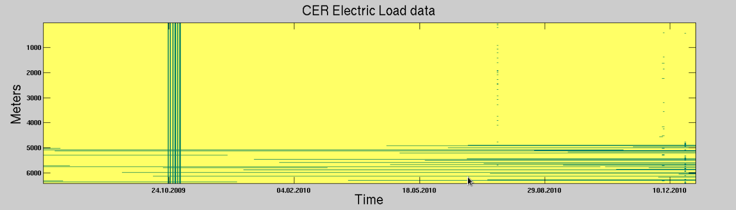
$$K^t(t_1, t_2) \cdot K^d(d_1, d_2)\,, \tag{11}$$

$$K^t(t_1, t_2) \cdot K^d(d_1, d_2) \cdot K^c(c_1, c_2)\,. \tag{12}$$

Outline

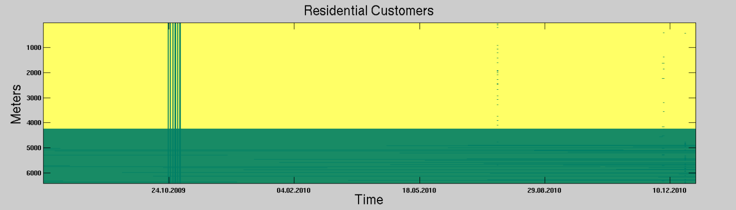
Commission for Energy Regulation (CER) Data

- 6435 smart meters
- 536 days (Jul 14, 2009 - Dec 31, 2010)
- Half-hour sampling
- 3 groups: residential, SME, others



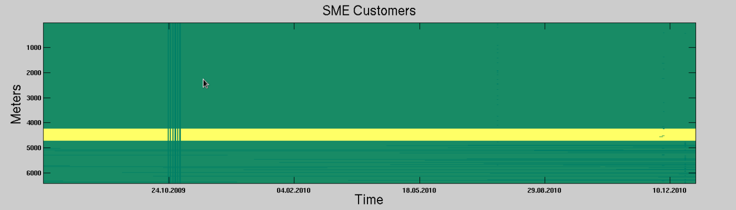
Commission for Energy Regulation (CER) Data

- 6435 smart meters
- 536 days (Jul 14, 2009 - Dec 31, 2010)
- Half-hour sampling
- 3 groups: **residential**, SME, others



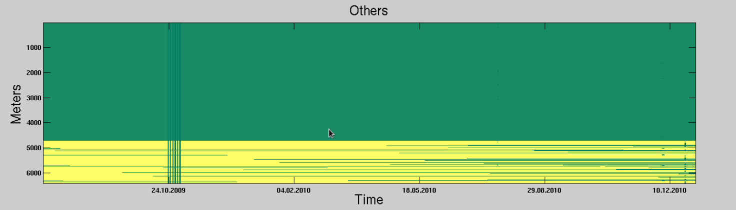
Commission for Energy Regulation (CER) Data

- 6435 smart meters
- 536 days (Jul 14, 2009 - Dec 31, 2010)
- Half-hour sampling
- 3 groups: residential, **SME**, others



Commission for Energy Regulation (CER) Data

- 6435 smart meters
- 536 days (Jul 14, 2009 - Dec 31, 2010)
- Half-hour sampling
- 3 groups: residential, SME, **others**



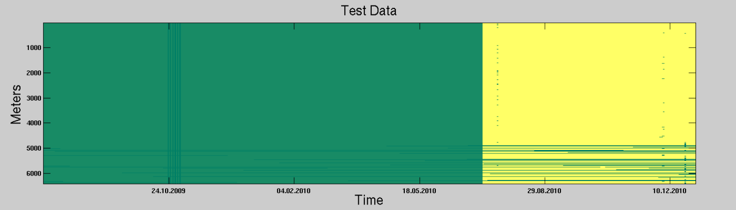
Pre-processing

- Removed two corrupted meters
- Corrected DST measurements
- Downsampled to 3-hour resolution
- Final dataset:
  - $m = 6433$ smart meters
  - $\ell = 4288$ time slots

Customer group	**Meters**	**Sparsity**
Residential	4225	0.028%
Industrial (SME)	485	0.035%
Others	1723	17%

Learning the Models

- Data split
  - 1 year (2920 obs.) used for training (80%) and validation (20%)
  - $\sim$ 0.5 year (1368 obs.) used for testing



- Independent Kernel Ridge Regression using the 6 kernels

- Output Kernel Learning using MM2
  - 1 model for $\{\text{residential}\} \cup \{\text{others}\}$, $p = 200$ to fit in memory
  - 1 model for $\{\text{SME}\}$, full rank ($p = 485$)

Qualitative Analysis

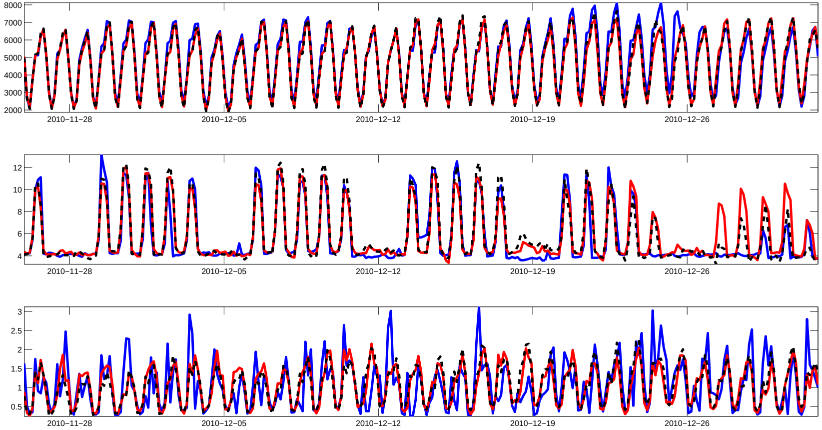


Figure: Measured load (blue), indep. KRR (red) and multi-task OKL (black) forecasts for the aggregated demand (top), a single SME meter (middle), and a single residential meter (bottom).

Performance Metrics (1/2)

Given a group of meters $\mathcal{G}$ and observation i, we define

- Absolute percentage error (APE)

$$\text{APE}(i, \mathcal{G}) = 100 \left| \frac{\sum_{j \in \mathcal{G}_i} y_{ij} - \sum_{j \in \mathcal{G}_i} f_j(t_i, d_i, c_i)}{\sum_{j \in \mathcal{G}_i} y_{ij}} \right| , \tag{13}$$

  where $\mathcal{G}_i$ is the subset of meters with available observations at i.

- Normalized absolute error (NAE)

$$\text{NAE}(i, \mathcal{G}) = \frac{\sum_{j \in \mathcal{G}_i} |y_{ij} - f_j(t_i, d_i, c_i)|}{\sum_{j \in \mathcal{G}_i} y_{ij}} , \tag{14}$$

Performance Metrics (2/2)

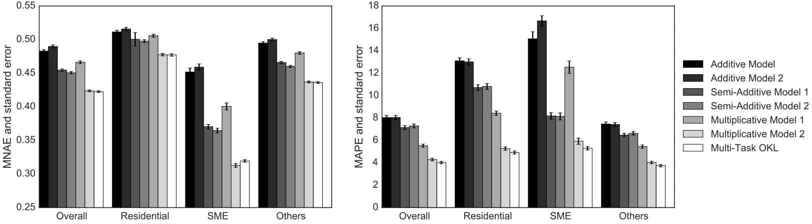
- Mean absolute percentage error (MAPE)

$$\mathrm{MAPE}(\mathcal{G}) = \frac{1}{\#\ T} \sum_{i \in T} \mathrm{APE}(i, \mathcal{G})\,, \tag{15}$$

- Mean normalized absolute error (MNAE)

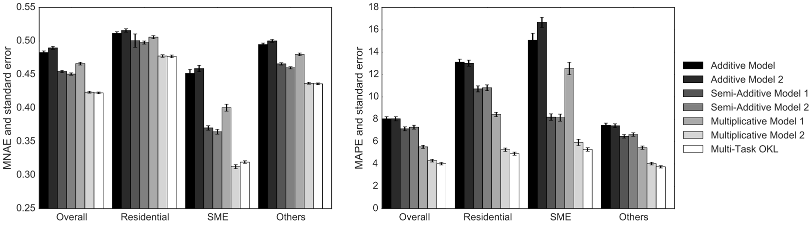
$$\mathrm{MNAE}(\mathcal{G}) = \frac{1}{\#\ T} \sum_{i \in T} \mathrm{NAE}(i, \mathcal{G})\,. \tag{16}$$

Prediction Accuracy (1/2)



1. Multiplicative kernels outperform (semi-)additive models.
   - Multiplicative kernels lead to a stricter selection of training obs.
   - EUNITE winners discarded $\geq 90\%$ of the dataset.

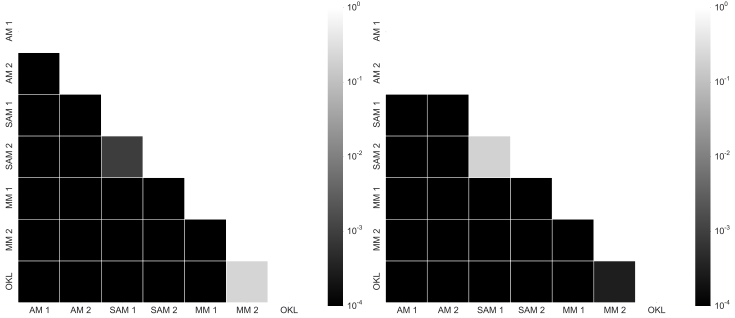
Prediction Accuracy (1/2)



1. Multiplicative kernels outperform (semi-)additive models.
   - Multiplicative kernels lead to a stricter selection of training obs.
   - EUNITE winners discarded $\geq 90\%$ of the dataset.

2. Multi-task OKL outperforms independent kernel ridge regression
   - The multi-task approach efficiently exploits the similarities
   - 44% improvement of σ_{APE} for SME against 2nd best method

Prediction Accuracy (2/2)



(a) NAE (b) APE

Figure: p-values of Welch t-test between the overall accuracies of all methods on the CER dataset

Basis Load Profiles g_k

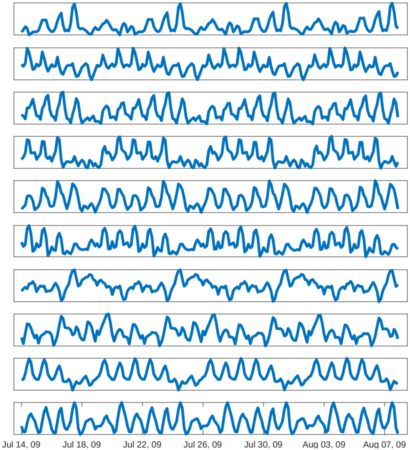


Figure: CER Data: Typical load profiles displayed over the horizon of one month, obtained from a low-rank OKL model with $p = 10$.

Number of Parameters

In this experiment, the OKL model is **4.24 times more compact**.

- Single-task: $\#$ params $=$ $\#$ obs. $= \sum_{j=1}^{m} \ell_j \approx 1.3 \cdot 10^7$
- Multi-task OKL: $\#$ params $= (\ell + m)p \approx 3 \cdot 10^6$

Relationships between Smart Meters

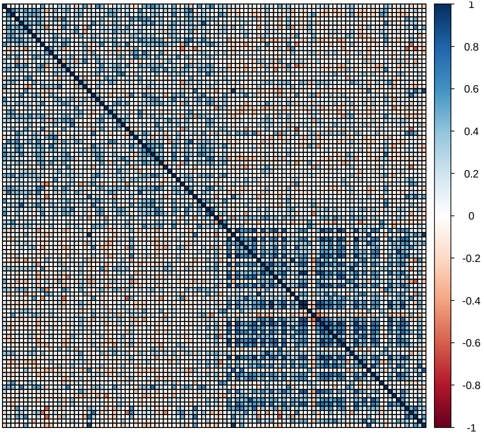


Figure: CER data: entries of the normalized output kernel $\mathbf{L}_n \in \mathbb{R}^{m \times m}$ for a subset containing 50 residential and 50 SME (small or medium enterprise) customers. $(\mathbf{L}_n)_{ij} = \frac{\mathbf{L}_{ij}}{\sqrt{\mathbf{L}_{ii} \times \mathbf{L}_{jj}}}, \qquad i, j = 1, \ldots, m.$

Outline

Contributions

1. We formulated the problem of forecasting the demand measured on multiple lines of the network as a **multi-task problem**.

2. We designed **kernels** able to capture the **seasonal effects** present in electricity demand data.

3. We exposed the performance limits of the very popular additive models, showing that they are often outperformed by **multiplicative kernel models**.

4. We showed how MTL can be used to gain **insights** and **interpretability** on real demand data

Thank You

- Any question?
- Contact details
  - Jean-Baptiste Fiot
  - jean-baptiste.fiot@centraliens.net