Industrial Applications of Kernel Spectral Clustering: Improving Power Load Forecasting and Detecting Sentence Similarity in Argument Construction

Carlos Alzate, Mathieu Sinn, Ruty Rinott

Smarter Cities Technology Centre



IBM Research - Ireland

Mini-Symposium:

"Complex Networks: Methods and Applications"

IBM

Outline

- Part I: Improving Power Load Forecasting
- Part II: Detecting Sentence Similarity in Argument Construction
- Part III: Demonstration

© 2012 IBM Corporation

Part I

© 2012 IBM Corporation

The Problem

- Improve forecasting accuracy of aggregated energy consumption signals
- Short-term analysis (prediction for next days)
- Essential for control and planning in electrical grids and electricity markets

The Idea

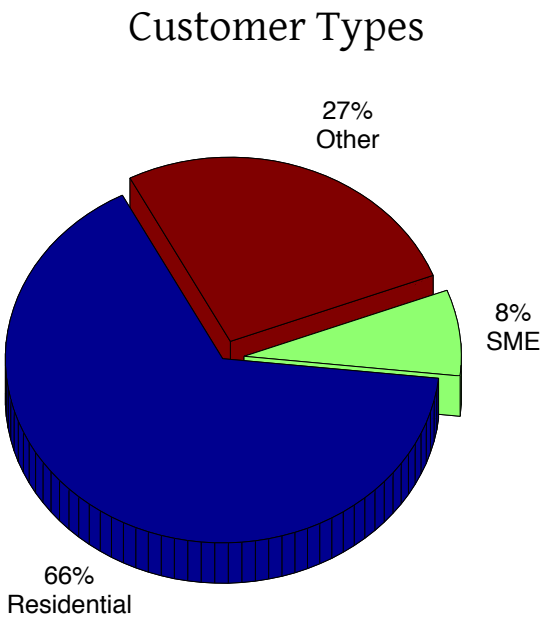
- Use smart meters and clustering to improve the global forecast
- Methodology: Kernel spectral clustering
- Forecasting model: PARX

© 2012 IBM Corporation

The Data - Preliminaries

- Electricity consumption measured at $6,435$ customers
- Individual time series (not aggregated)
- Values are kW consumed during 30 minute interval
- Total number of days is 536 (Jul. 14, 2009 - Dec. 31, 2010)





SmarterCities Technology Centre

© 2012 IBM Corporation

The Data - Preliminaries

Initial Preprocessing

- Time series < 100 missing values $\rightarrow$ linear interpolation
- Pruned dataset:
  - Residential $\rightarrow 4255$ time series
  - SME $\rightarrow 485$ time series
  - Other $\rightarrow 952$ time series

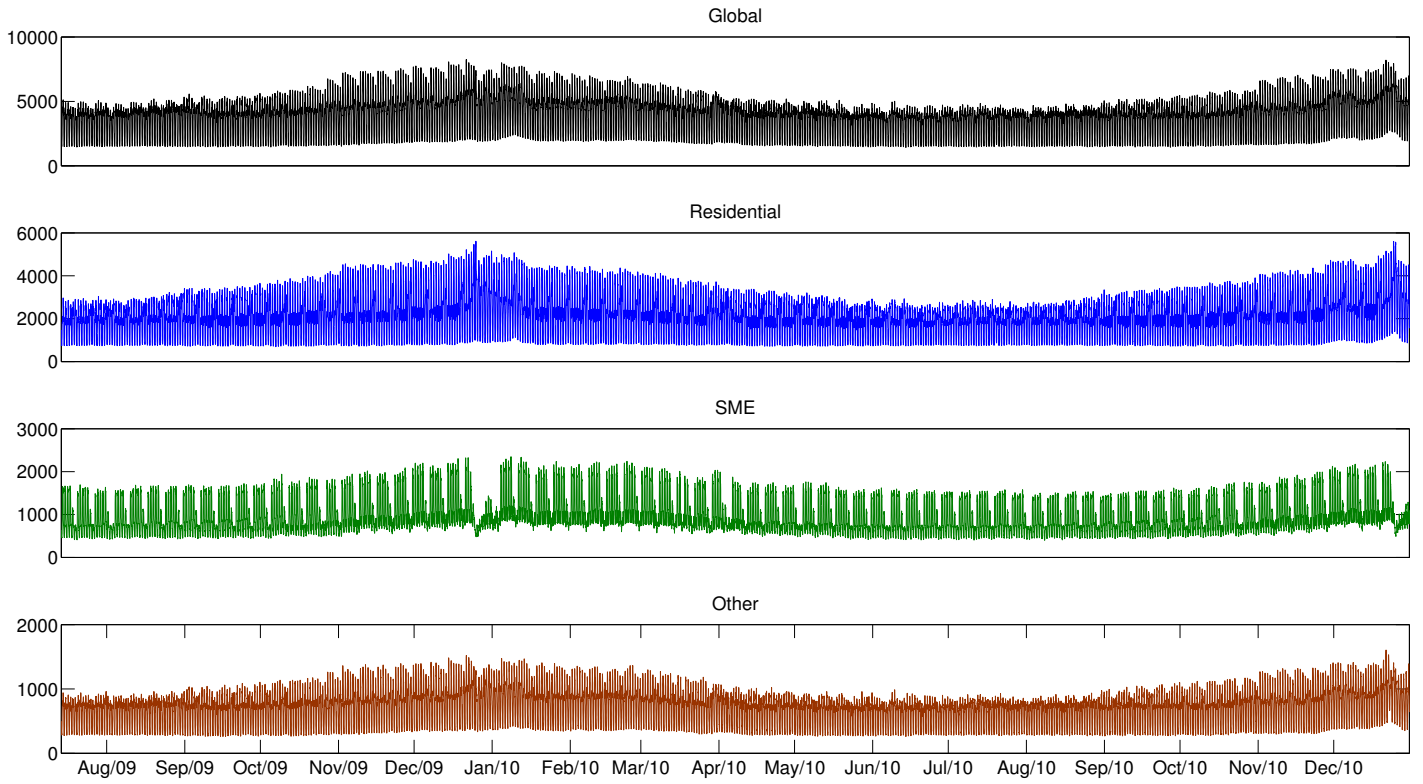
Outlier Detection

- Wavelet details capture high frequency information
- Outliers $\rightarrow$ heavy tails $\rightarrow$ high kurtosis

© 2012 IBM Corporation

The Data - Outlier Removal

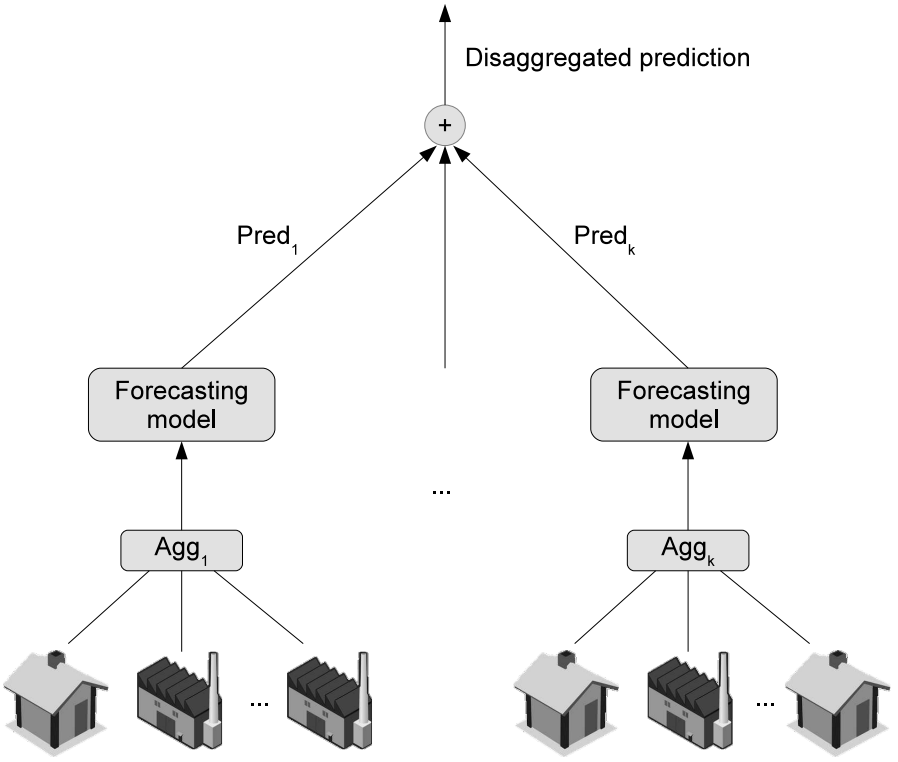
- Final dataset: $5,335$ time series
  - $3,994$ residential
  - 444 SME
  - 897 other



SmarterCities Technology Centre

© 2012 IBM Corporation

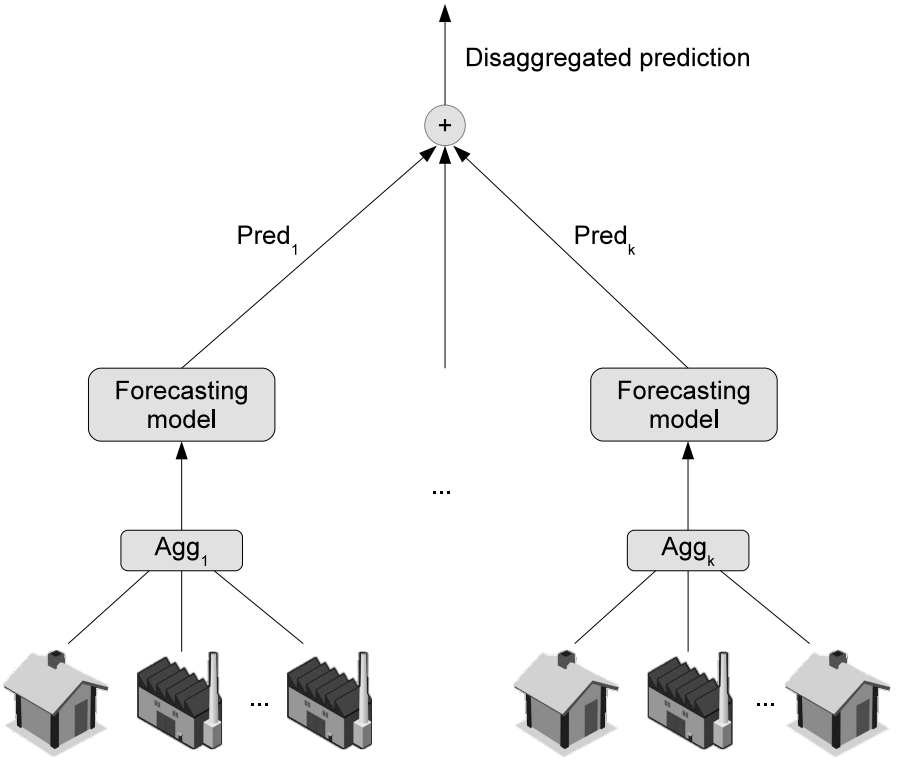
The Methodology



SmarterCities Technology Centre

© 2012 IBM Corporation

The Methodology



- **Existing work:** [Misiti et al., 2010]
  - Industrial data only
  - Hierarchical clustering
  - SARIMA for forecasting
  - **Initial** clusters via wavelets
  - **Iteratively** merge via cross prediction error

- Computationally expensive
- Ad-hoc param. determination
- Arbitrary # clusters
- Cross-prediction error not a proper distance

SmarterCities Technology Centre

© 2012 IBM Corporation

The Methodology - Clustering Method

Kernel Spectral Clustering (KSC)[Alzate and Suykens, 2010]

- Nonlinear clustering technique with predictive capabilites
- Bottleneck is an eigendecomposition problem
- Deliver better results than classical techniques (e.g., k-means)
- Modularity: Just plug in your favorite metric (valid kernel)
- Extendability (sparseness, prior information, partial labels, ...)

© 2012 IBM Corporation

The Methodology - Clustering Method

Kernel Spectral Clustering (KSC)[Alzate and Suykens, 2010]

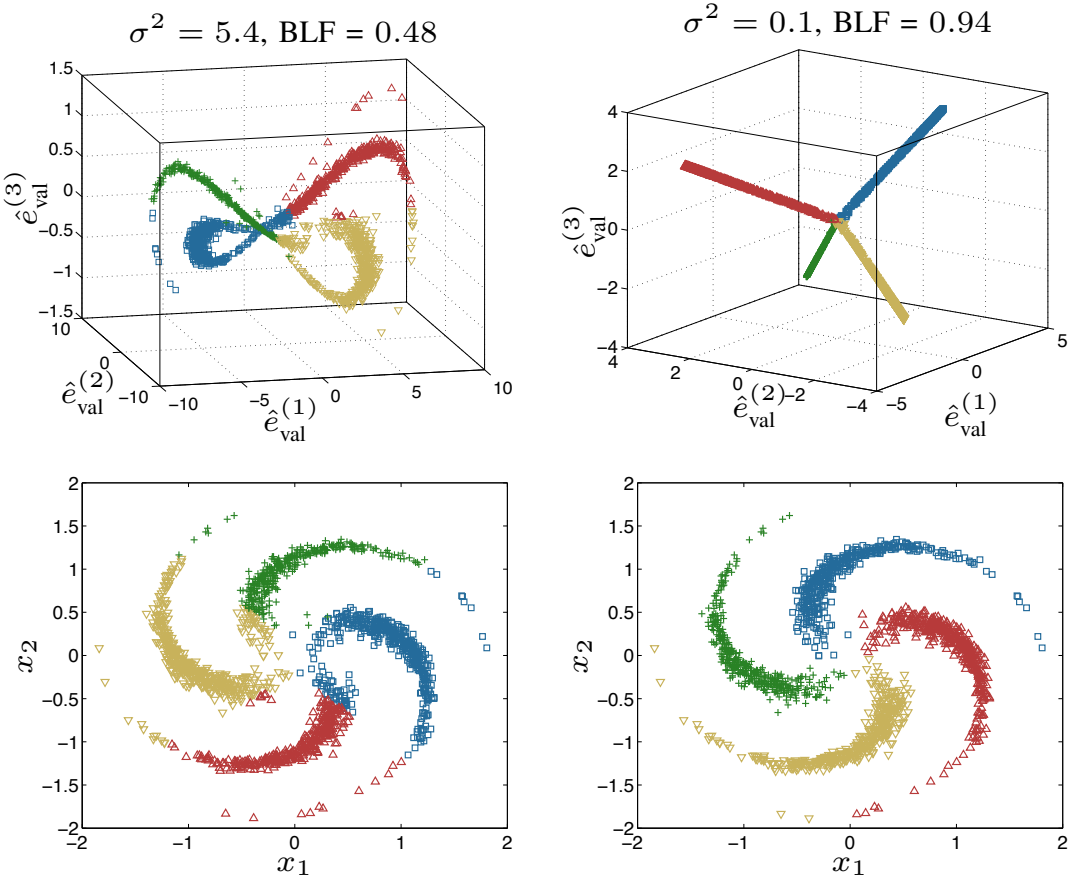
- Nonlinear clustering technique with predictive capabilites
- Bottleneck is an eigendecomposition problem
- Deliver better results than classical techniques (e.g., k-means)
- Modularity: Just plug in your favorite metric (valid kernel)
- Extendability (sparseness, prior information, partial labels, ...)

Technicalities

- Eigendecomposition of a modified Laplacian matrix
- Weighted version of PCA in high-dimensional feature spaces
- Out-of-sample extensions (uncommon in (spectral) clustering)
- Model selection (e.g., number of clusters, kernel params)

© 2012 IBM Corporation

KSC - Model Selection - Balanced Line Fit (BLF)



SmarterCities Technology Centre

© 2012 IBM Corporation

The Methodology - Selecting An Appropriate Kernel

Four kernels of type $K(x, z) = \exp(-\phi(x, z)^2/\sigma^2)$ are explored:

- **Wavelet feature extraction:** x, z are transformed to wavelet space. $\phi(x, z)$ is the Euclidean distance between wavelet features. Non-parametric, used in
  [Antoniadis et al., 2010, Misiti et al., 2007, Misiti et al., 2010]
- **VAR kernel:** x, z are multivariate time series. $\phi(x, z)$ is a distance between distributions of the VAR parameters. Semi-parametric, used in **[Cuturi, 2011a]**
- **Global Alignment kernel**: x, z are multivariate time series. $\phi(x, z)$ is a generalization of the DTW distance. Non-parametric, used in **[Cuturi, 2011b]**
- **RBF with Spearman's distance**: $\phi(x, z)$ is $\sqrt{1 - \rho^2_{\text{ranked}}}$. Measures correlation with respect to a monotonic function. Non-linear, non-parametric

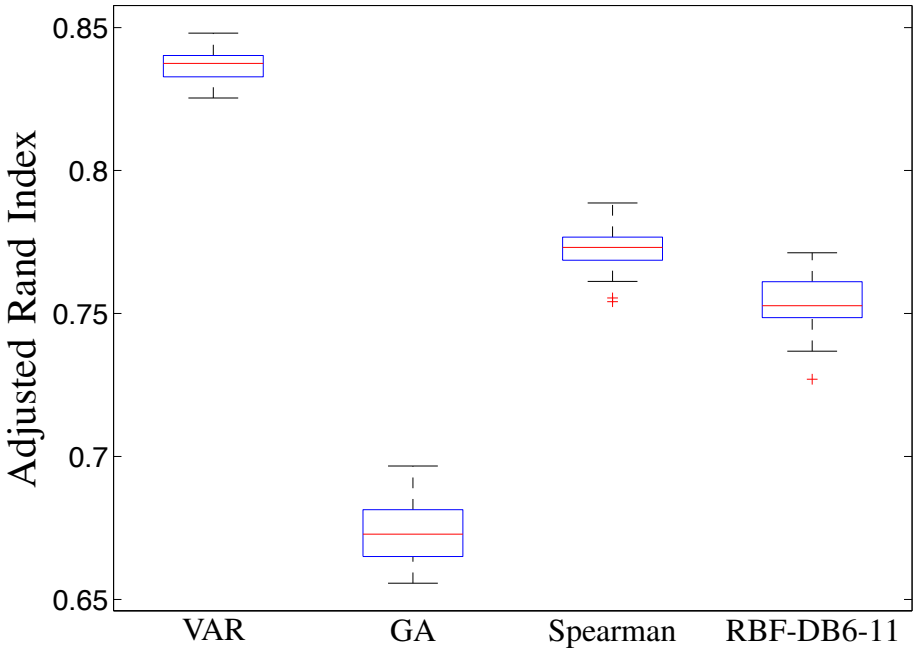
SmarterCities Technology Centre

© 2012 IBM Corporation

The Results - Residential vs. SME

Setup

- Assess discriminatory power of the kernels
- Residential vs. SME ("other" left out)
- 600 training samples (300 res, 300 SME), 50 randomizations
- VAR kernel outperforms (Spearman's runner-up)

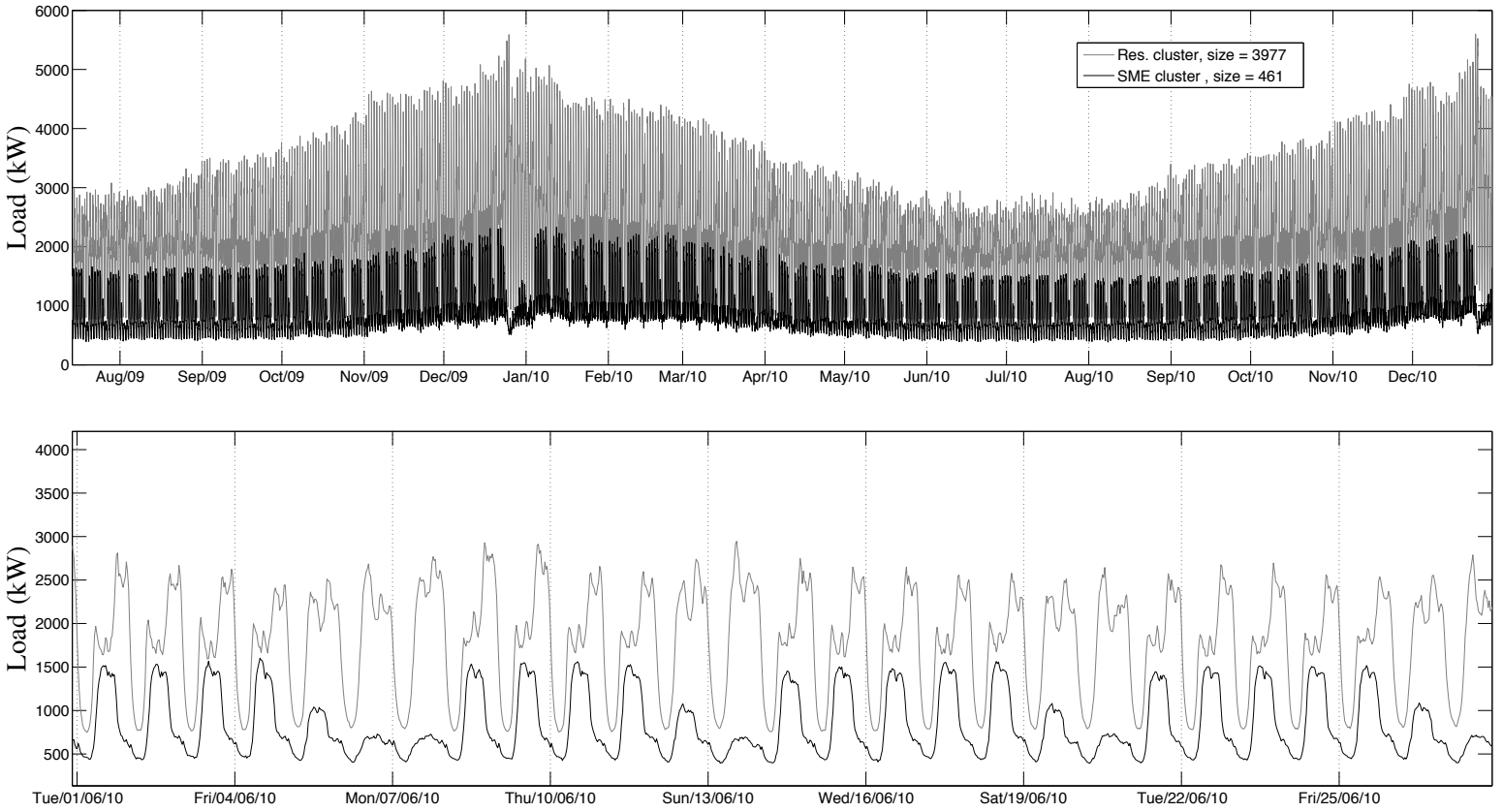


SmarterCities Technology Centre

© 2012 IBM Corporation

The Results - Residential vs. SME

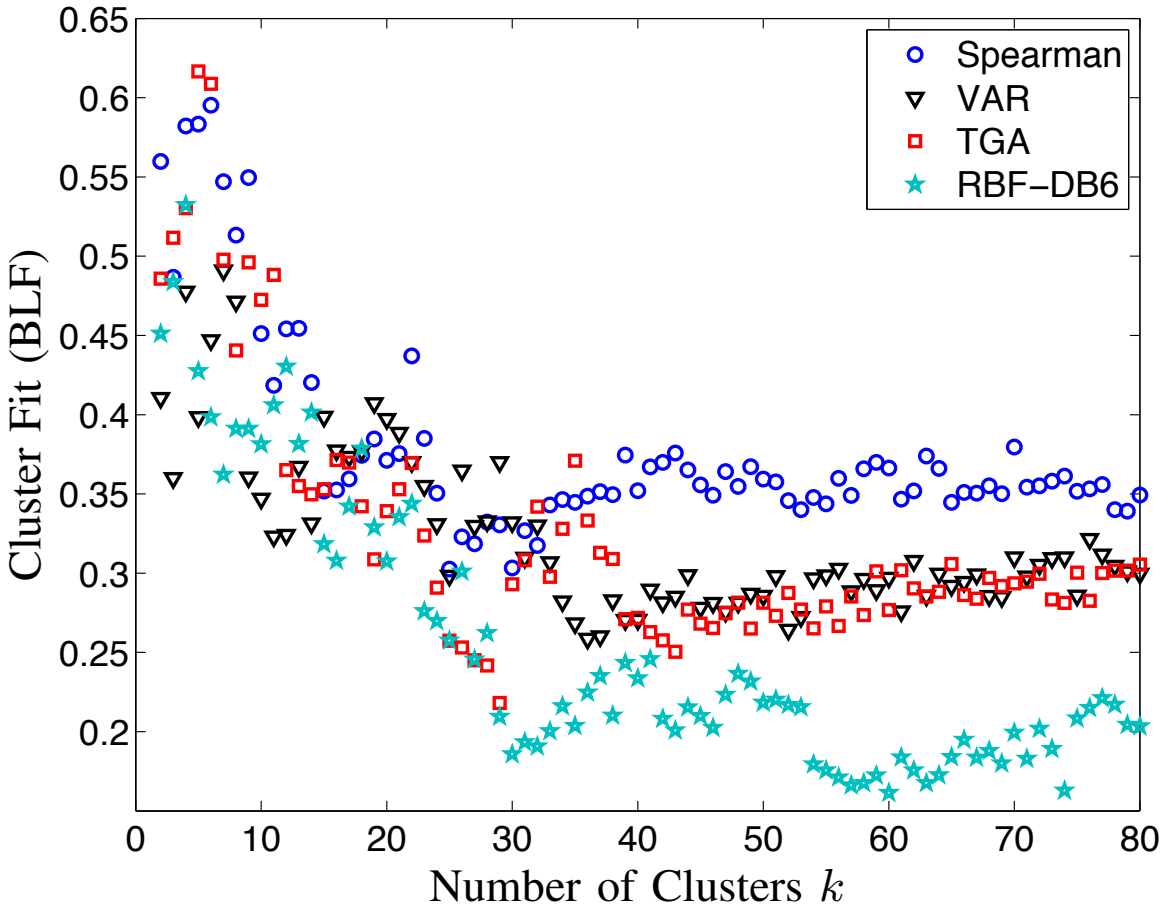
Cluster purity: Residential $\rightarrow 98.55\%$, SME $\rightarrow 91\%$



SmarterCities Technology Centre

© 2012 IBM Corporation

Determining the number of clusters k:



© 2012 IBM Corporation

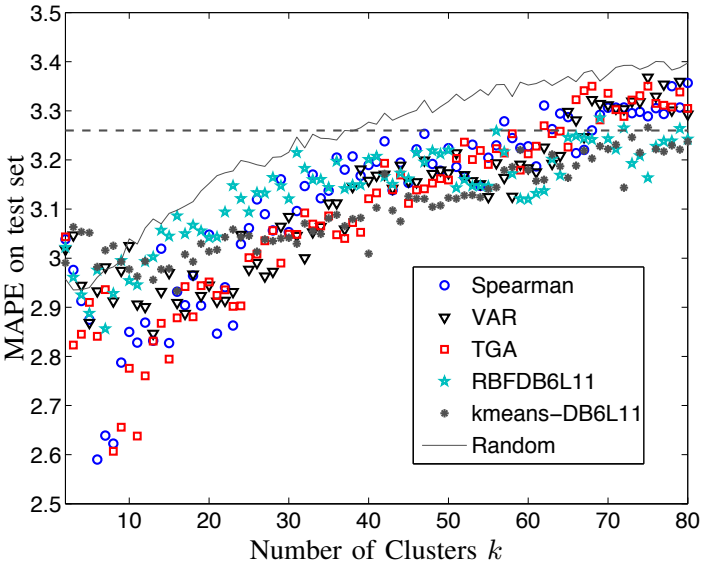
The Results - Clustering For Forecasting

- Forecasting model: PARX
  - Simple and fast
  - Calendar and temperature
  - Solved by OLS
  - Used in [Espinoza et al., 2005]
  - 400 days $\rightarrow$ training
  - 136 days $\rightarrow$ test

© 2012 IBM Corporation

The Results - Clustering For Forecasting

- Forecasting model: PARX
  - Simple and fast
  - Calendar and temperature
  - Solved by OLS
  - Used in [Espinoza et al., 2005]
  - 400 days $\rightarrow$ training
  - 136 days $\rightarrow$ test

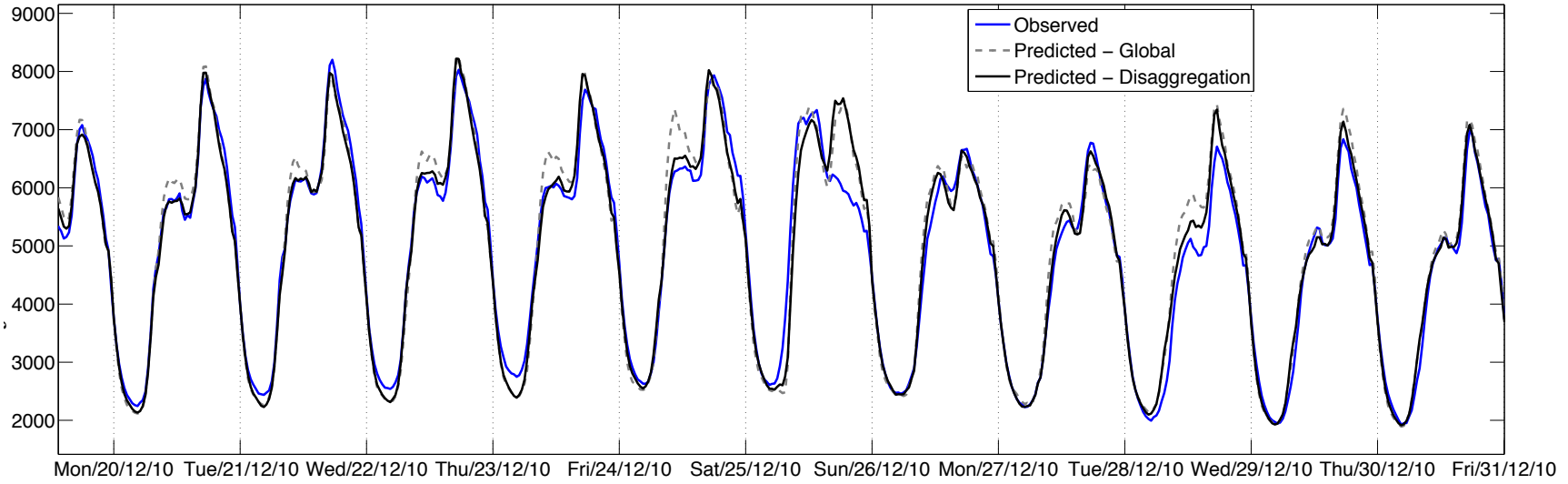


Kernel	Model Selection		Disaggregated Forecast Baseline MAPE = 3.26%		
	$k^\star$	BLF($k^\star$)	$k^\star$	MAPE($k^\star$)	Gain
VAR	7	0.49	13	2.85%	12.68%
TGA	5	0.63	8	2.61%	20.04%
Spearman	**6**	**0.55**	**6**	**2.59%**	**20.55%**
RBF-DB6-11	4	0.53	5	3.02%	7.36%
$k$means-wav			16	2.93%	
Random			3	$2.93\% \pm 0.03$	

SmarterCities Technology Centre

© 2012 IBM Corporation

The Results - Forecasting Details



SmarterCities Technology Centre

© 2012 IBM Corporation

Conclusions

- Disaggregated forecasting via clustering improves by 20.55%
- Number of clusters is rather small $k = 6$ for maximal gain
- Forecasting deteriorates rapidly with an increasing k
- Simple metrics seem to outperform fancy ones
- Spectral clustering shown to perform well in this taskdd

SmarterCities Technology Centre

© 2012 IBM Corporation



End of Part I

© 2012 IBM Corporation


References I


Alzate, C. and Suykens, J. A. K. (2010).
Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(2):335–347.


Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J. (2010).
Clustering functional data using wavelets.
In *COMPSTATS*, pages 697–704.


Cuturi, M. (2011a).
Autoregressive kernels for time series.
Technical report, arXiv:1101.0673.


Cuturi, M. (2011b).
Fast global alignment kernels.
In *Proceedings of the International Conference on Machine Learning (ICML)*.


Espinoza, M., Joye, C., Belmans, R., and De Moor, B. (2005).
Short-term load forecasting, profile identification and customer segmentation: A methodology based on periodic time series.
IEEE Transactions on Power System, 20(3):1622–1630.

SmarterCities Technology Centre

© 2012 IBM Corporation

References II



Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2007).
Clustering signals using wavelets.
In *Computational and Ambient Intelligence*, volume 4507, pages 514–521.



Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2010).
Optimized Clusters for Disaggregated Electricity Load Forecasting.
REVSTAT - Statistical Journal, 8(2):105–124.

© 2012 IBM Corporation