Forecasting and Clustering

Dr Richard Mitchell

Cybernetics Intelligence Research Group

Department of Cybernetics

The University of Reading, UK

– R.J.Mitchell@reading.ac.uk

Work done on Sabbatical at National Grid

The University of Reading

Cybernetics

National Grid

Overview

- Main work on sabbatical has concerned
  - clustering data
  - forecasting, using clusters
- Presentation will describe
  - new hybrid cluster algorithm
  - how to forecast using clusters
  - systematic method to select factors
  - success on Electricity Demand

Clustering

- Data are *n* points in *d* dimensions
  - e.g. Electricity Demand depends on Temperature, Time, Day of Week, etc.
- Clustering is process of grouping together similar data points
  - Similar means ‘close together’
- Numerous algorithms exist
  - k-Means is most popular
  - But there are problems

k-Means Algorithm

- Assume *K* and *K* initial cluster centres

  REPEAT

    Allocate each point to nearest centre

    Centres := Mean(points in cluster)

  UNTIL Centres don't move

  QError := Mean(distances from centre)

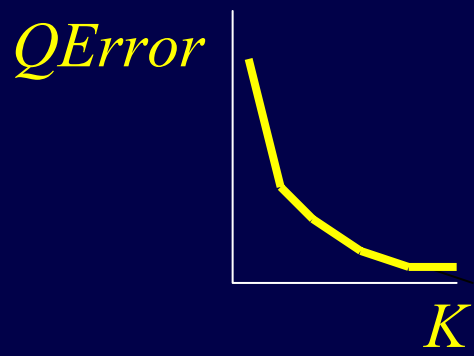
- BUT What is *K*?

  It is sensitive to initial positions

  Uses (slow) distance calculations

Solutions

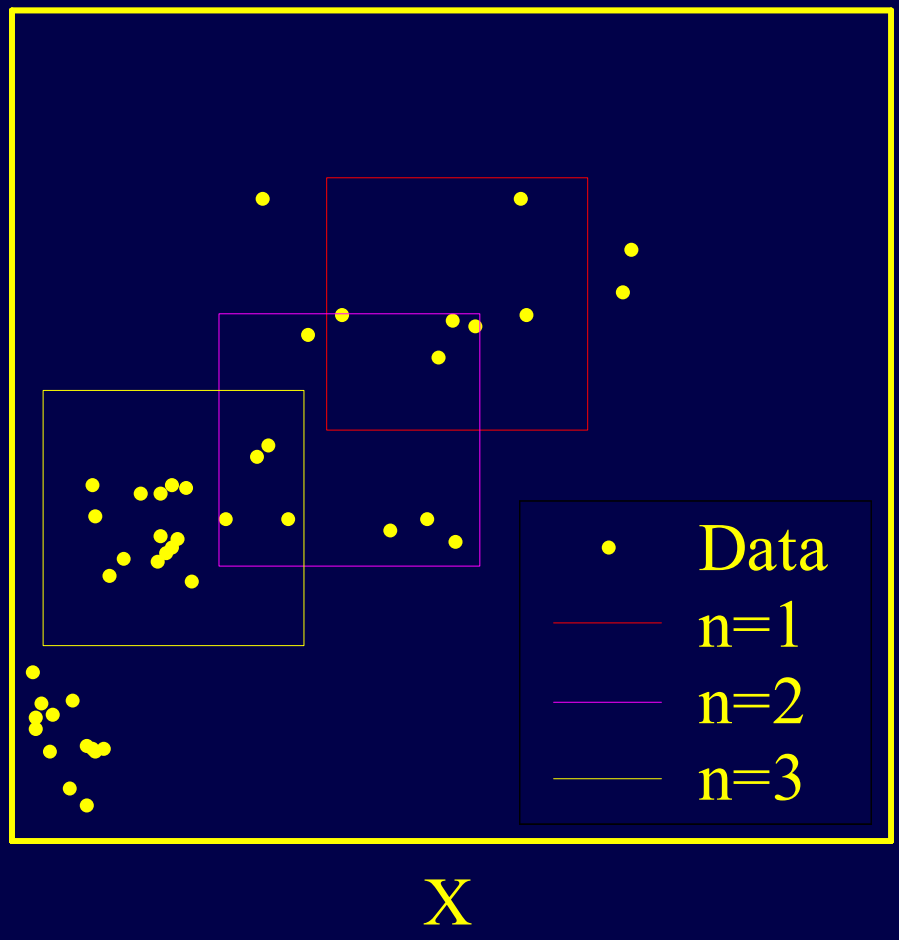
- Sensitivity to search position
  - run many times
  - use answer with smallest *QError*
- To find *K*
  - run with *K* = 1, 2, 3, …..
  - until *QError* just better
- But, slow to do once,
  slower to do many times,
  even slower to do for many *K*s



Enter .... Mean-Tracking

- Developed for High Speed Machinery:
  - Operators set controls differently
  - Various measurements can be made
  - Some combinations good, some bad
  - Need to identify different states
- Also used to find centres of Radial Basis Function Networks
- Finds number of states, and where
  - Just what is needed for k-means
- so MT is pre-processor for k-means

Basic Mean-Tracking



Rectangular window

REPEAT

find data in window

(by comparisons)

centre := mean(data)

UNTIL small move

Window at dense area

Here, have found one area, but are others ...

So have many windows

Initialise each window suitably

REPEAT

Move each window (as above)

IF *n* windows identical, discard *n*-1

IF *n* windows close, combine using weighted average of points in window

UNTIL movement of all windows is small

- Start with many windows
- End with fewer windows, at dense areas

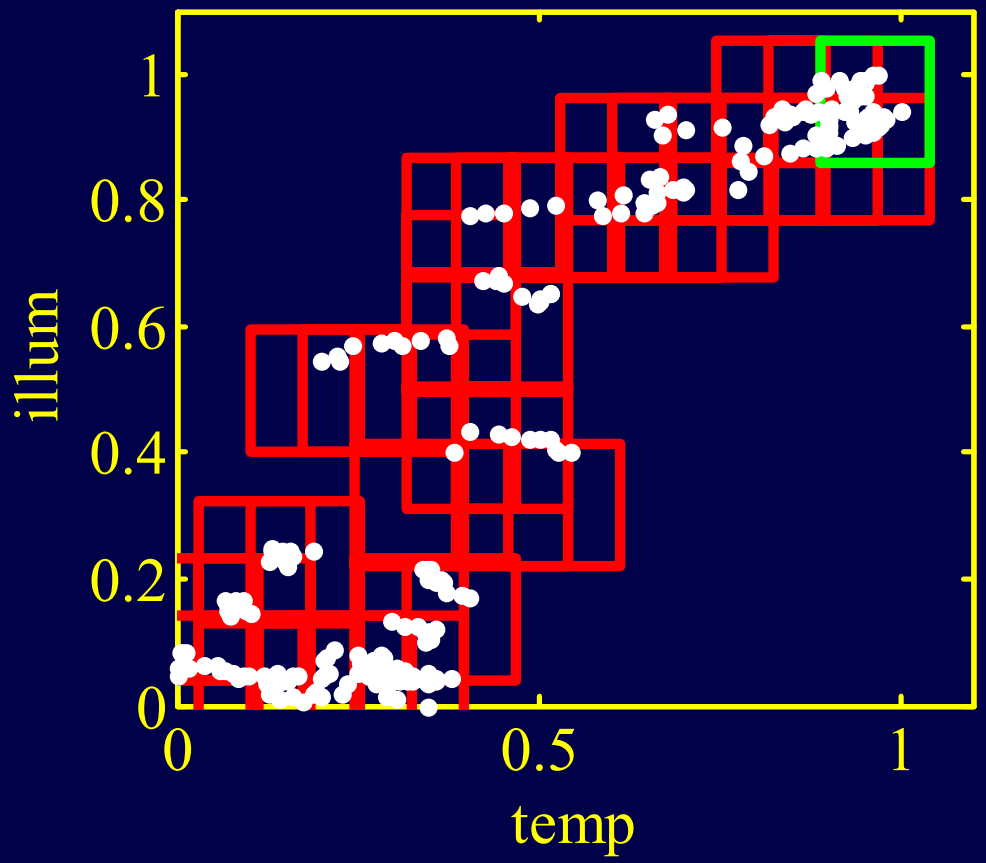
Merging Close Windows

- Find each pair of windows to merge
- Find all groups of such window pairs
  - group is where each window is to be merged with each of the others
  - i.e. find Maximal Cliques
  - NP complete problem
  - Classic Bron-Kerbosch algorithm converted from obscure Algol 60 implementation into efficient MATLAB
- Merge all cliques

Parameters for Mean Tracking

- Window size
  - *f* * std (of each dimension)
  - *f* say 0.5, but needs further work
- Number of Initial Windows
  - 1/3 data set chosen randomly
    - different results each time
  - Linearly spaced overlapping
    - all points covered
    - repeatable result USE

Experimentation



Simulated Data

240 points

values of

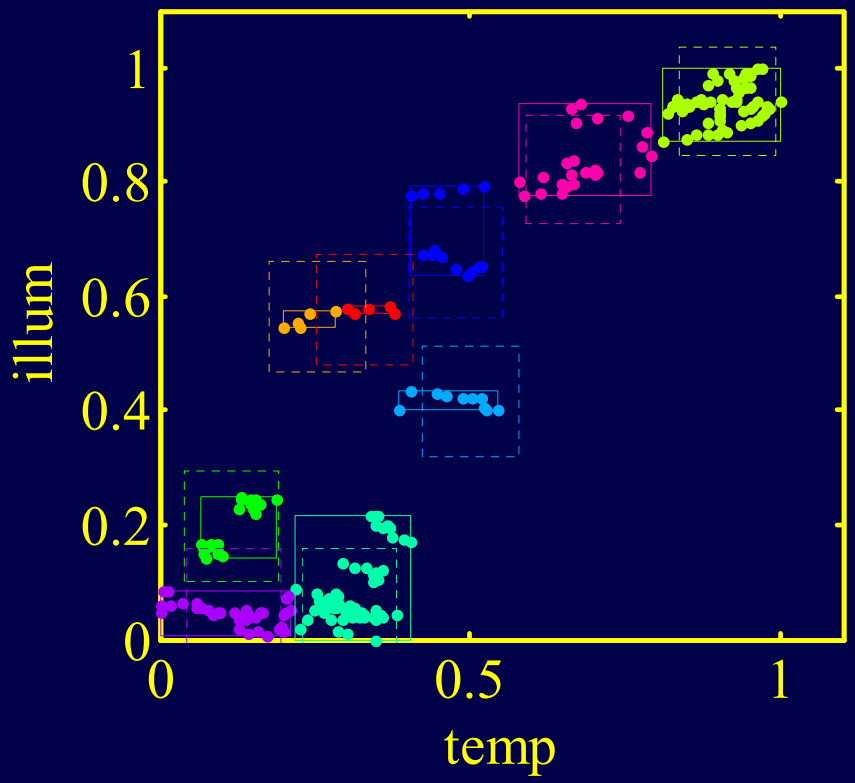
temperature

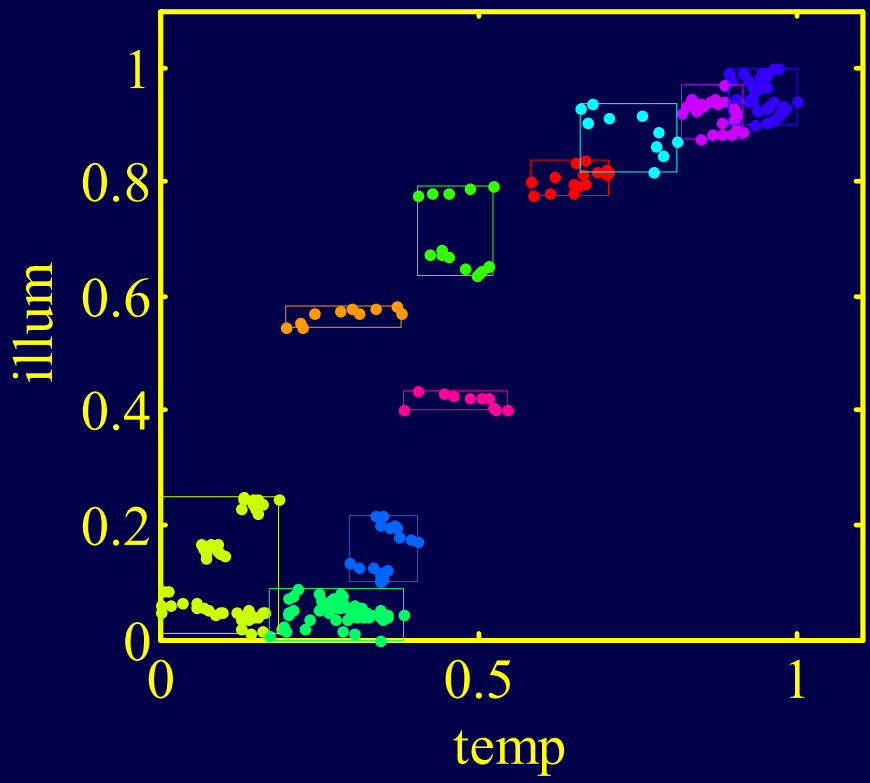
illumination

Shows data and initial positions found for MT windows

Results of clustering

- MT+K-Means
- Best Auto K-Means

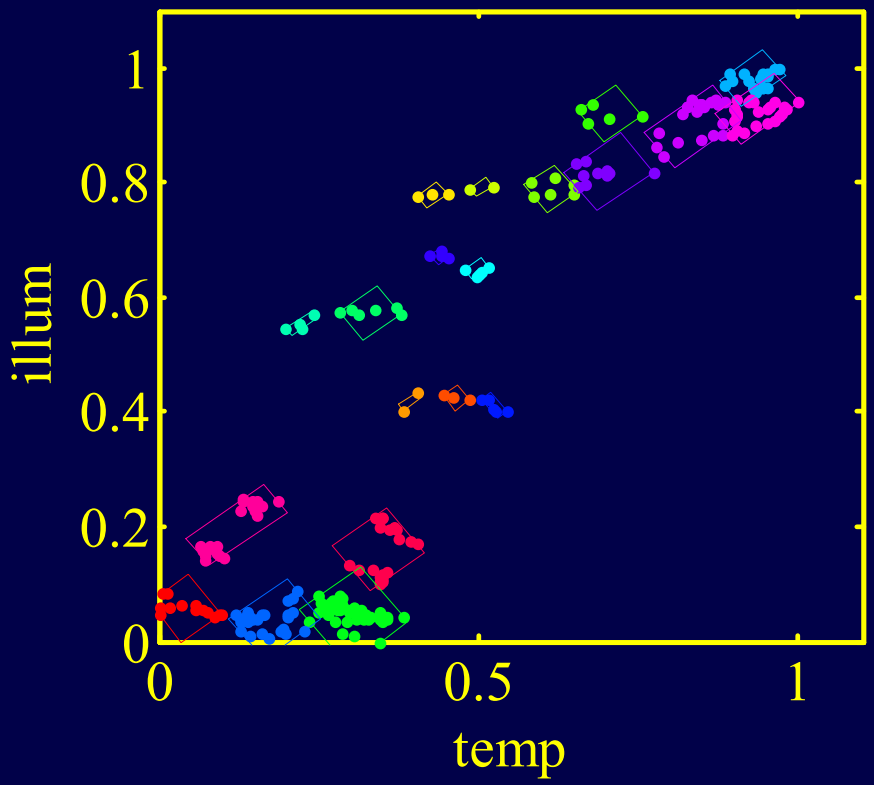


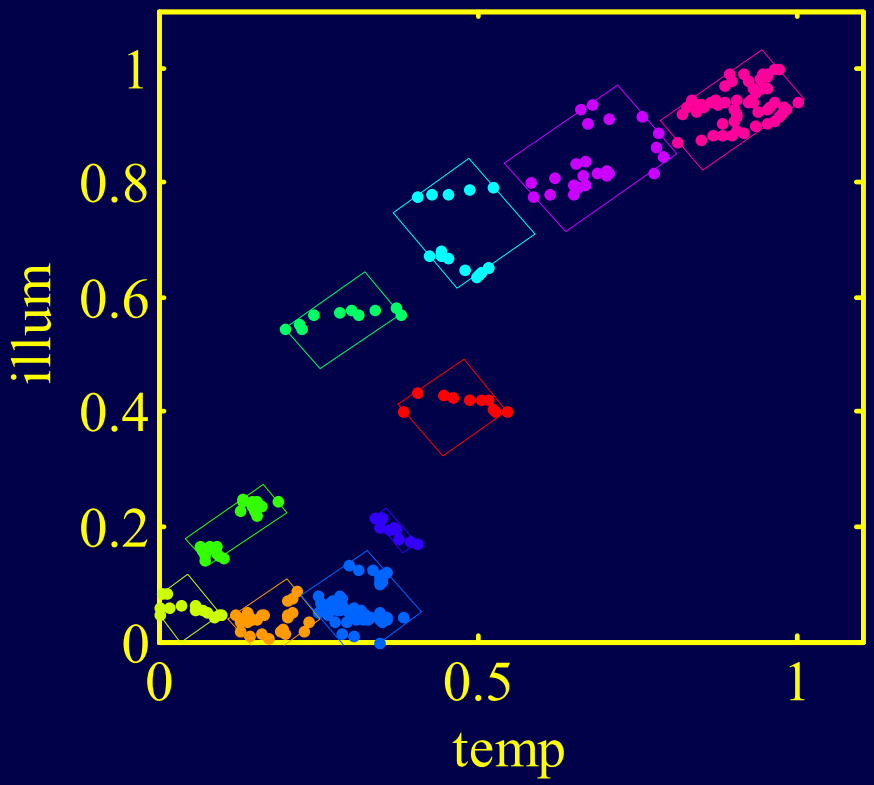


Solid lines round clusters, dotted are MT windows

If rotate using PCA, then cluster

- $f = 0.5$
- $f = 0.75$





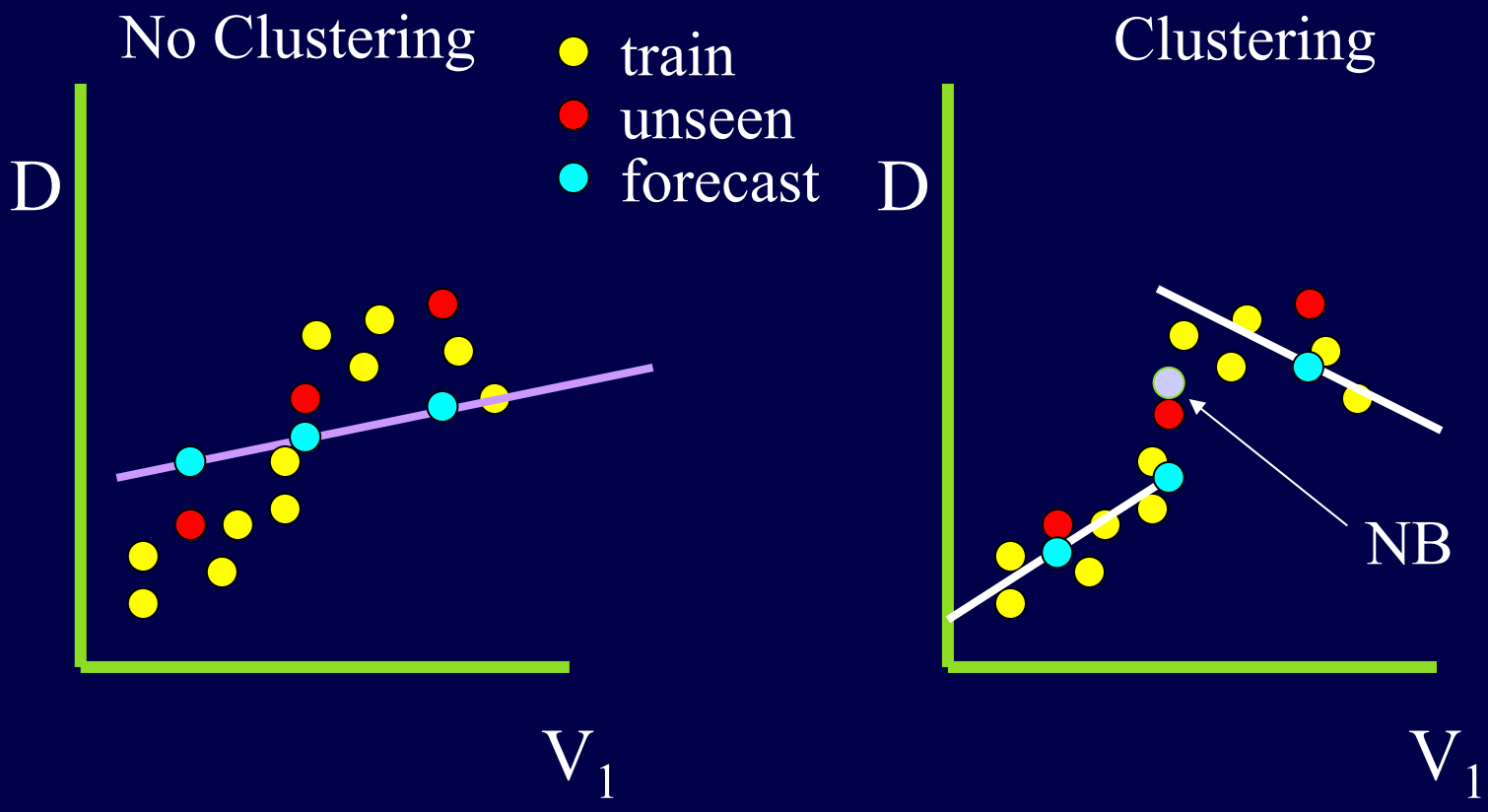
So Clustering successful, let's use it ...

Using Clusters to Forecast

- Have ‘training’ data and ‘unseen’ data (only training data has var to be forecast)
- Cluster training data using algorithm
- For each cluster form linear model
- For each item in unseen data
  - Find clusters of *n* nearest points
  - Find forecasts of each point
  - Return average of *n* forecasts

  (Use *n* nearest points as unseen point may be between training set clusters)

Why Cluster - 1 variable example



Use, $D = c_0 + c_1 * V_1$ **Clearly better if 2 models**

East Slovakian Data

- Need suitable (publishable) data
- East Slovakian Demand Competition

  For EUNITE 2001 conference

  To forecast maximum demand Jan 1999 given data in Jan 1996 and all 1997-1998

Best 3 Methods	SVM	ALN	Average
MAPE *	1.98%	2.15%	2.5%
MaxAE †	51MW	40MW	61MW

*Mean Abs Percentage Error †Max Abs Error

Data Provided and Added

- *Given:* Half hourly demand and
  Average daily temperature
  for all 1997/8 and January 1996
- *Add:* date and time information
  + sin/cos(time) and sin/cos(day)
  (conditions around midnight similar)
  day of week; weekend/weekday
  daily/ half hourly max illumination
- 16 variables - daily or half hourly
- Jan 99 temp: average of Jan96..98

Factors to decide - I

- What is the training set?
- Could use January 1997/1998. Too little?
- Better Jan & Feb; or Dec, Jan & Feb

- Data has half hourly or daily information
- Aim is to forecast max daily demand
- So train only on records where demand at maximum
  - time when this occurs varies
  - can have two maximums in one day

Factors to decide - II

- When forecasting January 1999
- Have 48 forecasts per day; use which?
  - Choose Maximum
  - Choose First
  - Choose Forecast for record closest to training set
  - Ignore any time of day variables, so only one forecast per day
- Handle Weekdays and Weekends separately or together

Factors to Decide - III

- Which variables to use for clustering and how many?
  - Use variables most correlated with maximum demand
  - Use Principle Component Analysis

    use first few principle components

    useful as PCA aligns data with axes as are the Mean-Tracking windows
- For unseen data: find *n* nearest points in training set, but what is *n*?

Factors to decide - IV

- How many forecast variables to use?

  Having decided,

  which variables to use can be found by exhaustive search

  can do as standard matrix method finds model coefficients quickly

  can run on all combinations of up to 3 variables

  Matrix method takes fraction of second
  Clustering takes a few seconds

How to determine these factors

- Train on (part of) 1997 and 1998 data
- Treat Jan 1996 as ‘unseen set’
  making forecasts of max demand
  for each combination of factors
- Determine combination whose forecasts have minimum error (MAPE)
- Then, using this combination,
  make forecasts for Jan 1999
- Effectively, Jan 1996 is a validation set

Implementation

- Matthew Roberts (National Grid) and I developed many MATLAB functions
- Data storage and handling, Statistics, Clustering, Visualisation, etc. + Demos
- Thus for this project, few simplish MATLAB scripts written which called this library

Results : Times Taken to Cluster

Algorithm	Min	Mean	Max
Simulated Data (2 variables)			
MT+KM		0.25	
MT+KM (PCA) *f*=0.5		0.62	
MT+KM (PCA) *f*=0.75		0.26	
Auto K-Means	1.0	1.6	2.9
East Slovakia Data (3 variables)			
MT+KM		0.58	
Auto K-Means	1.3	2.1	3.3

Results : All together forecasts

Extract Method	Corr / PCA	Jan 96 MAPE	Jan 96 MaxAE	Jan99 MAPE	Jan99 MaxAE
First	Corr	1.73%	36MW	2.62%	56MW
First	PCA	1.73%	49MW	2.35%	43MW
Nearest	Corr	1.73%	36MW	2.62%	56MW
Nearest	PCA	2.09%	63MW	2.56%	52MW
Max	Corr	1.73%	36MW	2.62%	56MW
Max	PCA	3.50%	103MW	5.11%	104MW
NoTime	Corr	1.73%	36MW	2.62%	56MW
NoTime	PCA	1.90%	47MW	2.99%	57MW

Results : Separate Weekend/day

Extract Method	Corr / PCA	Jan 96 MAPE	Jan 96 MaxAE	Jan99 MAPE	Jan99 MaxAE
First	Corr	1.70%	55MW	2.42%	56MW
First	PCA	1.56%	54MW	1.93%	47MW
Nearest	Corr	1.60%	45MW	4.07%	119MW
Nearest	PCA	1.44%	58MW	2.06%	45MW
Max	Corr	1.82%	39MW	3.21%	64MW
Max	PCA	2.00%	52MW	3.38%	55MW
NoTime	Corr	1.66%	33MW	3.06%	55MW
NoTime	PCA	*1.43%*	*57MW*	*1.86%*	*47MW*

Comments

- If not separate, Extract method no effect if correlate (time vars not used)
- Extract ‘first’ or ‘no time’ : same MAPE
  But ‘no time’ smaller MaxAE
- Separate Weekends / Weekdays better
- PCA generally better than correlation
- Best result on Jan 96 does indeed give best (MAPE) result for Jan 99
  so validation set has worked

Details of Best Result

- Weekdays
  - Train on Jan, Feb and Dec 1997/8
  - Use first 5 principal components
  - Use cluster mean
  - Use 10 nearest points
- Weekends
  - Train on all of 1997/1998
  - Use first 4 principal components
  - Use model with 3 variables
  - Use 15 nearest points

Detailed Performance

	Jan 96		Jan 99	
	MAPE	MaxAE	MAPE	MaxAE
Weekday	1.39%	55MW	2.16%	47MW
Weekend	1.55%	57MW	1.32%	20MW
Overall	1.43%	47MW	1.86%	47MW
Competition Winner			1.98%	51MW

Conclusion

- The hybrid Mean-Tracking k-Means algorithm is repeatable, successful and more computationally efficient than the popular k-Means algorithm.
- A systematic methodology has allowed forecasts of maximum demand to be made more accurately than any entrants in the competition.
- Used Principal Component Analysis, Correlation, Clustering, Linear Models.
- It could be (and has been) applied to other (confidential) problems.

Acknowledgement

The author would like to thank

- The Uni for allowing the sabbatical,
- National Grid for supporting the sabbatical,
- David Esp and Matthew Roberts of National Grid for useful comments and assistance

References

J.B.MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 1967, 281 -297.

E.L.Sutanto and K.Warwick, Multivariable cluster analysis for high-speed industrial machinery, *IEE Proc Sci Meas. Technol*., *142*(5), 1995, 417-423.

E.L.Sutanto, J.D.Mason and K.Warwick, Mean-tracking clustering algorithm for radial basis function centre selection, *Int J. Control, 67*(6), 1997, 961-77.

East Slovakian Demand Forecasting Competition *http://neuron.tuke.sk/competition/*

C. Bron and J. Kerbosch, Finding all cliques of an undirected graph. Comm ACM, 16(9): 575-577, 1973