Traffic state prediction using Markov chain models

Constantinos Antoniou[1], Haris N. Koutsopoulos[2] and George Yannis[1]

[1]National Technical University of Athens
[2]Northeastern University

ECC 2007, 3 July 2007
Kos, Greece

Outline

- Motivation
- Methodology
  - Model-based clustering
  - Variable-length Markov chains
  - Nearest neighbors classification
- Application
- Main findings

Model-based clustering (I)

- Finite mixture models have been studied in the context of clustering
- Each component probability distribution in finite mixture models corresponds to a cluster
- Problems of determining the number of clusters and choosing appropriate clustering method can be recast as statistical model choice models
- Models that differ in number of components and/or component distribution can be compared
- Outliers can be explicitly handled (through additional distributions)

Model-based clustering (II)

- Cluster analysis
  - Initialization via model-based hierarchical agglomerative clustering
  - Maximum likelihood estimation via the EM algorithm
  - Selection of model and number of clusters using approximate Bayes factors (BIC approximation)

(Full) Markov chains

- One of the most general models for stationary categorical process
- Several applications in many fields, including transport-related
- Rather inflexible in terms of the number of parameters that it can represent
  - For a model with 4 states, chains with 0 to 5 parameters have dimensions of 3, 12, 48, 192 and 768
  - "dimensionality curse"

Variable-length Markov chains

- Allow memory of the Markov chain to have a variable length, depending on the observed past values
  - Computes a huge tree and then prunes it
- Fitting a vlmc from data involves estimation of the structure of the variable length memory
  - Can be reformulated as a problem of estimating a tree
  - Rather efficiently using the so-called context algorithm

Nearest neighbors classification

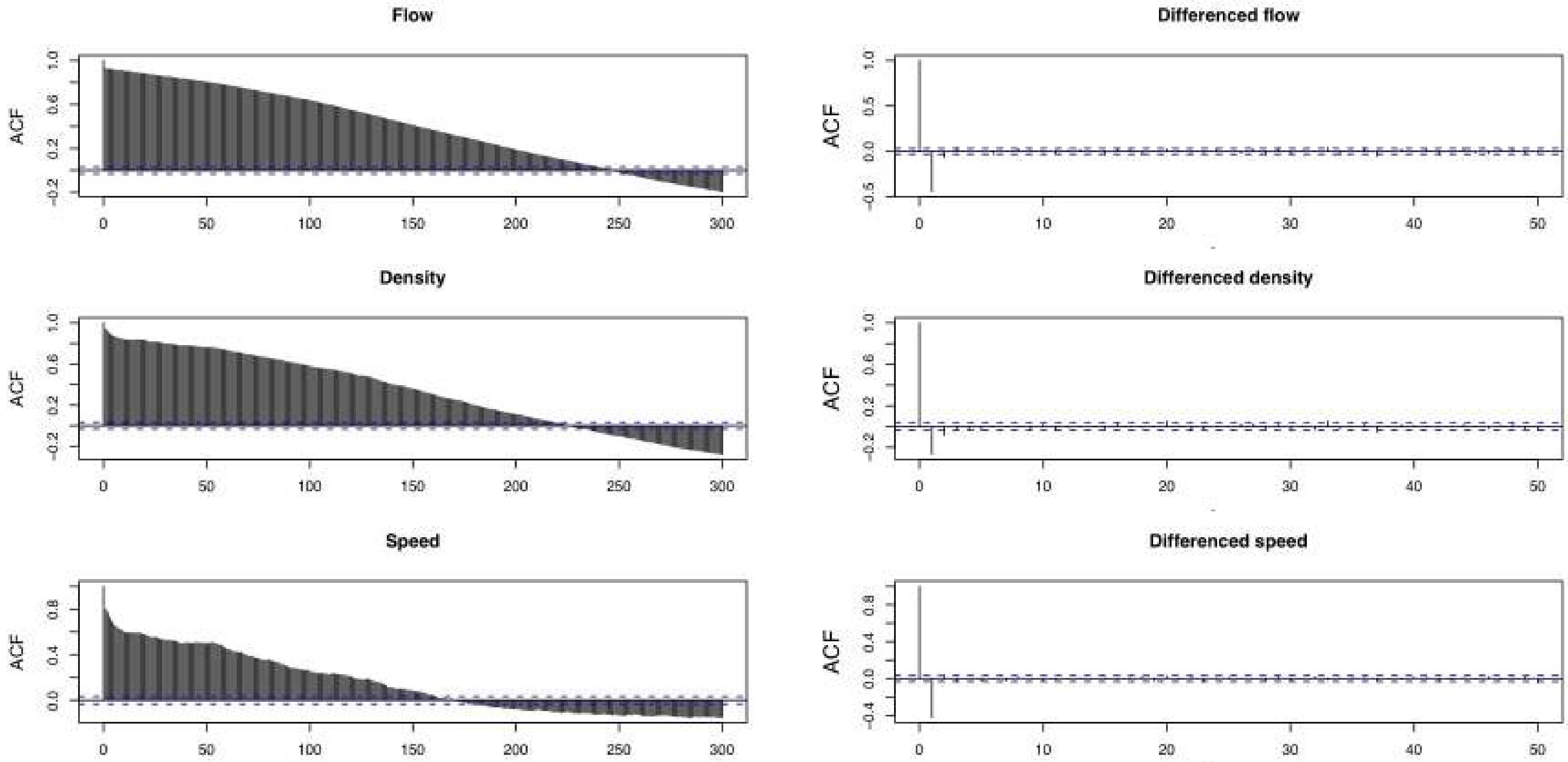
- Used to classify observations into the most appropriate clusters
- K-nearest neighborhood learning is the most basic instance-based method
  - Nearest neighbors are defined in terms of standard n-dimensional Euclidean distances

Application setup

- Freeway I-405 in Irvine, CA
- Morning period data (4am-10am)
- Speed, occupancy and flow data over 2-minute intervals
- Four days of data used for training
  - Model based clustering
  - Variable length Markov chain estimation
  - K-nearest neighbor training
- One different day used for application
- Locally weighted regression has been used for speed estimation/prediction

Stationarity assumption

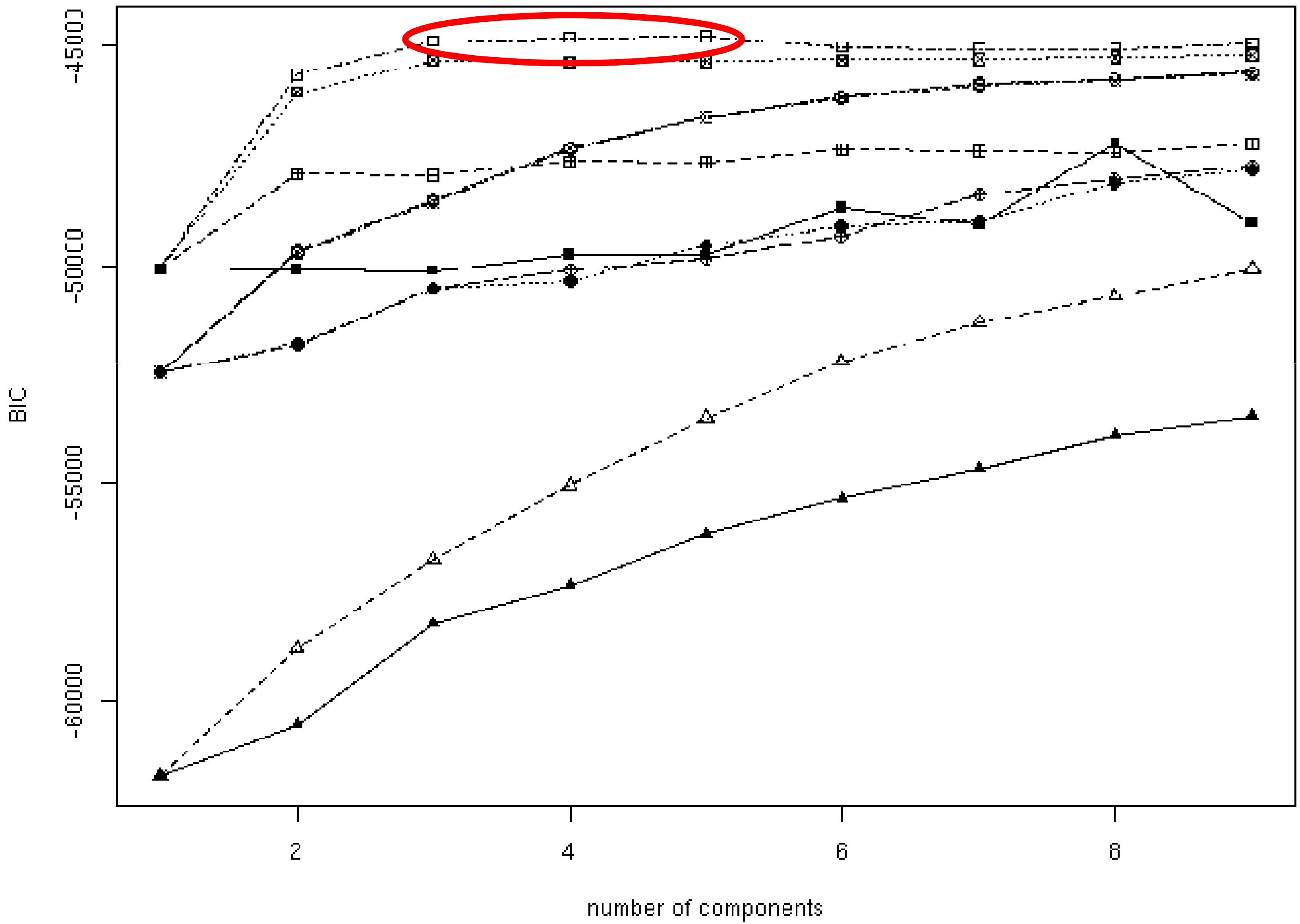
Original data | Differenced data



Clustering and classification

- Considered mixture models
  - "EII": spherical, equal volume
  - "VII": spherical, unequal volume
  - "EEI": diagonal, equal volume and shape
  - "VEI": diagonal, varying volume, equal shape
  - "EVI": diagonal, equal volume, varying shape
  - "VVI": diagonal, varying volume and shape
  - "EEE": ellipsoidal, equal volume, shape, and orientation
  - "EEV": ellipsoidal, equal volume and equal shape
  - "VEV": ellipsoidal, equal shape
  - "VVV": ellipsoidal, varying volume, shape, and orientation

Optimal number of clusters

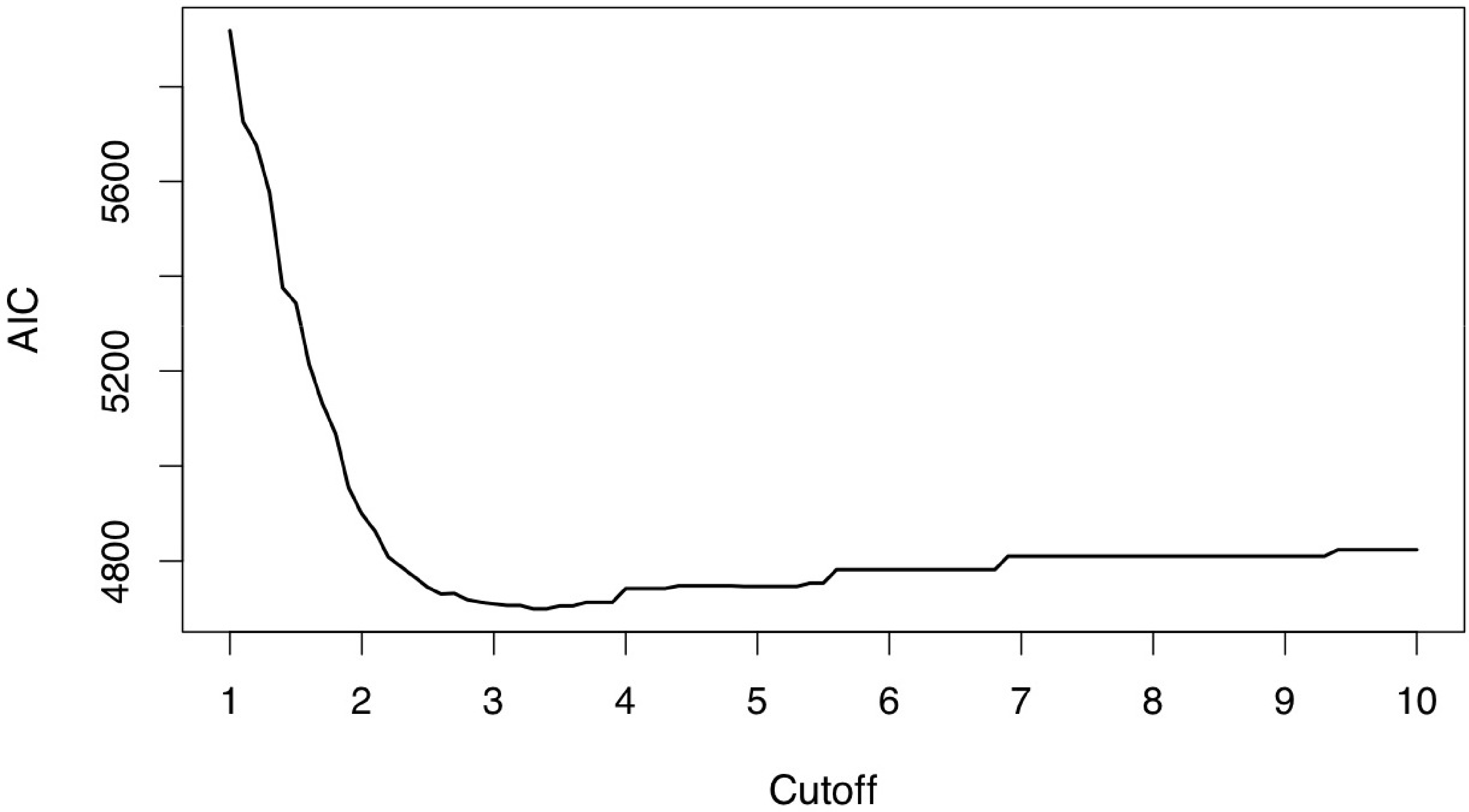


3- vs. 5-mixture models

		3 clusters		
		A	B	C
5 clusters	1	1641	*62*	0
	2	*3*	1217	0
	3	*0*	440	0
	4	*7*	0	172
	5	57	0	0

- 98% of observations are clustered intuitively from the 5 to the 3 clusters
  - 1 and 5 into A
  - 2 and 3 into B
  - 4 into C
- The parsimonious 3 cluster model is retained

“Pruning” cutoff parameter

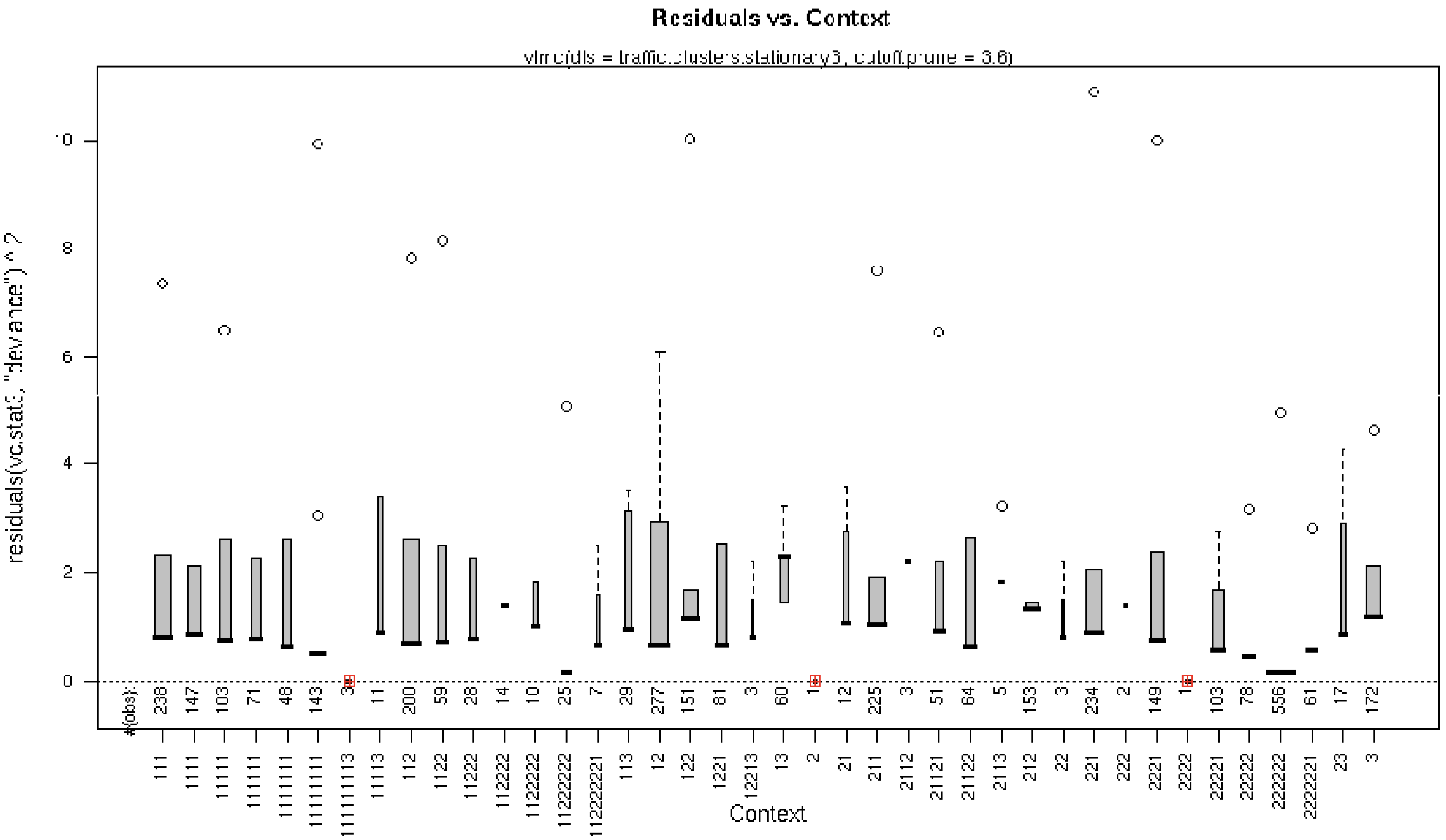


Speed prediction

- Given the estimated variable-length Markov chain
- Predict the state of the traffic for the next interval
- Use locally weighted regression (loess) -trained on data from this cluster only- to compute speed
- Reference case: loess trained on all data – already a powerful approach

	Estimation	Reference prediction	Markov-based prediction	Prediction improve-ment
RMSN	0.0449	0.0883	0.0801	9.3%
RMSPE	0.0504	0.1049	0.0989	5.7%
MPE	0.0044	0.0085	0.0033	60.7%
U	0.0221	0.0434	0.0396	8.8%
U^M	0.0177	0.0046	0.0017	63.6%

Model diagnostics



Conclusion

- A methodology for identification and short-term prediction of traffic state
  - Model-based clustering
  - Variable length Markov chains
  - Nearest neighbor classification
- Application in a freeway network in Irvine,CA
- Potential uses may include automated incident detection, indirect capacity estimation etc

Thank you for your attention!

Questions?