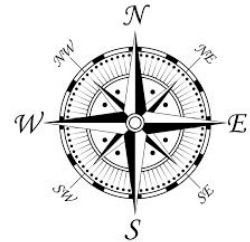


STATS R US[®]

Alan Min, Jack VanSchaik, Karan Samel, Sam Eschker
Kent Gauen



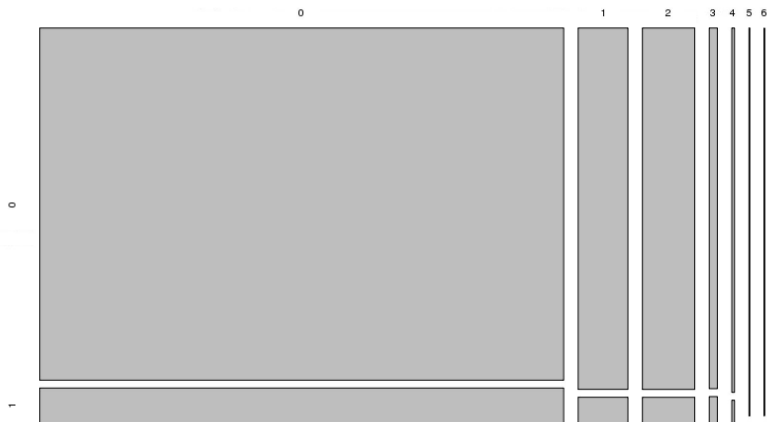
$$\varphi : \mathbb{R} \rightarrow \mathbb{R}$$



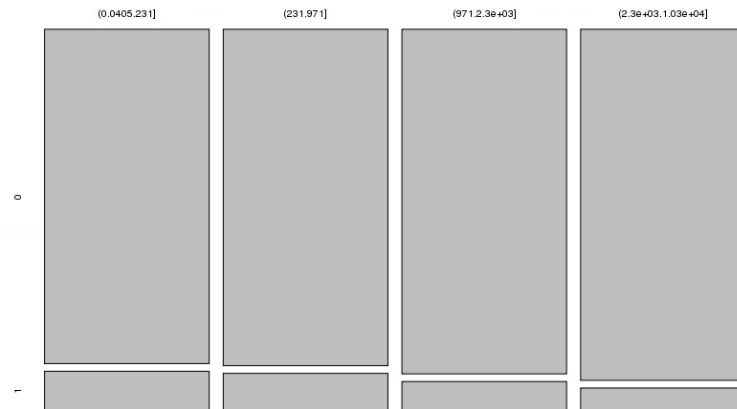
Feature Creation and Selection with Naive Bayes

```
[15] "Iteration: 21 Score: 0.91193 With vars: hour"  
[1] "site_name" "cut_orig_destination_distance"  
[3] "is_mobile" "is_package"  
[5] "channel" "srch_adults_cnt"  
[7] "srch_children_cnt" "prop_is_branded"  
[9] "prop_starrating" "distance_band"  
[11] "hist_price_band" "popularity_band"  
[13] "month" "day"  
[15] "hour"
```

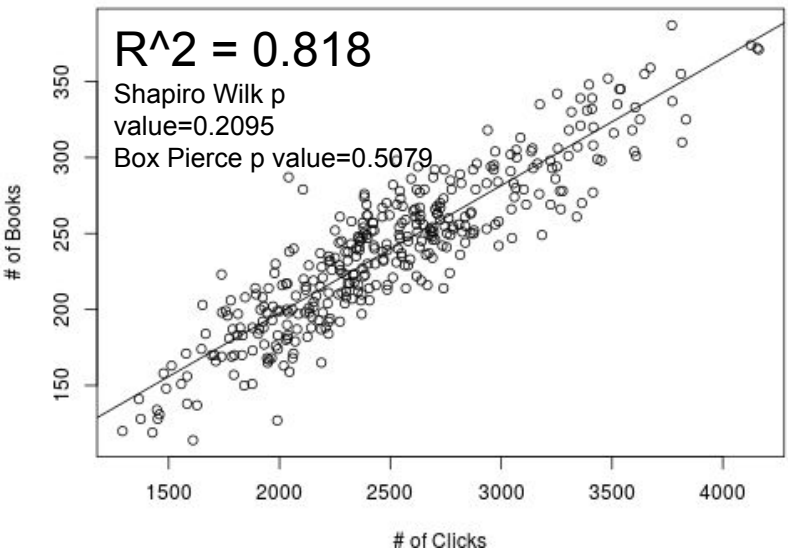
Number of Children vs Tickets Bought



Distance to Destination vs Tickets Bought



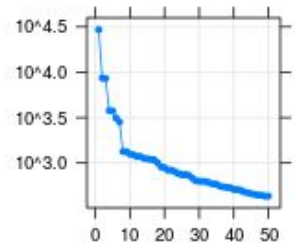
Daily Books vs Clicks



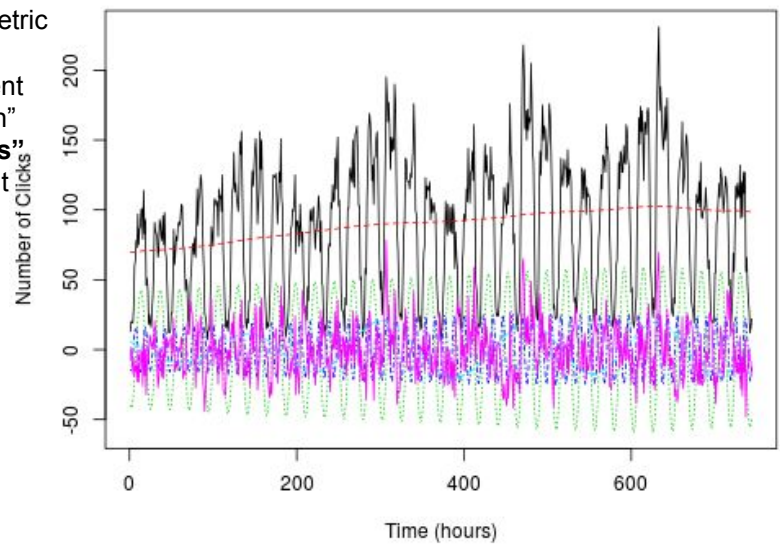
Singular Spectral Analysis - nonparametric time series modeling.

- 1 trend component
- 1 weekly "season"
- **3 daily "seasons"**
- Noise component

Component norms

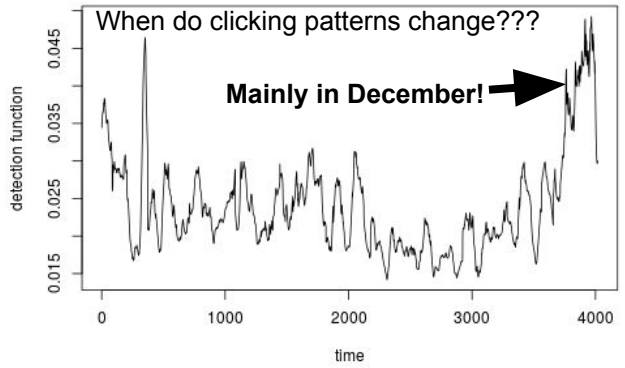


Reconstructed Series

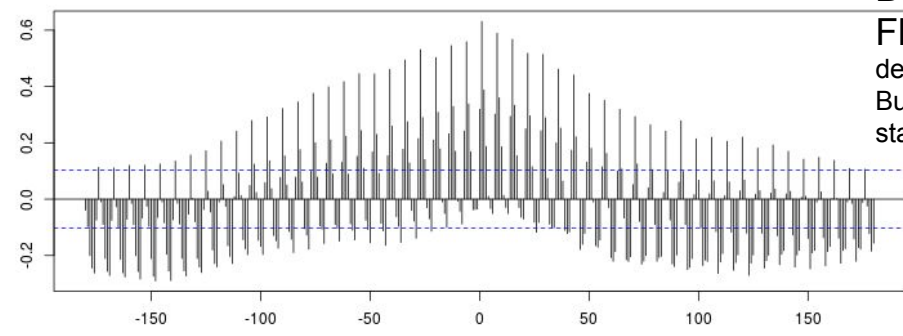


SSA Detection Function-

When do clicking patterns change???



CCF of daily number of bookings and flights flown



Bookings and Flights Flown - Used departure delay data from Bureau of transportation statistics

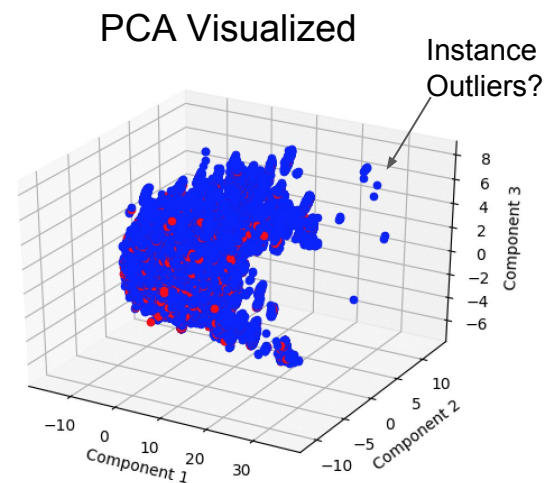
- CCF of books vs flights flown
- Share weekly pattern
- Book as well as 100 days in advance

Cleaning, Reduction, and Fitting

- Joined the datasets given to create a richer dataset
- Wanted to convert the data to numeric values so we could predict bookings
 - Categorical values were assigned indices or one hot encoded (192 total features)
- Resulting data was normalized and reduced

	Accuracy*	ROC
RF	0.93805	0.49720
MLP	0.92763	0.51366
KNN	0.94382	0.50147
SVM	0.94565	0.5

PCA Components Used	% Total Variance Preserved
2	0.77348
5	0.81977
10	0.86558
15	0.89945
25	0.95336
50	1



*baseline model accuracy \approx 0.94796