Siemens 2017 Wind Analytics Contest

March 2017

UCF

Team Members (alphabetically)

- Mahsa Almaeenejad
- Kanak Choudhury
- Md Jibanul Haque Jiban
- Taha Mokfi

Faculty supervisor:

- Alexander Mantzaris

Content

- Data understanding
- Modeling
  - Exploratory Data Analysis
  - ANOVA Model
  - Markov Chain Model
  - Path Analysis
  - Clustering Approaches
  - Network Analysis
  - Pattern Mining
    - Association Rules
    - Sequential Rules
- Summary

Siemens Wind Power contest

- Produce power using wind turbine
- During operation, wind turbines automatically generate **event information, warnings, and faults (Code)**
- Some of which then cause the turbine to shut down and require intervention before restart
- Wind turbines are maintained by technicians, who visit the turbines when some action is needed
- Stored all those information in database

UCF

Introduction

- Descriptive task
  - Fundamental Statistical analysis
  - Unsupervised learning methods
- Visualization was important
- Wrote more than 2000 lines of codes in R
- Created 17 different functions (350 lines of code)

Data Preparation and Exploration

Data Outline

- All variables are categorical except Visit Duration
- 37 Parks information
- 1614 Stations in all Parks
- 642 different Codes
- 7653 different visit information

Error Codes can appear in three different time points

- Before the Visit Start
- Within the Visit Duration
- After the Visit Duration

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Park_Name	FactorA	FactorB	FactorC	FactorD	StationID	VisitType	VisitId	ManualSto	VisitStartTime	VisitDurMin	Code	ManualSto	TimeOn
2	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	3173	FALSE	8/24
3	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	63027	FALSE	8/24
4	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/2
5	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/2
6	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/2
7	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/2
8	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/2
9	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	1001	TRUE	8/24/2
10	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	63027	FALSE	8/24/2
11	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	2	FALSE	8/24/2
12	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	7	FALSE	8/24/2
13	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	18	FALSE	8/24/2
14	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13902	FALSE	8/24/2
15	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14001	FALSE	8/24/2
16	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	15001	FALSE	8/24/2
17	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	8	FALSE	8/24/2

UCF

Data Construction

- Different techniques has been used to construct data
  - Chain based count data construction (Transition matrix)
  - Frequency based data construction for each visit ID

	VisitId	Code	EventWarningStop	Time_On
1	178	63027	Stop	2016-08-24 08:33:00
34	178	3173	Stop	2016-08-24 08:33:00
33	178	13	Event	2016-08-24 11:33:00
31	178	14	Event	2016-08-24 11:39:00
61	178	13	Event	2016-08-24 11:39:00
47	178	14	Event	2016-08-24 11:44:00
20	178	1001	Stop	2016-08-24 11:49:00
51	178	13	Event	2016-08-24 11:49:00
19	178	2	Event	2016-08-24 11:50:00
30	178	18	Event	2016-08-24 11:50:00
35	178	63027	Stop	2016-08-24 11:50:00
66	178	7	Event	2016-08-24 11:50:00
9	178	13902	Stop	2016-08-24 11:51:00
14	178	14001	Warning	2016-08-24 11:51:00
18	178	8	Event	2016-08-24 11:51:00

(a)

(b)

	Stop	Event	Warning
Stop	1	3	1
Event	3	5	0
Warning	0	1	0

(c)

	Stop	Event	Warning
Stop	0.200	0.600	0.2
Event	0.375	0.625	0.0
Warning	0.000	1.000	0.0

(d)

	Stop	Event	Warning
Stop	0.07142857	0.21428571	0.07142857
Event	0.21428571	[illegible]	0.00000000
Warning	0.00000000	0.07142857	0.00000000

(e)

	Stop	Event	Warning
Stop	0.25	0.3333333	1
Event	0.75	0.5555556	0
Warning	0.00	0.1111111	0

	VisitId	1001	1002	1003	1004	1005	1007
1	178	3	0	0	0	2	0
2	252	4	0	4	0	3	0
3	450	3	0	0	0	1	0
4	534	0	0	0	0	0	0
5	691	1	0	0	0	1	0
6	896	1	0	0	0	0	1
7	904	5	0	0	0	2	0
8	1004	0	0	0	0	0	0
9	1034	2	0	1	0	0	0
10	1082	0	0	0	0	0	0

(a) Subset of data, (b) Transition matrix based on the data (a), (c) Conditional probability matrix (For Example Probability of appearance of error code that is Event given that the present state is Stop is 0.60), (d) Probability matrix (for example, probability of appearance of consecutive error codes that are Stop and Event is 0.214), (e) Conditional Probability with respect to column factor.

EXPLORATORY DATA ANALYSIS (EDA)

Objective:
Whether there is any association between variables

- Showed dependence between variables
- For example, number of type of codes are associated with Parks
- Indicates codes pattern among parks are not same
- It is better to analyze by parks

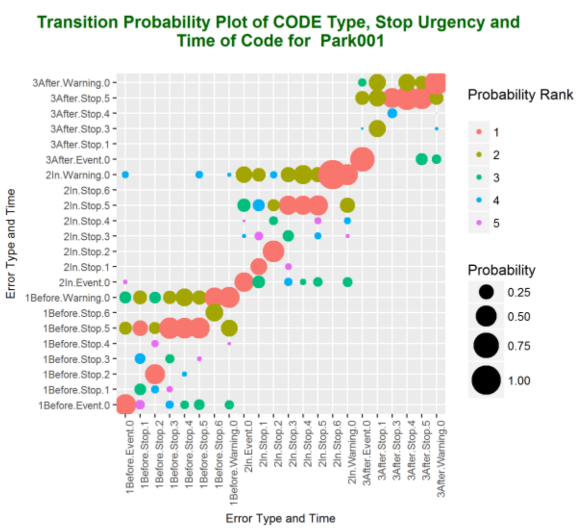
Chi-squared test for Association

Association between	Test statistic	P-value
Parks and Codes	436308	<0.0001
Manual Stop Vs Code	3487	<0.0001
Codes and Factor A	124101	<0.0001
EventStopWarning and Parks	3718	<0.0001
EventStopWarning and Codes	143698	<0.0001
EventStopWarning and Codes	98412	<0.0001
EventStopWarning and Factor A	1183	<0.0001
EventStopWarning and Factor C	980.12	<0.0001
EventStopWarning and Factor D	193.68	<0.0001
EventStopWarning and Manual Stop	11992	<0.0001

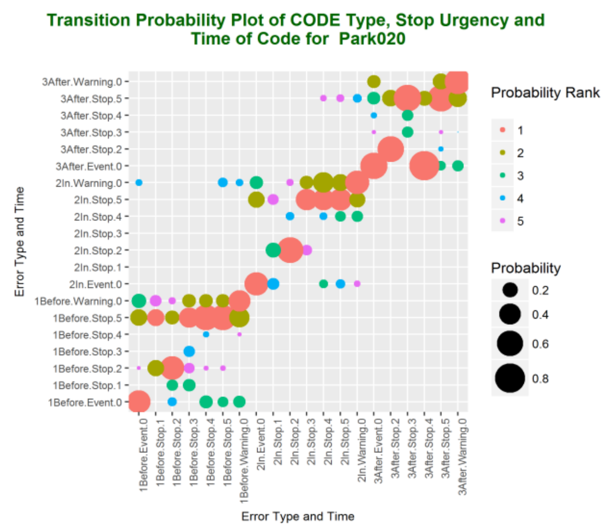
Conditional Probability Based on Markov Model

Conditional Probability Model Based on Markov Chain

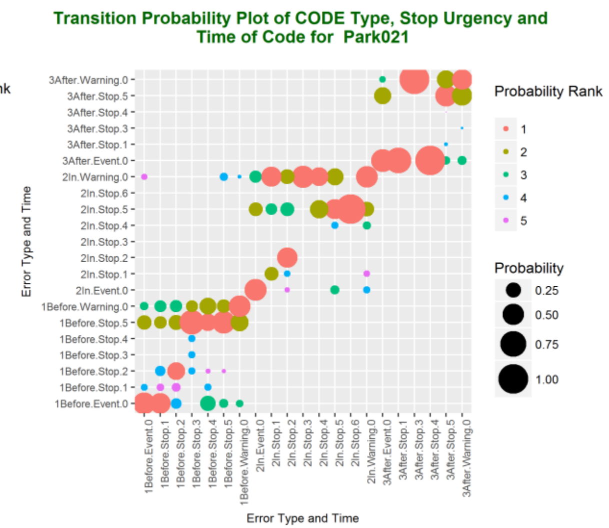
- For example, if there is a warning code and it appears before the visit start, the largest probability that the next code that may appear before the visit is warning code
- Parks with small number of assets show different pattern than those with larger number of assets (park1→ 87, park20 → 112, park21 → 66)



(a)



(b)

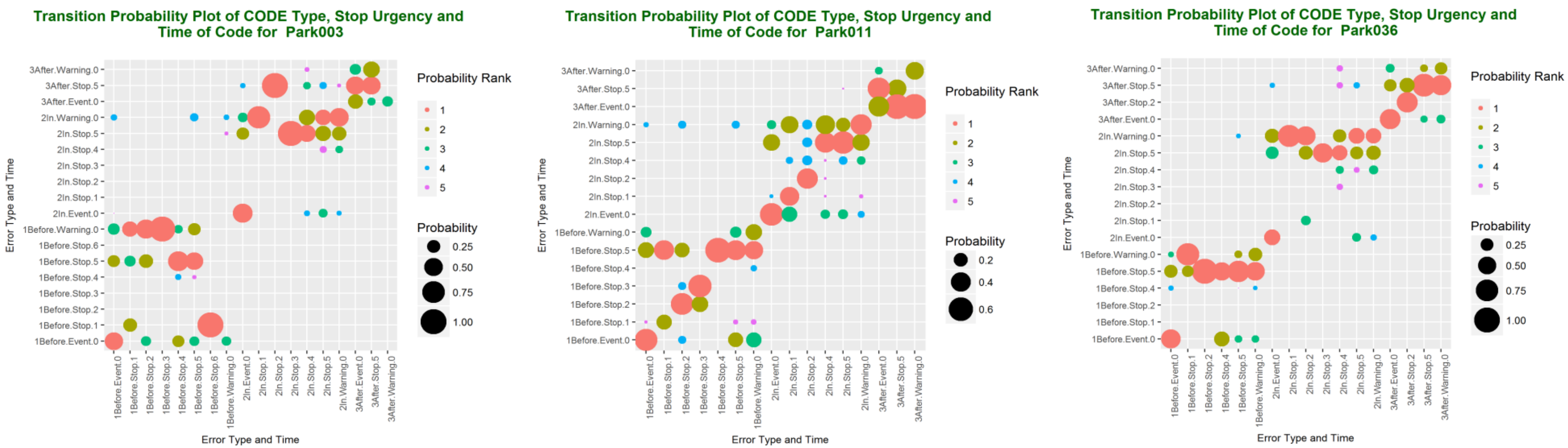


(c)

Conditional Probability Plot. X-axis represents the present state and y-axis represents the future state. Color of the graph represents the rank of the probability (Rank 1 (largest probability) – Rank 5) and size of bubble represents the value of the probability of the state transitions. This plot represents the probability of appearance of type of code with urgency given the present appearance of the type of code with another urgency.

Conditional Probability Model Based on Markov Chain

- Parks with small number of assets show different pattern than those with larger number of assets (park3 →21, park11 → 14 and park36 → 30)



(d) (e) (f)

Conditional Probability Plot. X-axis represents the present state and y-axis represents the future state. Color of the graph represents the rank of the probability (Rank 1 (largest probability) – Rank 5) and size of bubble represents the value of the probability of the state transitions. This plot represents the probability of appearance of type of code with urgency given the present appearance of the type of code with another urgency.

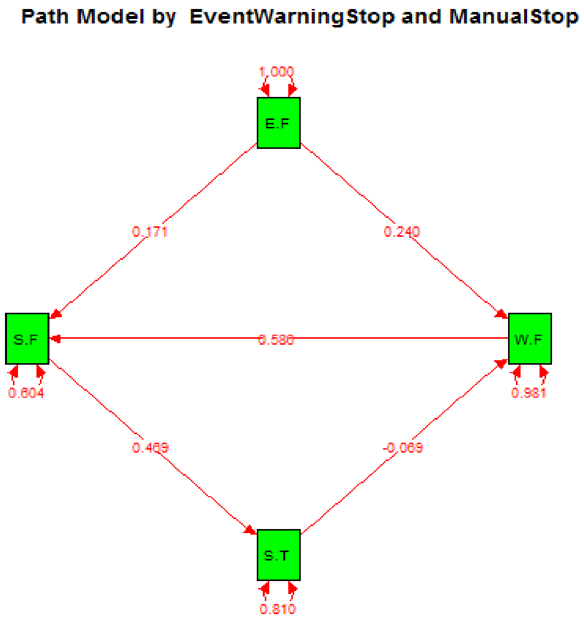
UCF

Path Model

Path Model

- What is path of number of different type of codes?
- Is there any direct and indirect effect on type of codes that is caused by other type of codes?

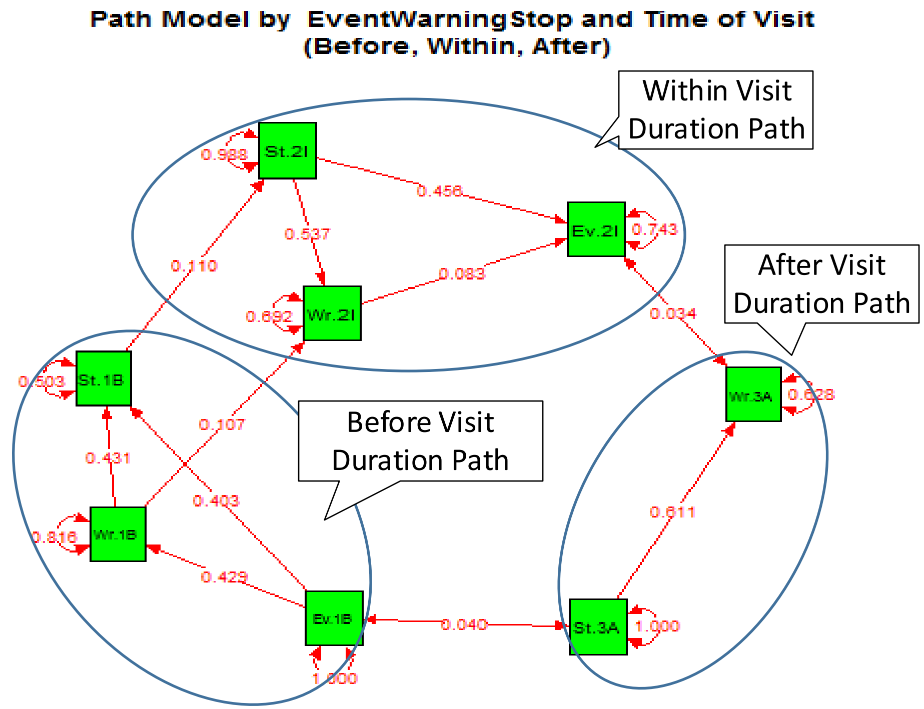
- Number of Stop.FALSE type of codes caused by number of Event.FALSE (0.171) and Warning.FALSE type of codes (0.586)
- Event.FALSE and Warning.FALSE are positively associated with Stop.FALSE.
- Similarly, number of Warning.FALSE type of codes can be predicted by Event.False (0.24) and Stop.TRUE (-0.069).



Path model for the model A. Arrow indicates the direction of the variable and the value indicates coefficient of the parameter. S.F, E.F, W.F, and S.T indicate Stop.FALSE, Event.FALSE, Warning.False, and Stop.True respectively.

Path Model

- Event_Before → Stop_Before
- Event_Before → Warning_Before → Stop_Before
- However, within the visit duration, the path is inverse
- Stop_Within → Event_Within
- Stop_Within → Warning_Within → Event_Within



Path model for the model C. Arrow indicates the direction of the variable and the value indicates coefficient of the parameter. Ev, St, and Wr stand for Event, Stop and Warning respectfully and 1B, 2I and 3A stand for Before, Within and After visit duration.

Association and Sequential rules mining

Algorithms

- Apriori algorithm which is among 10 most used data mining techniques
- The cSPADE algorithm were applied. Lots of proved applications in market basket analysis and web pattern mining
- Three categories of patterns were extracted based on visit occurrences: Before, Within and After.

Association rules on Before Visit Dataset

- A total of 448 rules were extracted in this category.

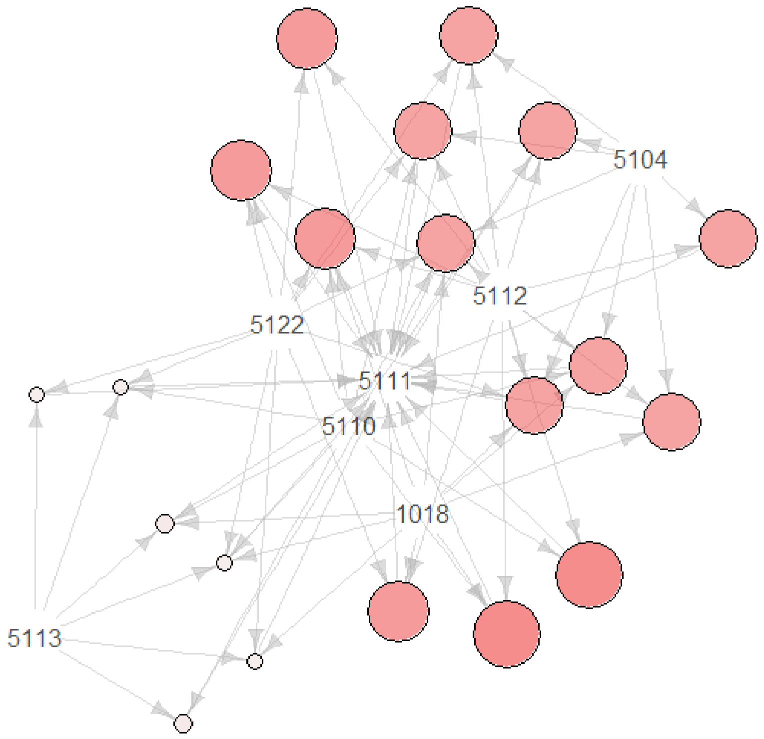
Number of error code in each rule	2	3	4	5	6	7
Number of rules	28	124	158	100	33	5

- Mean of the lift parameter calculated for this category of rules is about 4.4 which indicates that most of the rules are not happening in random pattern.
- Sequence doe not matter here

antecedent => consequent	support	confidence	Lift
{15001} => {14001}	0.17	0.99	5.55
{1018,5104,5110,5112,5122} => {5111}	0.17	1	5.60
{13902,15001} => {14001}	0.16	0.99	5.55
{1018,5112} => {5104}	0.17	0.99	5.52
{1022,7111}=>{1001}	0.02	1	5.77

Visualizing patterns

- Size: support
- Color: Lift
- Codes appeared in many of left hand side and right hand sides of the rules for before dataset.
- code 5113 occurred mostly in left hand side which means the patterns in which this code is caused other codes are more frequent than the patterns in which Code 5113 is the consequence.
- Code 5113 associates with less support and lift



Sequential Pattern Mining

- Data preparation consume more processing than the previous cases.
- New datasets were created based on the time sequence of code occurrence in each visit.
- All the codes which happed in the sequence of N minutes were combined.

VisitId	TimeOn	Code
178	8/24/2016 8:33	1001
178	8/24/2016 8:36	1002
178	8/24/2016 8:39	1003
178	8/24/2016 9:39	1005

VisitId	Code	Time Id
178	1001,1002,1003	1
178	1005	2

Sequential Pattern Mining

Rule 1:

If code 1007 happens then code 9 and 63003 will happens after this code within 6 minutes or more

Rule 3:

If codes 1001 and 1020 and 1022 happens in a sequence (order is not important) of less than five minutes of each other, then code 7111 will happen in next sequence which is in next 5 minutes of more

#	Dataset	Rule	Support	Combined based on N minutes
1	Before	{1007},{9,63003}	0.03	N=5
2	After	{10105},{3130}	0.02	N=5
3	within	{1001,1020,1022},{7111}	0.003	N=5
4	Before	{10105},{3130}	0.03	N=15
5	After	{59},{59}	0.02	N=15
6	Within	{ 1001,1020,1022}	0.03	N=15
7	Before	{3130},{10105}	0.02	N=30
8	After	{ 1111},{ 13,14}	0.01	N=30
9	Within	{ 1020,1023}	0.1	N=30
10	Before	{ 1007},{ 5113}	0.08	N=1
12	After	{ 13},{14}	0.06	N=1
13	Within	{1020,1023}	0.12	N=1
14	Before	{1007},{5113}	0.09	N<1
15	After	{13},{13},{14}	0.02	N<1
16	within	{1020,1023},{1001}	0.03	N<1

UCF

Sequential Pattern Mining

- Number of rules that are extracted is huge.
- Rule 6 has %2 support on before dataset and just %0.6 support in after dataset.
- This sequence is more probable to happen before the visits than after the visit.

rules	Rules	After	Before
1	{59},{59}	0.024	0.041
2	{2,7,8,18,13902,14001,15001}	0.007	0.036
3	{10105},{3130}	0.013	0.025
4	{3130},{10105}	0.008	0.024
5	{9,63003},{2,7,8,18,13902,14001,15001}	0.006	0.02
6	{5122,13140},{5122}	0.028	0.012

Clustering Visits

Clustering Visits

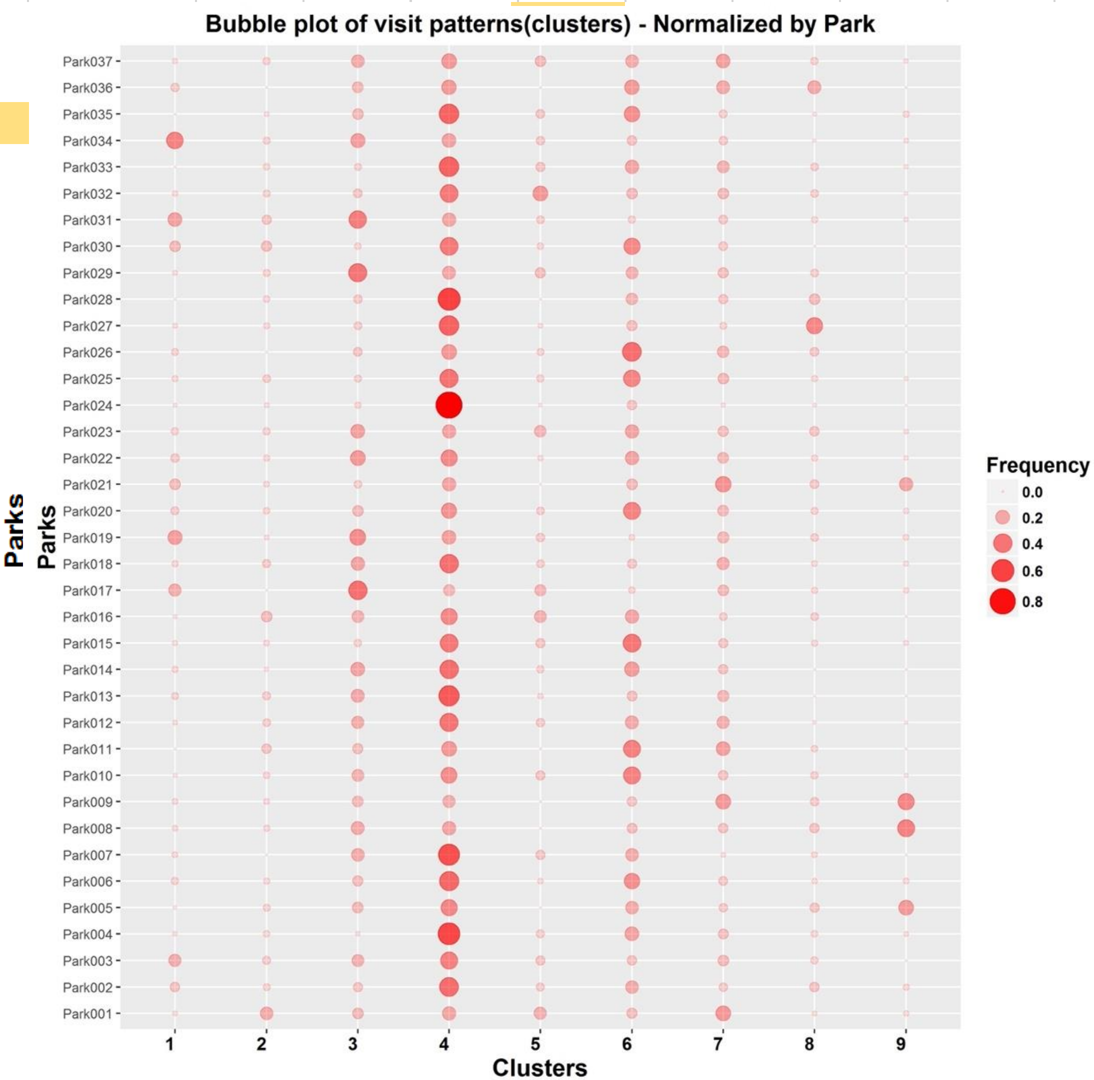
- Objective:
  - Find visits with the similar error code pattern
- Data Preparation:
  - Normalization
  - Binarization
  - Quantile
  - PCA
- Modeling step:
  - Hierarchical methods
  - K-means (9 clusters)
  - DBSCAN
- Evaluation:
  - Davies-Bouldin Index

VisitId	1001	1002	1003	…	Cluster
175	4	0	6	…	1
18	0	2	1	…	4
1002	3	0	4	…	1

7653×642

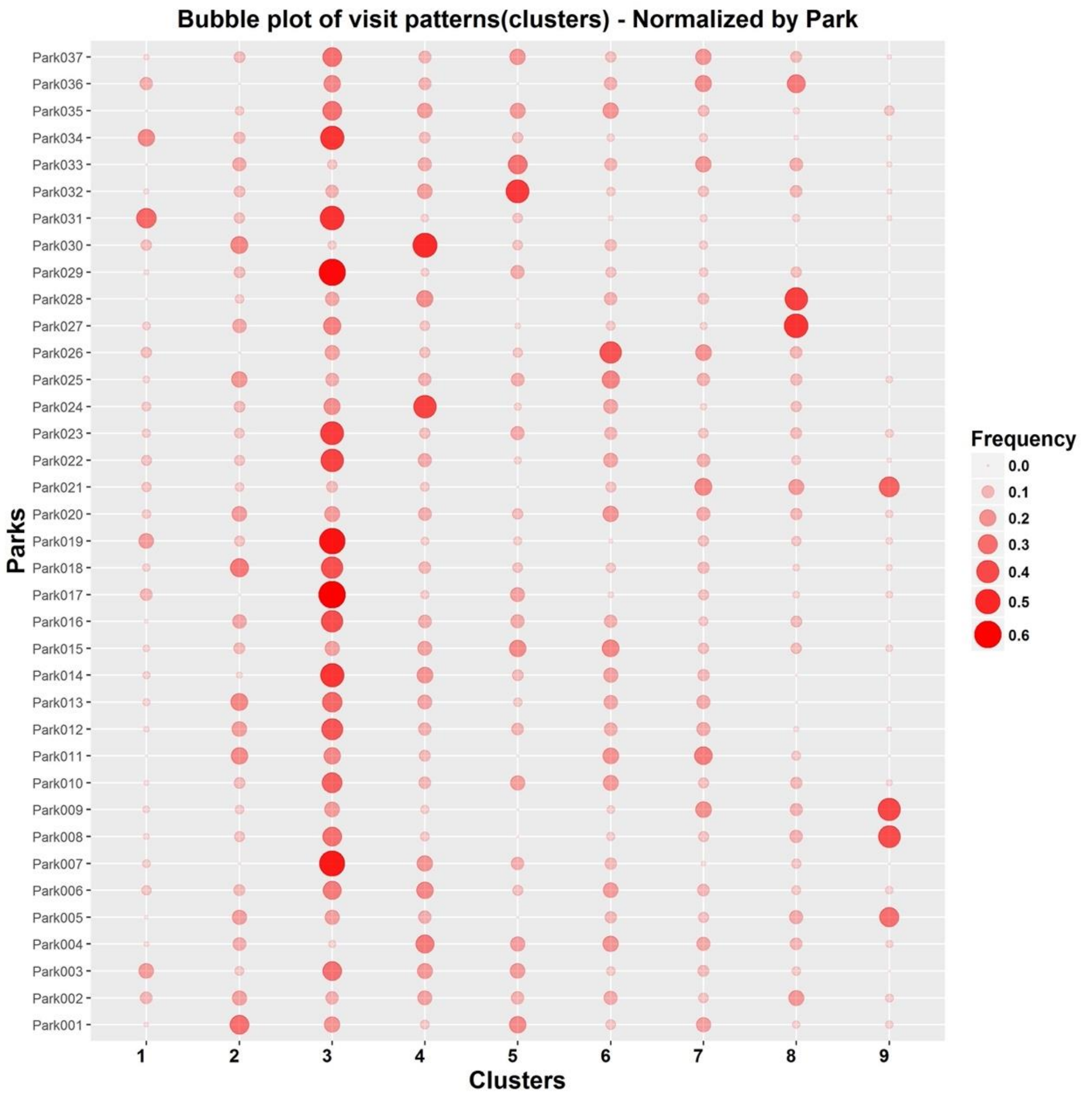
Clustering output

1) Aggregated table
2) Bubble plot – normalized by park
3) Bubble plot – normalized by cluster
4) Cluster Parks



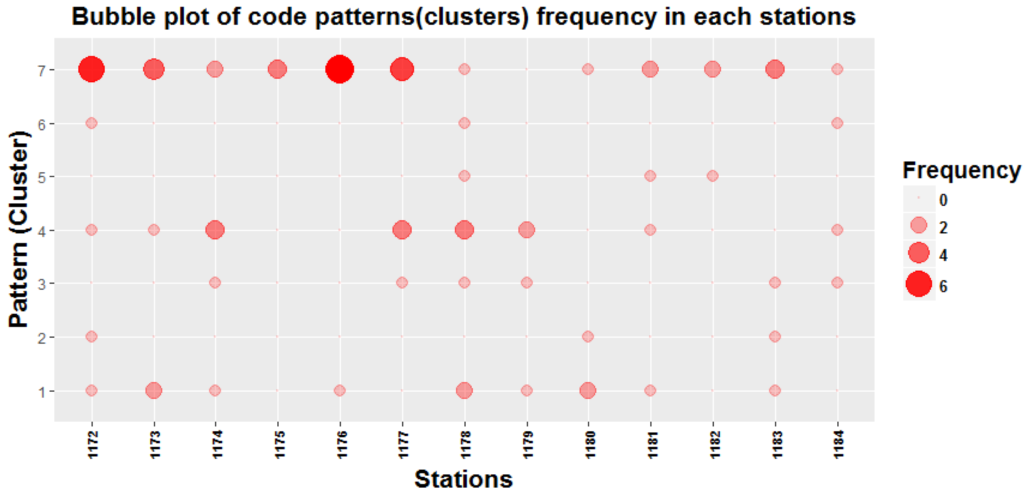
Before – After – In

1) Before – normalized by Park
2) In – normalized by Park
3) After – normalized by Park

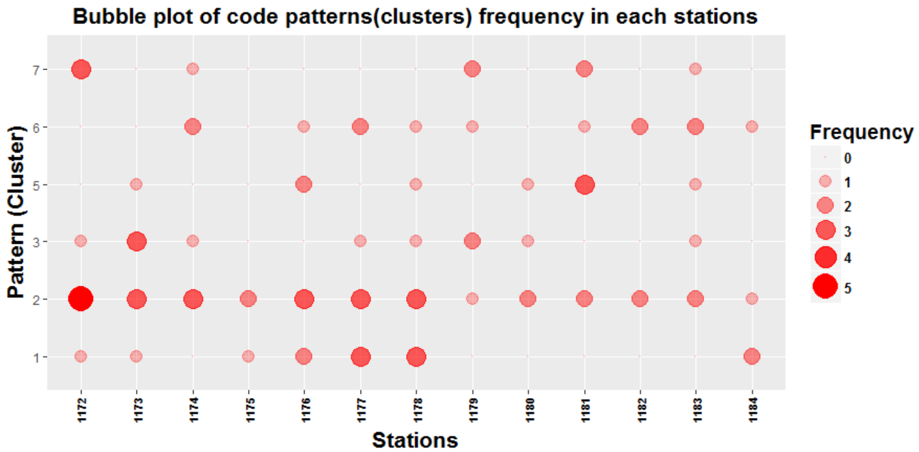


Analyze the stations based on Visit clusters

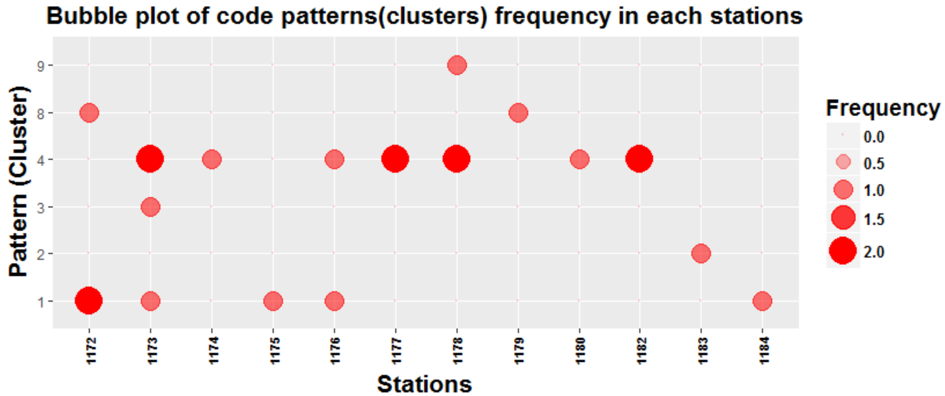
- Frequency of different visits' clusters in each station of Park 19



Before



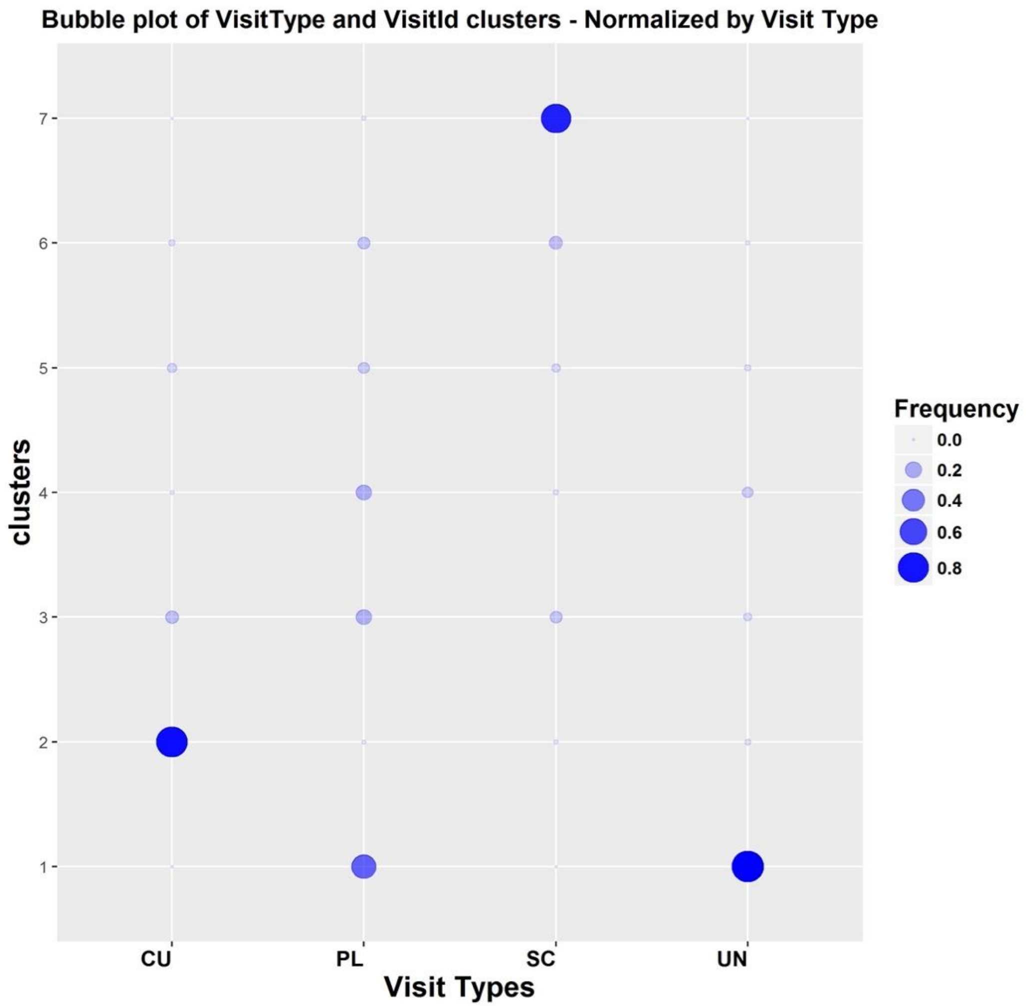
Within



After

Other plots

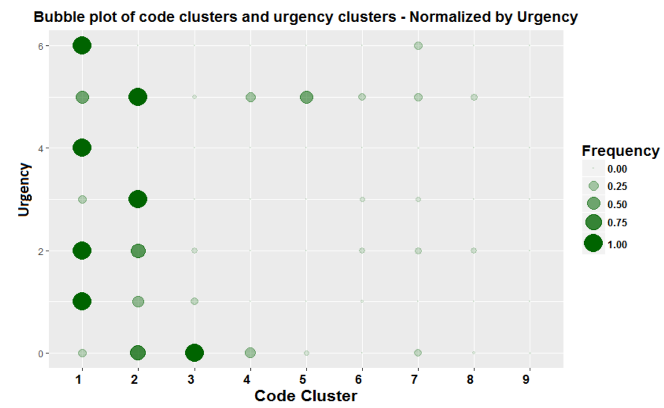
- All is about data aggregation



Clustering Codes

Clustering Codes

- Data Preparation:
  - Quantile based on visits
- Clustering methods:
  - Two level hierarchical clustering



Code	Cluster
1007	3
1014	3
1001	1
1002	6
1003	1
1008	1
1015	1
1017	1
1021	1

Clusters	1	2	3	4	5	6	7	8	9
Members	148	65	16	13	19	47	295	21	18
Frequency in dataset	29032	51935	38992	17538	13218	4974	9920	4250	2070

UCF

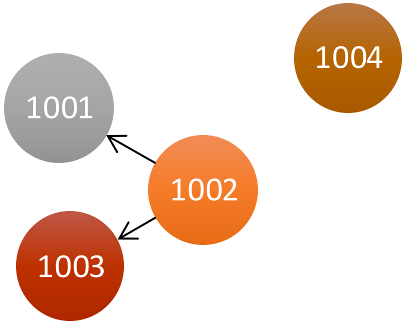
Network Analysis

Why network?

- Practically everything is a graph !
- We can consider each code as a node and a sequence of two codes as an edge between them

VisitId	Date	Error 2
175	03/20/16	1002
175	03/21/16	1001
1002	03/21/16	1002
1002	03/22/16	1003
10	03/23/16	1004

	1001	1002	1003	1004
1001	0	0	0	0
1002	1	0	1	0
1003	0	0	0	0
1004	0	0	0	0

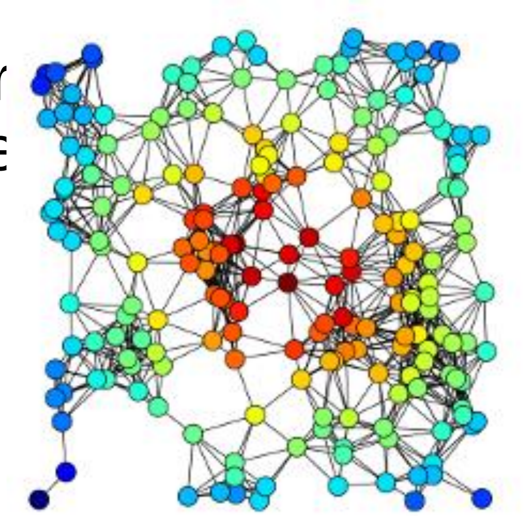


- We constructed tow matrices
  - Frequency based Matrix
  - Time interval based Matrix

UCF

Find the most central codes

- The betweenness is (roughly) defined by the number of shortest paths going through a vertex or an edge,
- when going from a code to another code, is there a code which is frequently visited?
- Node (which are codes here) with higher betweenness centrality would have more control over the network of codes
- In other words most of codes passes through this code



Most central codes

Occurrence	Data	Most Central Codes
Before	Frequency of codes	3130,13902,1001,13,14,8,13900,5122,59,10105
	Average time between code	13902,5112,5111,7,9303,2,1018,14001,64101,18
After	Frequency of code	3130,13902,13,14,1001,5122,14302,13900,10105,18
	Average time between codes	13902,8,13,1001,18,9303,7,2,3130,17027
Within	Frequency of code	1020,1001,1005,13902,1023,7,18,8,13900,2
	Average time between code	8,13902,9,1023,1001,1020,2,7,18,14001

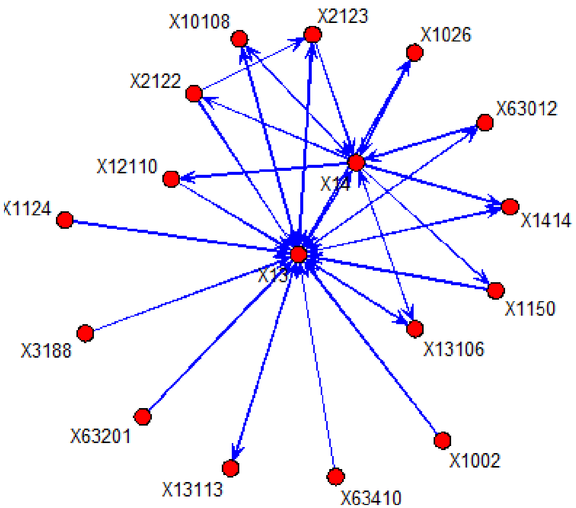
Community detection in graphs

- “There might be some codes which have strong relations with each other than with other groups of code”
- *Infomap* community detection algorithm

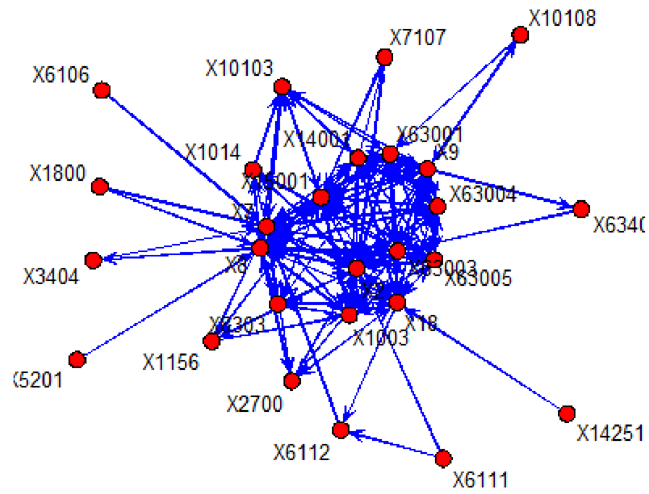
Occurrence	Data	# Communities	# Communities with one member
Before	Frequency of codes	57	3
	Average time between code	59	40
After	Frequency of code	64	21
	Average time between codes	119	80
Within	Frequency of code	54	14
	Frequency of codes	105	75

UCF

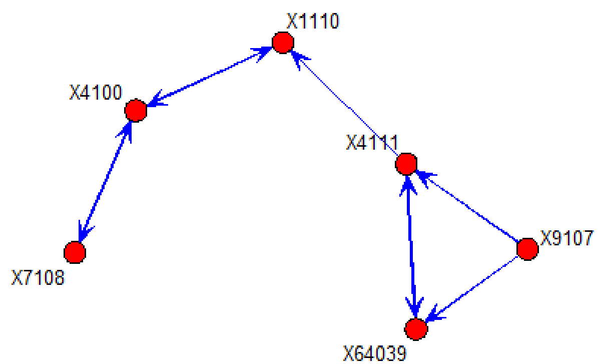
Detected communities



Before - Frequency of codes



Within - Frequency of codes



Before - Average time between codes

Summary

Challenges and Limitations

- Tasks for this competition were more descriptive than predictive so, there was no solid approach to evaluate the results (except domain knowledge expertise)
- There was no exact and detail information about the error codes. If there would be more general classification of each error code, categorizing similar codes and making better interpretations would be easier for analysts.

Executive Summary

- Conditional probability of the codes with respect to the previous state of the code
- Found direct and indirect paths among different types of codes as well as three different timelines
- Categorized similar parks and stations using clustering approaches
- Found different networks of codes which were highly related to each other in terms of frequency and time
- Discovered the probability of co-occurrence of the codes
- Found sequential patterns based on timing of the codes relative to each other

Question?

The End