UCSD Master of Advanced Study Data Science & Engineering Capstone: CalTrans Traffic Analysis

Kevin Dyer
John Gill III
Conway Wong

Advisor: Yoav Freund

June 10, 2016

UC San Diego
Jacobs School of Engineering

Outline

- Business Objective
- Project Execution Overview
  - Data Sources & Acquisition
  - Data Preparation
  - Exploratory Data Analysis
- Analysis and Results
- Demo
- Conclusion
- Future Work

Business Objective

Among traffic congestion in America, California areas ranks among the worst

1. Los Angeles
2. San Francisco
5. San Jose
14. San Diego
24. Riverside
25. Sacramento

Source: http://tomtom.com/en_us/trafficindex/list

Better Insight → Better Solutions

What traffic patterns exist?
What factors have most influence on traffic?

Project Execution Overview

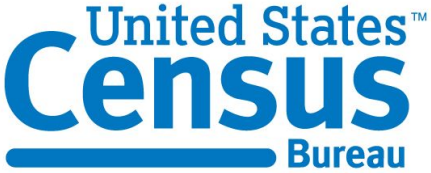
Data Source & Acquisition

Data Source & Acquisition

- Caltrans PeMS
  - 5-minute station readings
  - CHP incidents
  - Station Metadata

- Zillow
  - Monthly Home Value index (ZHVI)

- U.S. Census
  - Population (2010)
  - TIGER/Line Shapefiles



Data Source & Acquisition

1000s of Files & > 100 GB

PeMS

CHP

Python Web Scraper
BeautifulSoup & Mechanize

S3 (Simple Storage Service)

10s of Files & < 100 MB

Zillow

United States Census Bureau

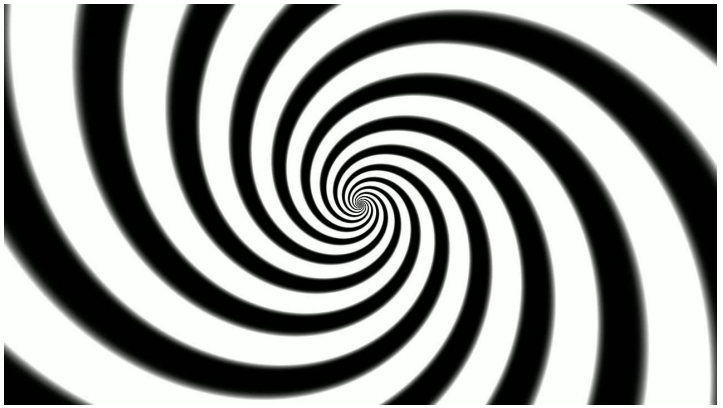
Manual Download

Data Preparation

Data Preparation

- Extract
  - Source Locations
    - S3
    - Websites
  - Source Formats
    - CSV
    - Shape Files
    - JSON

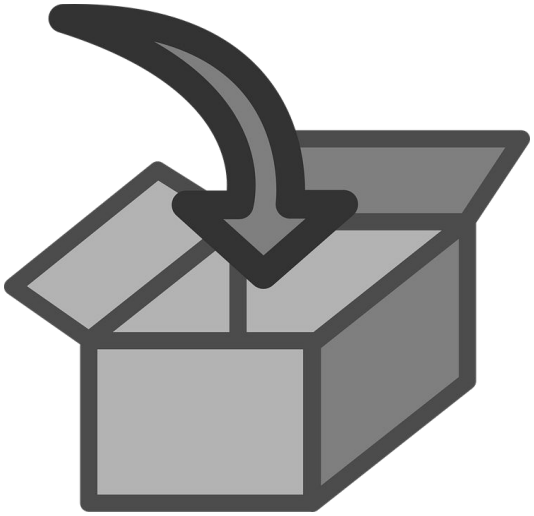
Data Preparation



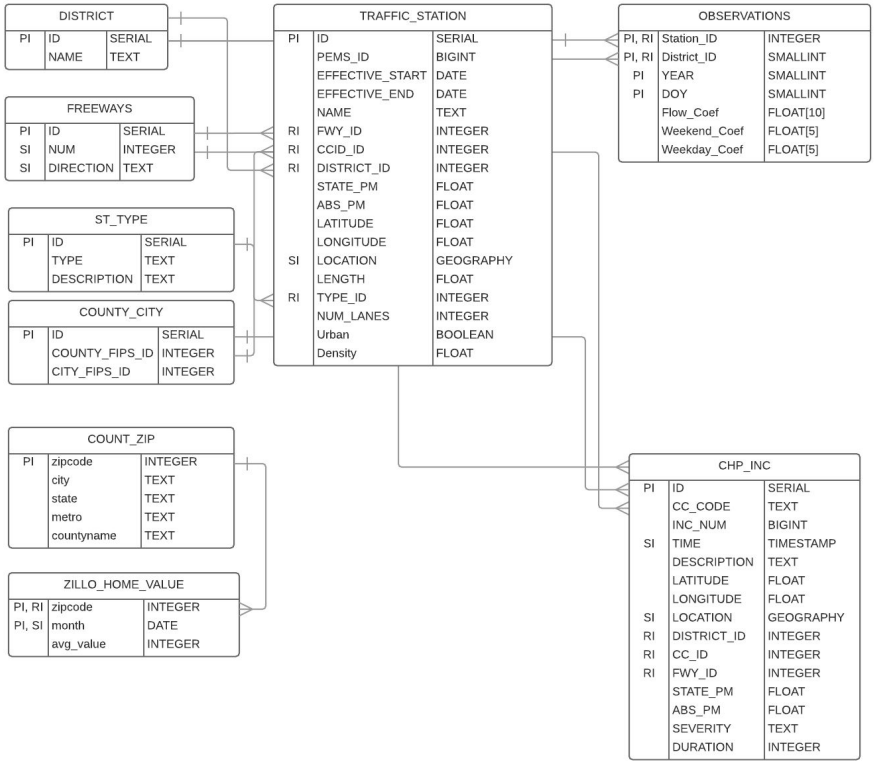
- Transform
  - Missing Fields
  - Bad Data Dropped
  - Time
    - Different Granularity per source
    - Effectivity Dates
  - Sensor Data Pivoted

Station Id	District Id	Year	Day of Year	Day of Week	Total Flow (288)	Avg. Occupancy (288)	Avg. Speed (288)

Data Preparation



- Load
  - Data Warehouse - Postgres
  - Target Schema - Snowflake



Exploratory Data Analysis

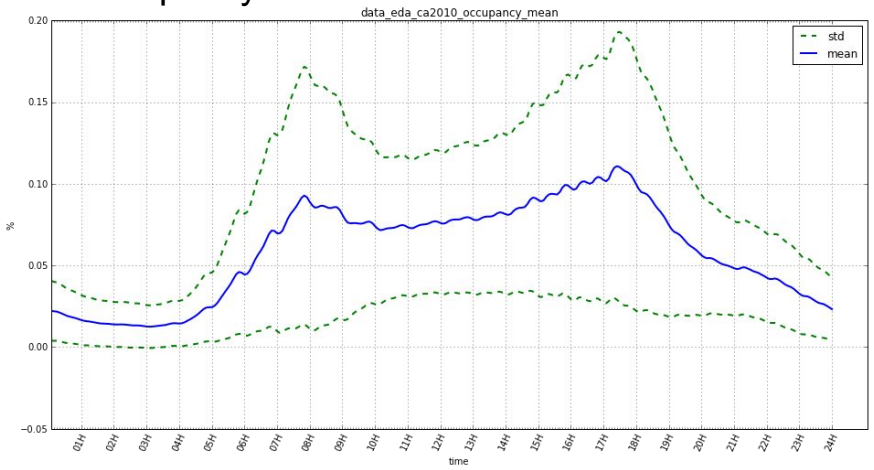
Exploratory Data Analysis

- Mean + Standard Deviation
- Top 4 Eigenvectors
- Scatter Density Plot (Total Flow)

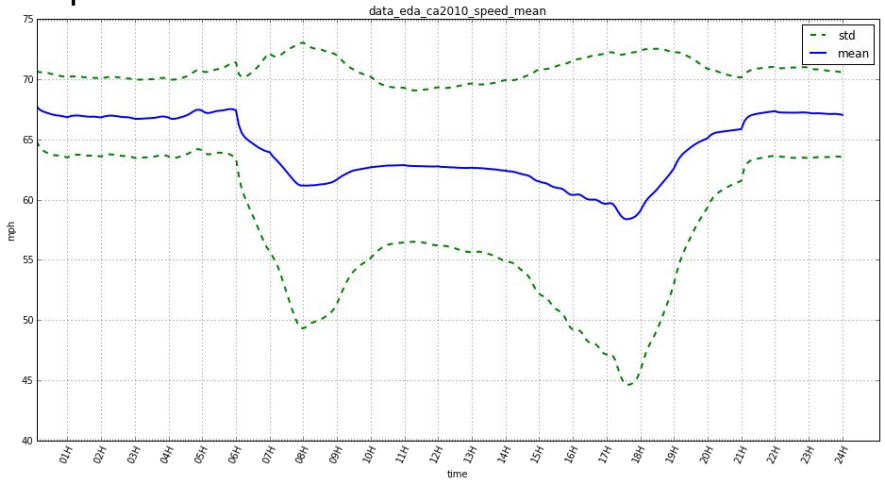
Exploratory Data Analysis

Total Flow

Occupancy

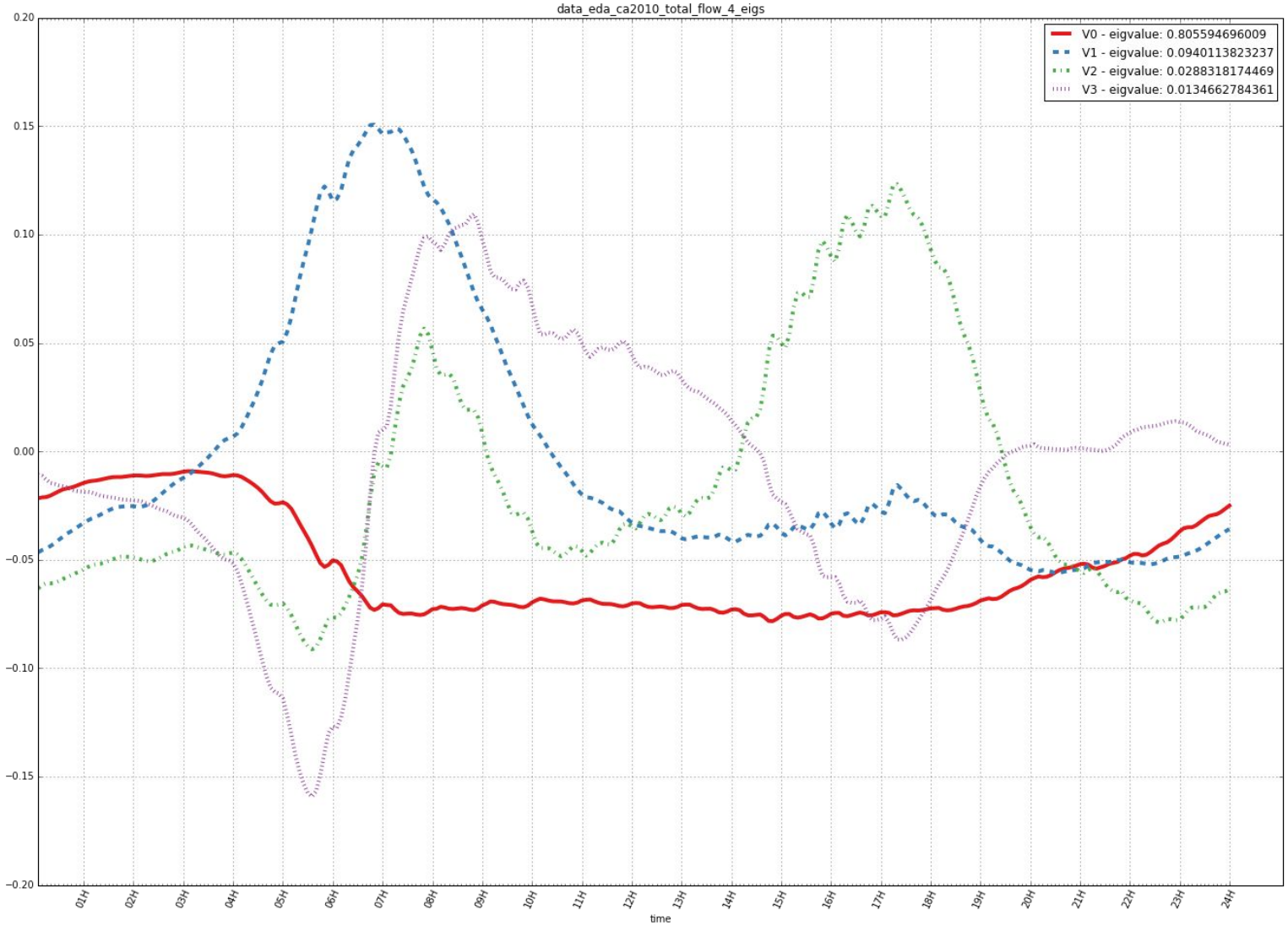


Speed

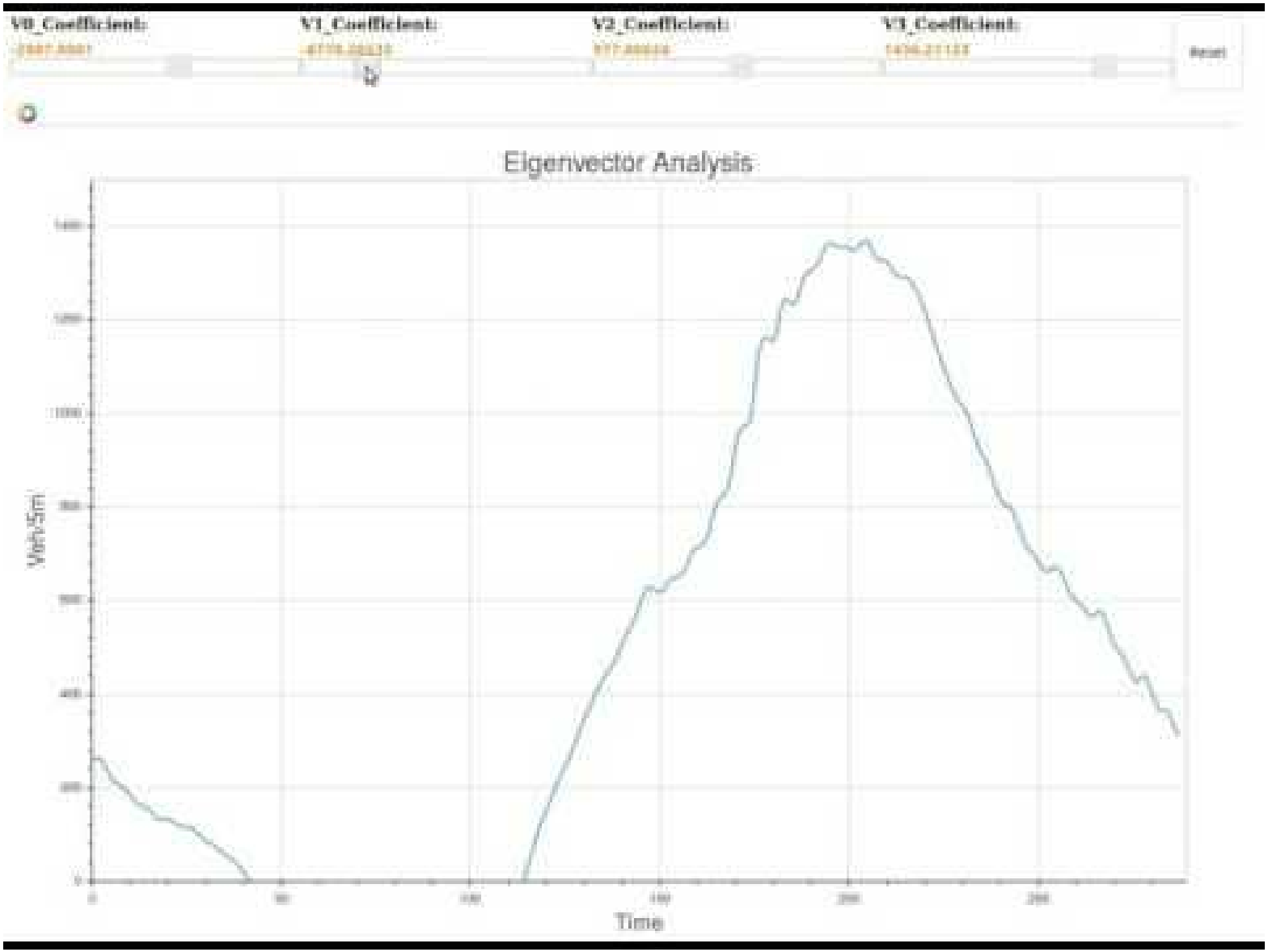


- Total Flow - measures traffic volume
- Occupancy - measures vehicle occupancy across lanes
- Speed - measures velocity of vehicles across lanes
- Preliminary analysis:
  - Total flow - directly proportional to occupancy
  - Total flow - inversely proportional to speed
  - Traffic volume - high between the hours of 6AM and 9PM

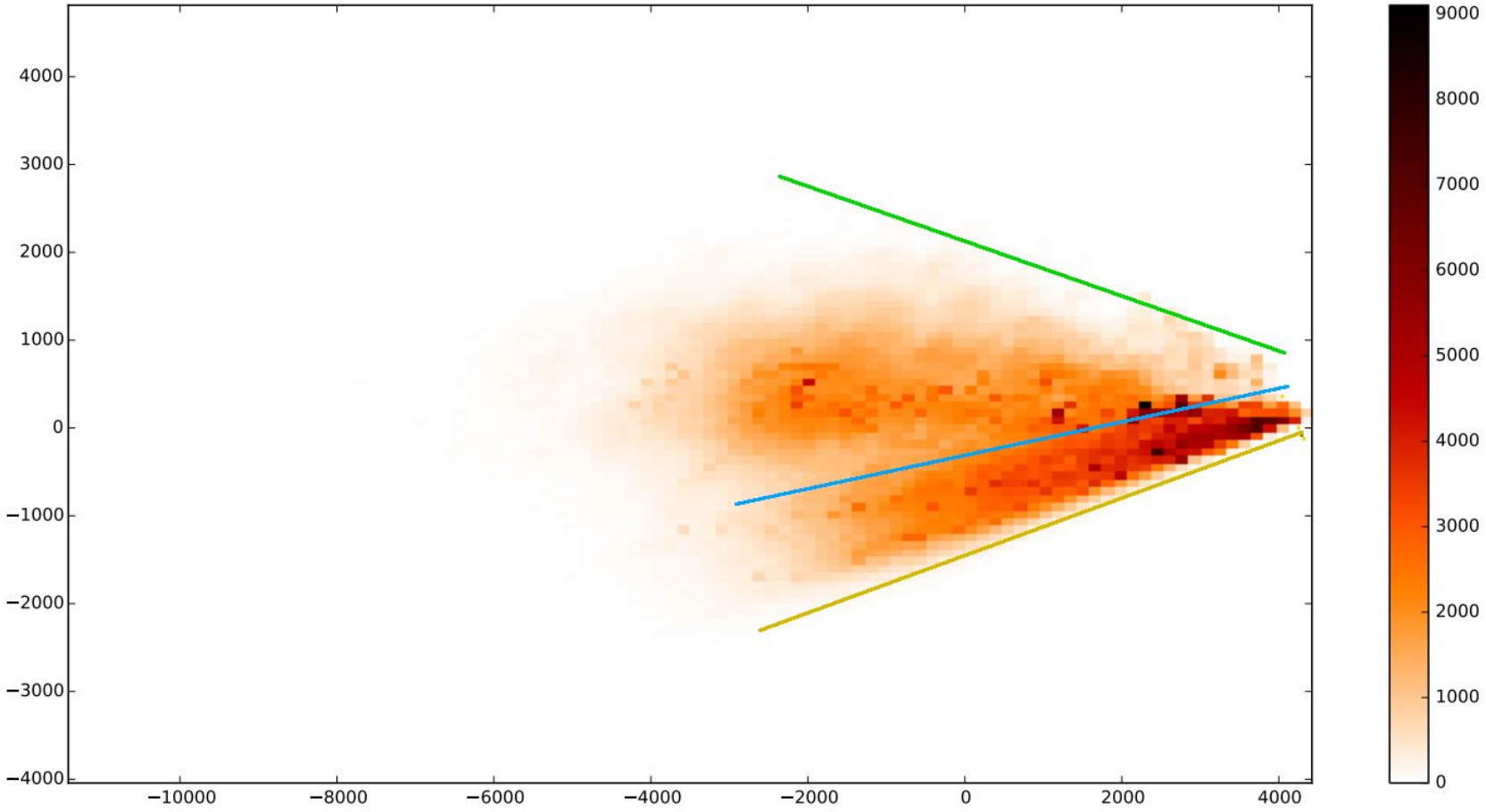
Exploratory Data Analysis



Exploratory Data Analysis



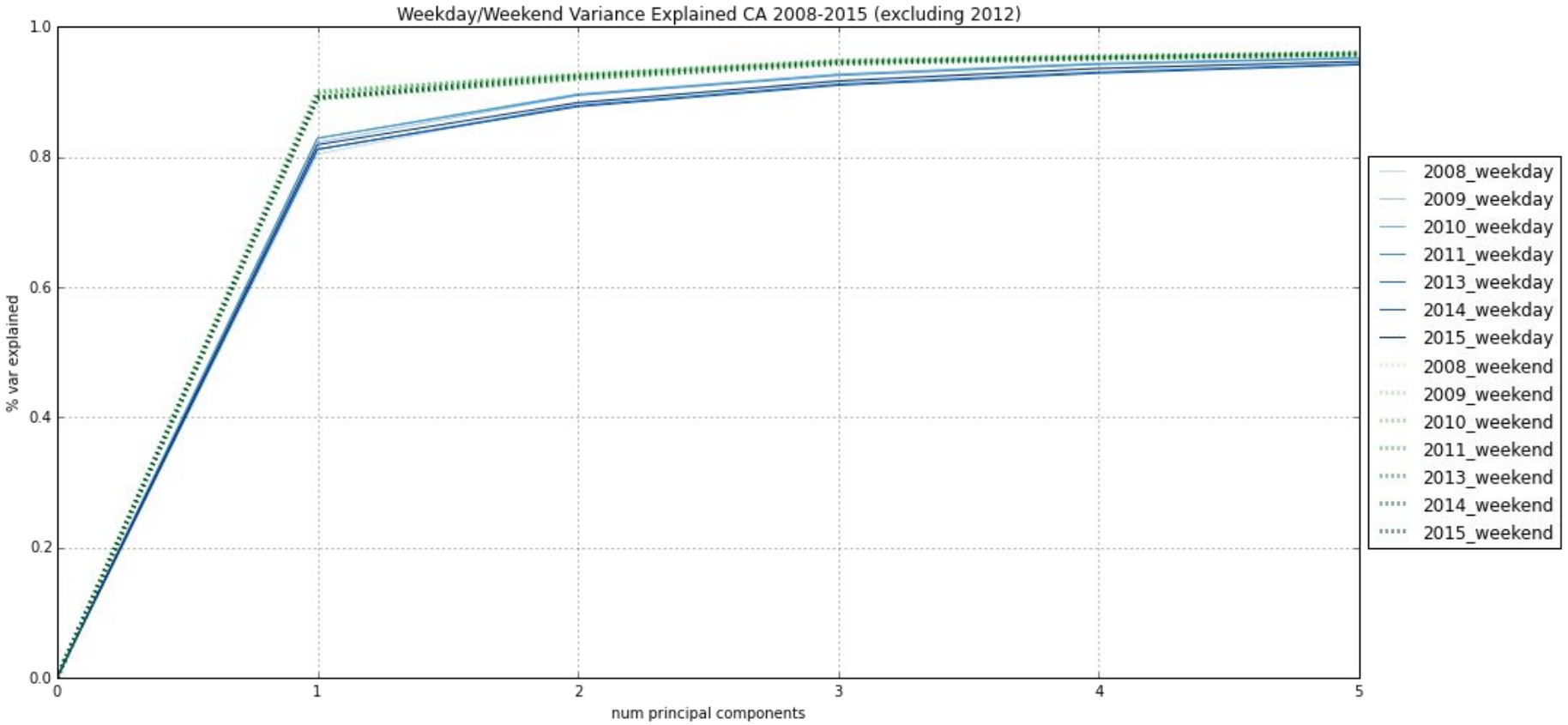
Exploratory Data Analysis



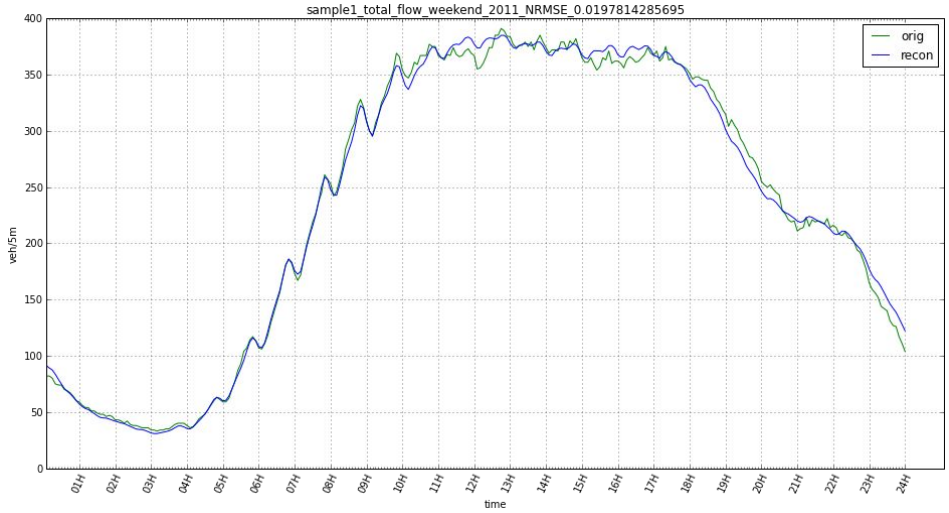
Analysis and Results

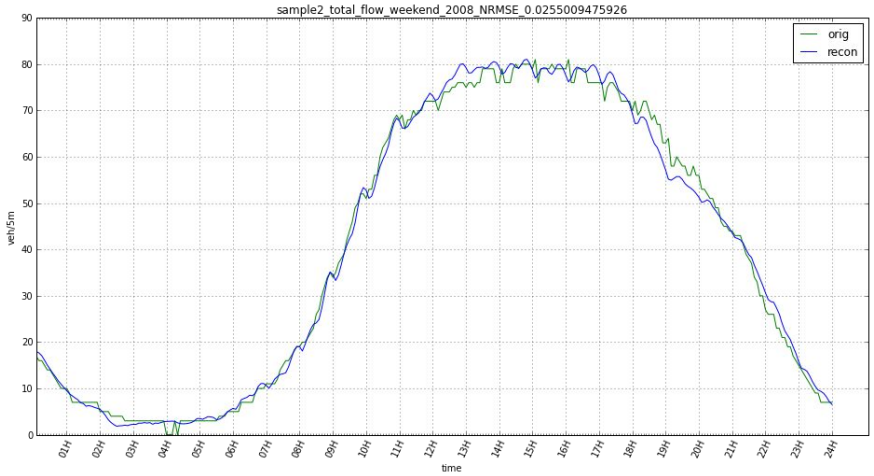
PCA

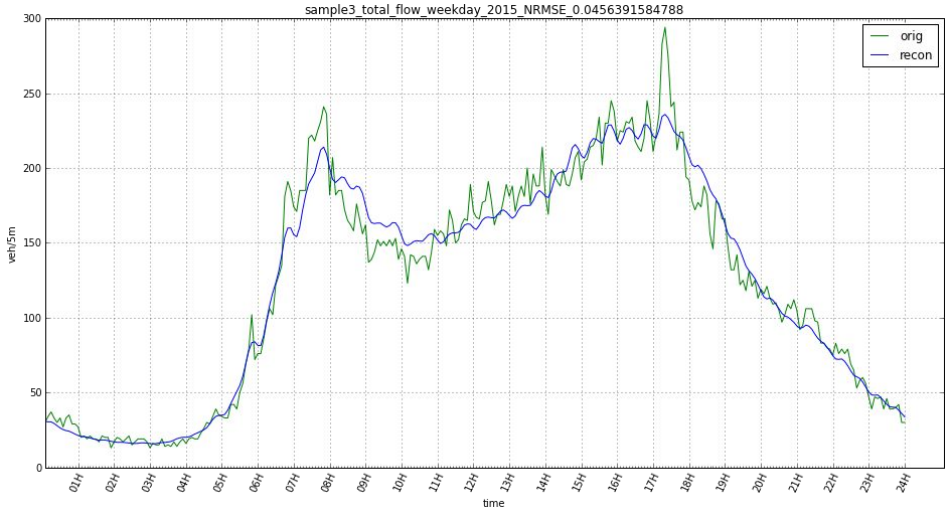
- Traffic Pattern Modeling and Prediction
- Assessment Criteria: NRMSE



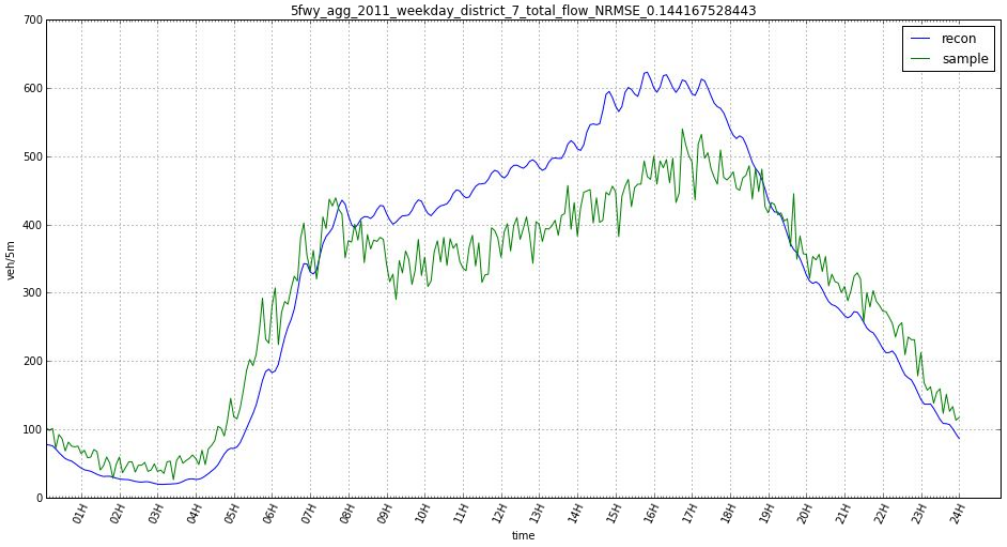
- 5 eigenvectors
  - Over 90% variance explained
  - Sufficient in modeling total flow

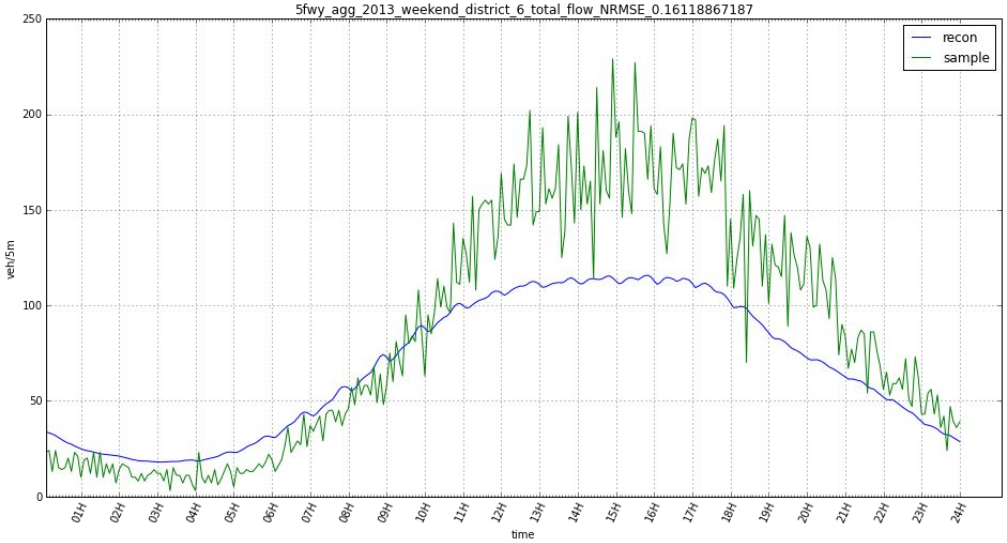






- Execution
  - station+day combination
  - Random samples taken from station+day combinations
  - Projection and Reconstruction using top 5 eigenvectors
- Low NRMSE → Model Total Flow w/ Sufficient Accuracy



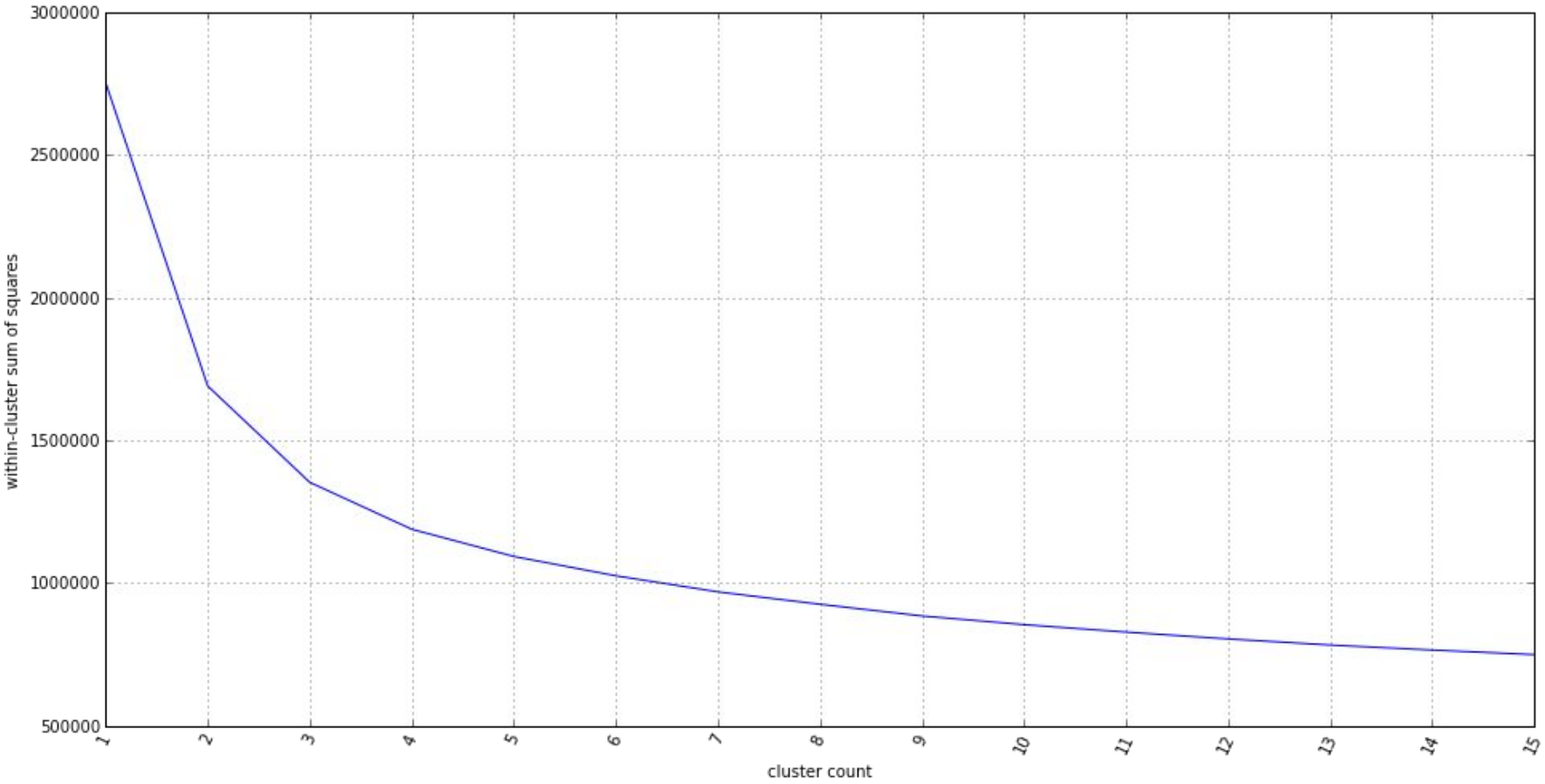


- Execution
  - station+day combination grouped by district+freeway; aggregated across 5 Fwy to obtain 5 Fwy eigenvector
  - Random samples taken from station+day combinations from 5 Fwy
  - Projection and Reconstruction using 5 Fwy eigenvector

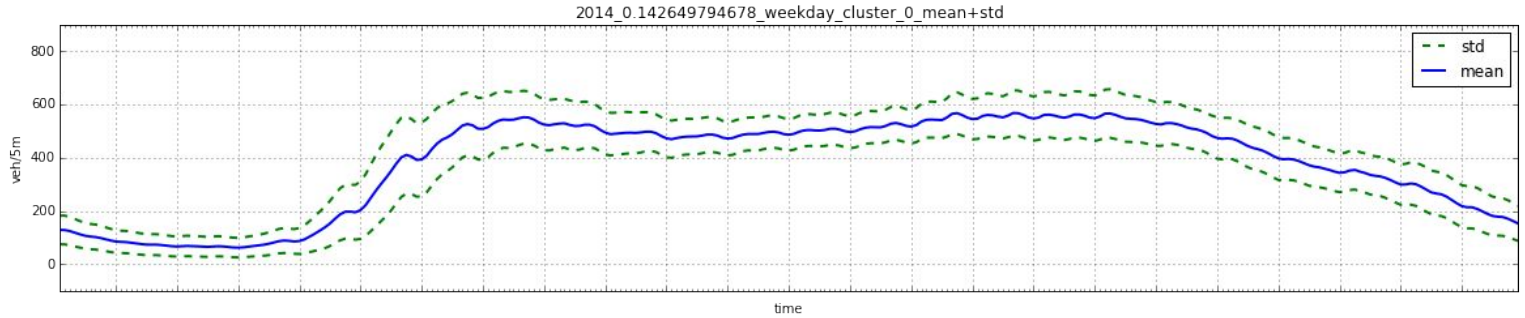
- Reasonable NRMSE → Model Total Flow for a Highway w/ Sufficient Accuracy

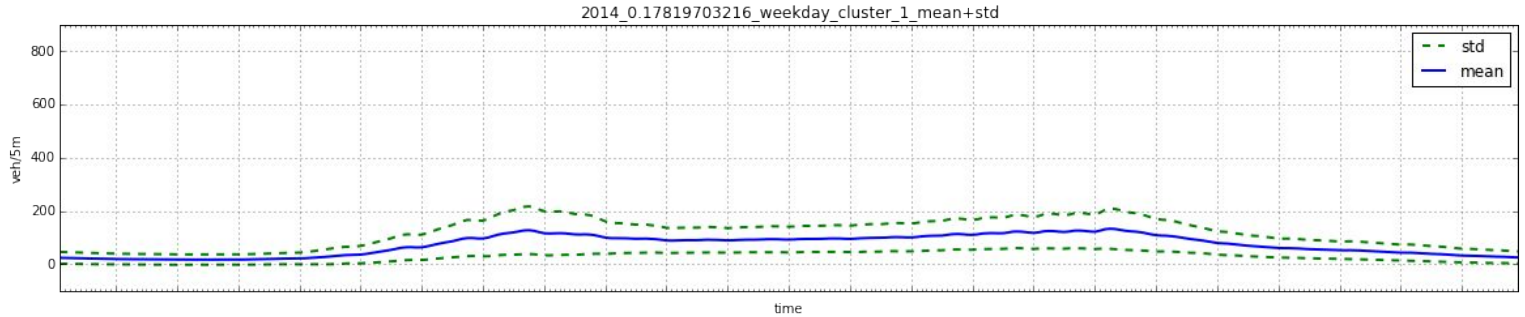
KMeans++

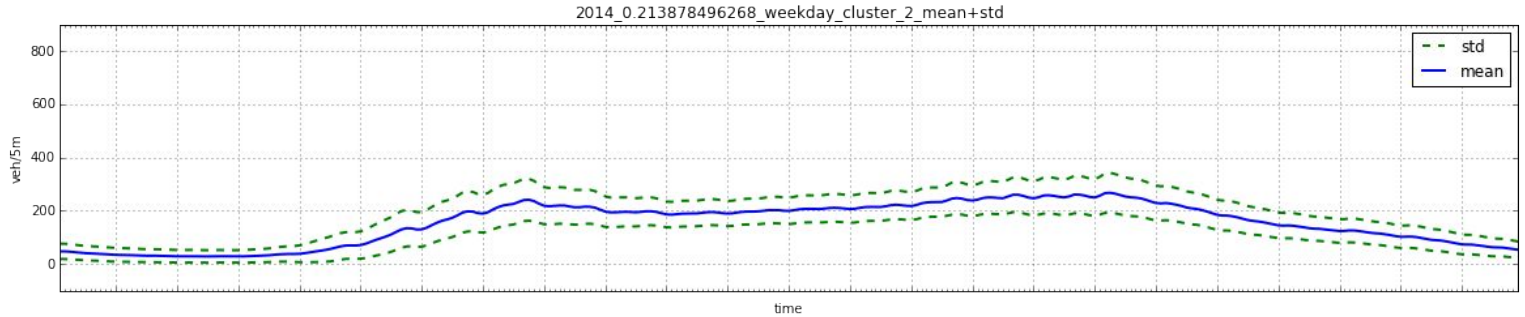
- Traffic Pattern Behavior Clustering
- Assessment Criteria: None

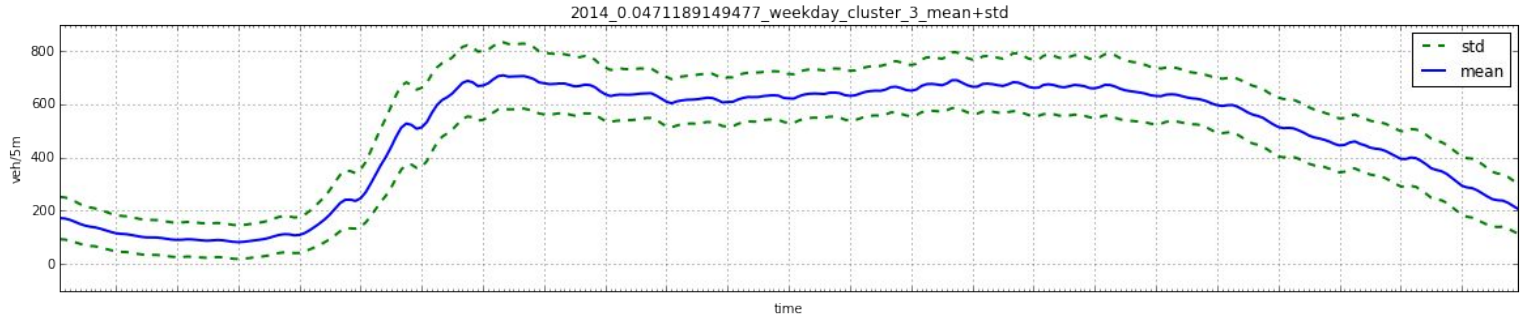


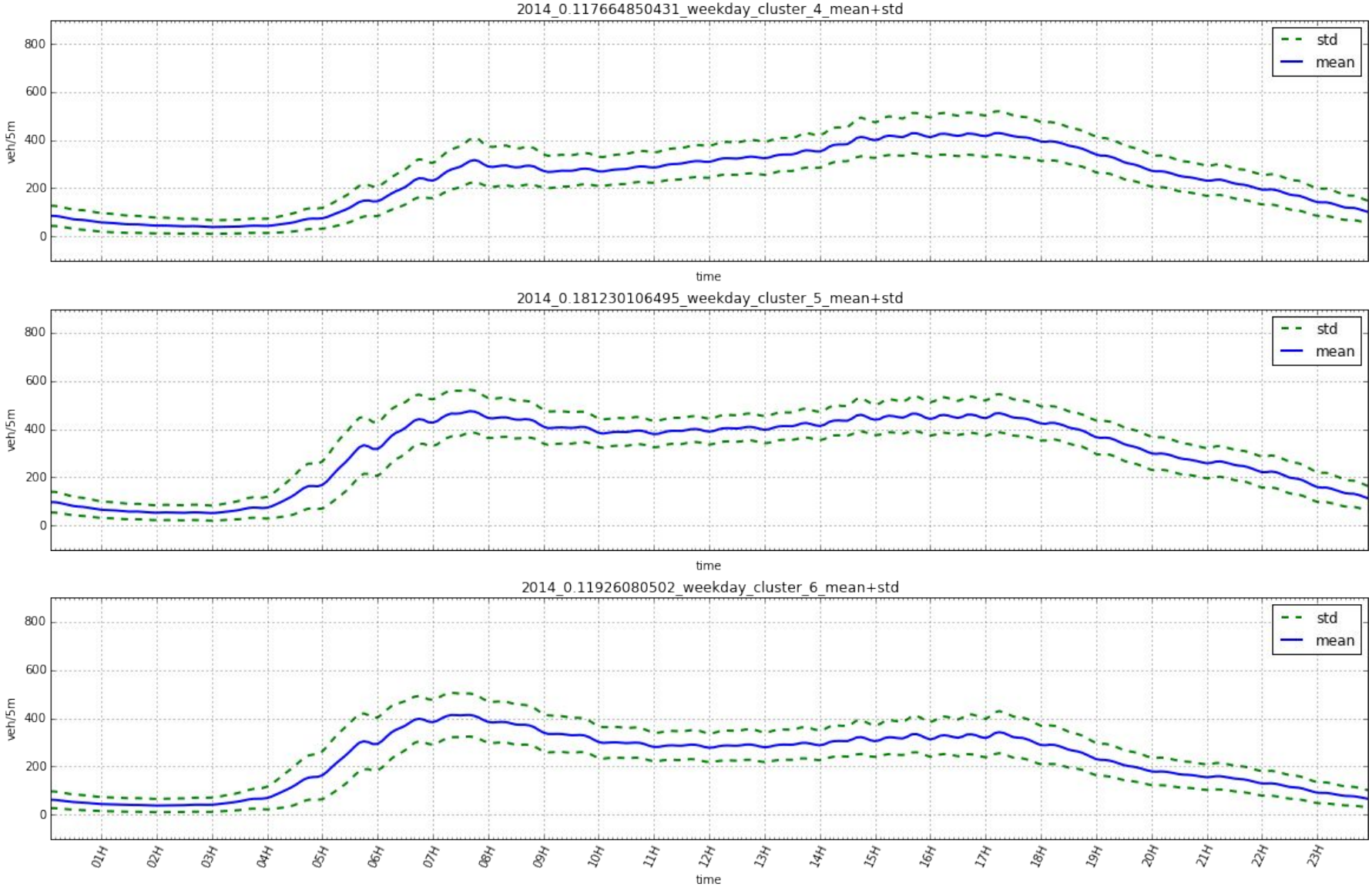
- Using CA 2014 as a sample, determine optimal cluster count
  - Identify “elbow” in within-cluster sum of squares plot → about 7 clusters


2014_0.142649794678_weekday_cluster_0_mean+std
std
mean
veh/5m
time

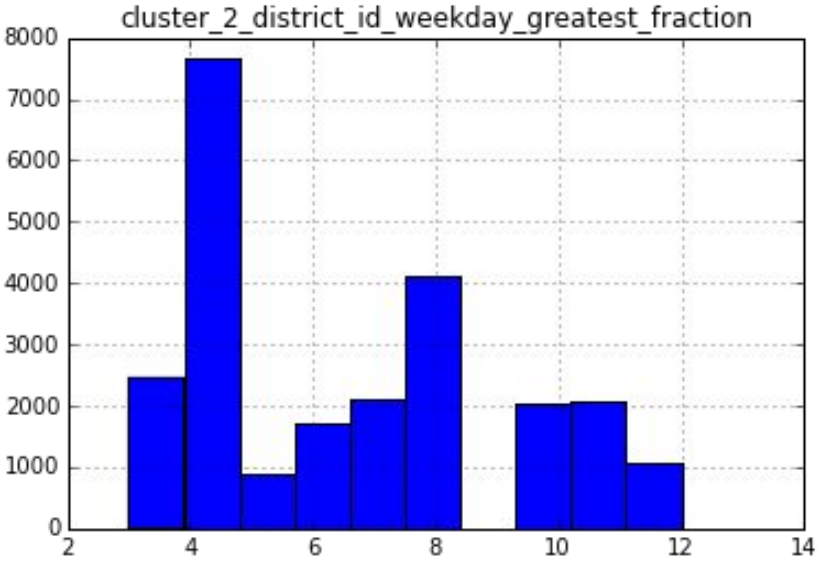

2014_0.17819703216_weekday_cluster_1_mean+std
std
mean
veh/5m
time

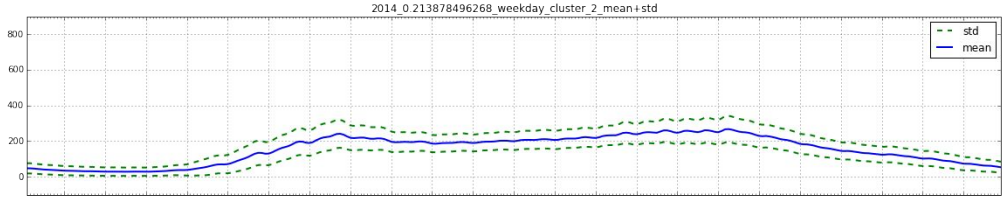

2014_0.213878496268_weekday_cluster_2_mean+std
std
mean
veh/5m
time

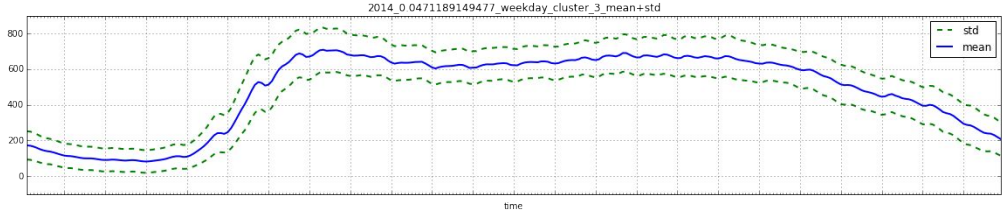

2014_0.0471189149477_weekday_cluster_3_mean+std
std
mean
veh/5m
time

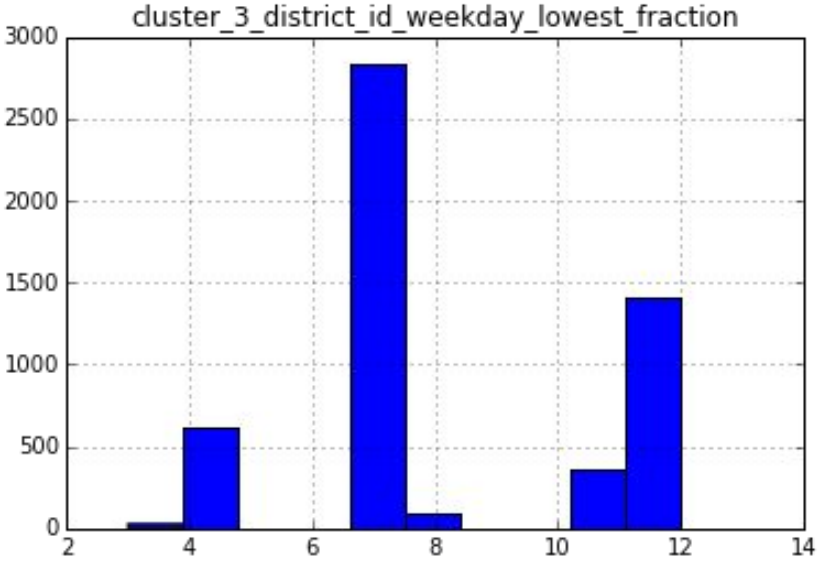

2014_0.117664850431_weekday_cluster_4_mean+std
std
mean
veh/5m
time
2014_0.181230106495_weekday_cluster_5_mean+std
std
mean
veh/5m
time
2014_0.11926080502_weekday_cluster_6_mean+std
std
mean
veh/5m
time
0
200
400
600
800
01H
02H
03H
04H
05H
06H
07H
08H
09H
10H
11H
12H
13H
14H
15H
16H
17H
18H
19H
20H
21H
22H
23H

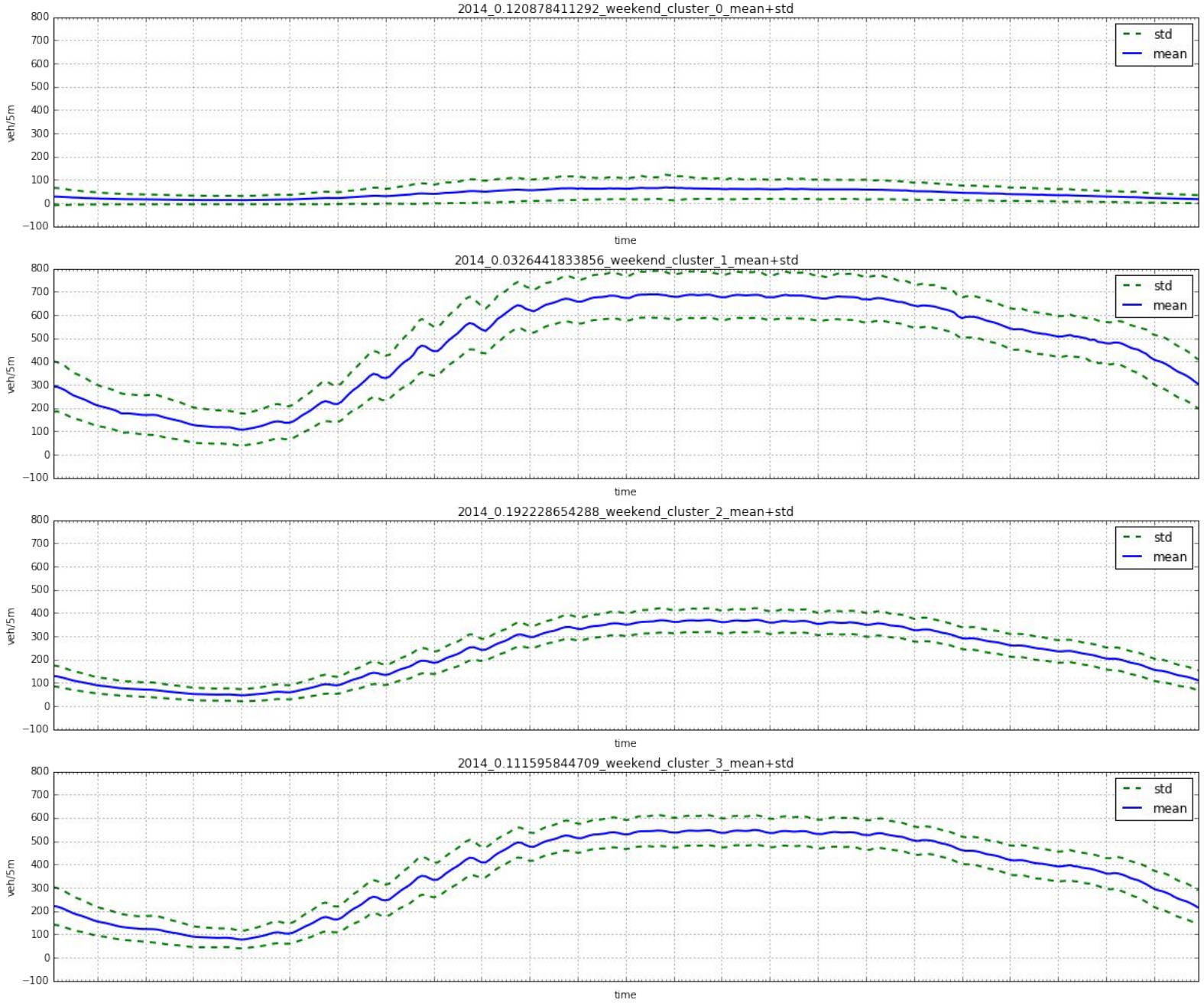

DN
Siskiyou
Modoc
2
Trinity
Shasta
Lassen
Hum
1
Tehama
Plumas
Men
Glenn
Butte
Sierra
Lake
Col
Placer
3
El Dorado
Yolo
Nap
Son
Sac
Alp
Sol
Mrn
4
CC
Cal
San Joaquin
Tuolumne
10
Mono
SF
Ala
Stan
Mariposa
San Mateo
SCl
Merced
Santa Cruz
Madera
Fresno
9
Inyo
Tulare
Kin
6
Mon
5
San Luis Obispo
Kern
Santa Barbara
Ven
Los Angeles
7
San Bernardino
8
Riverside
Ora
12
San Diego
11
Imperial

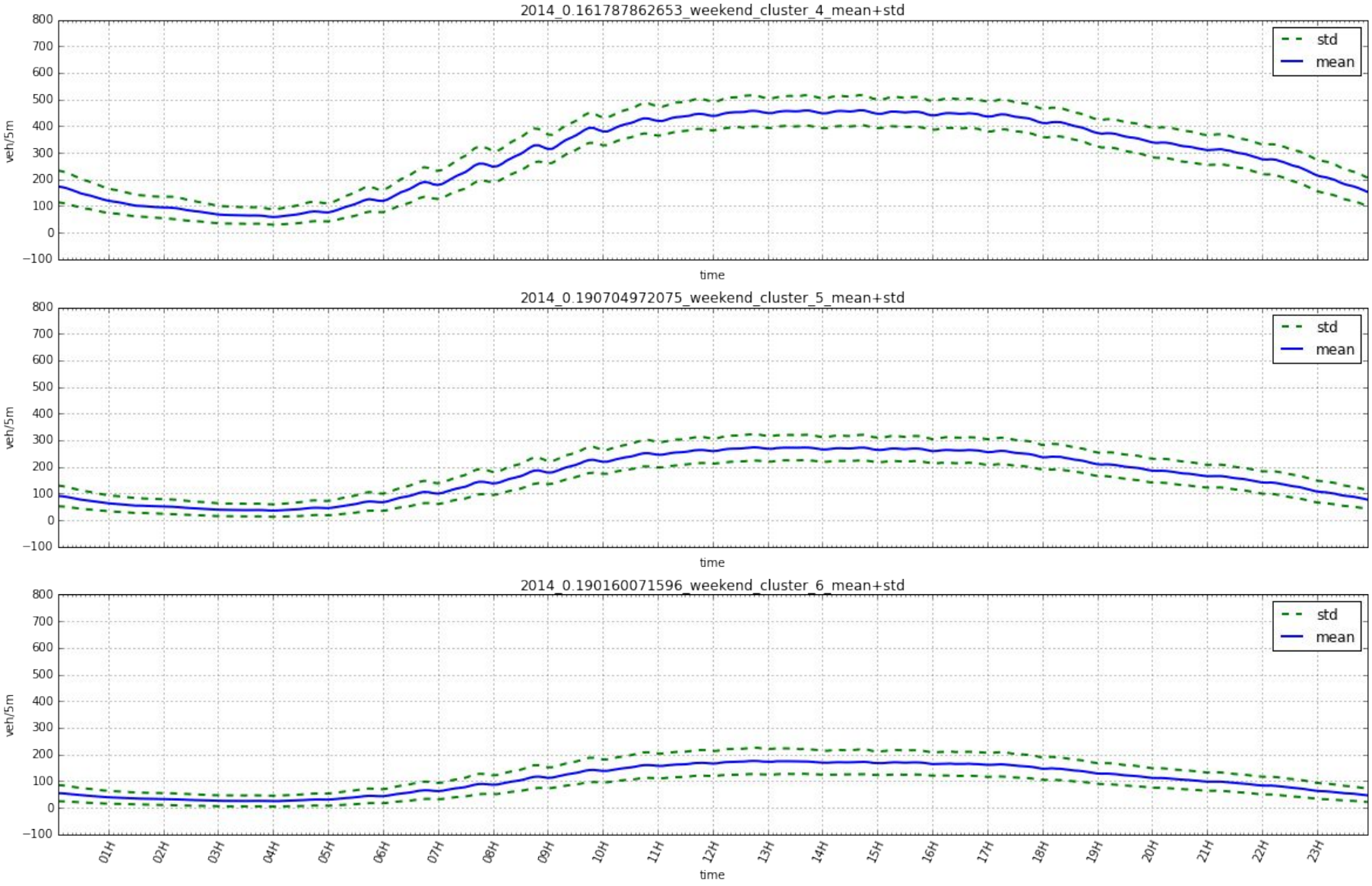

cluster_2_district_id_weekday_greatest_fraction
8000
7000
6000
5000
4000
3000
2000
1000
0
2
4
6
8
10
12
14


2014_0.213878496268_weekday_cluster_2_mean+std
std
mean
800
600
400
200
0
time

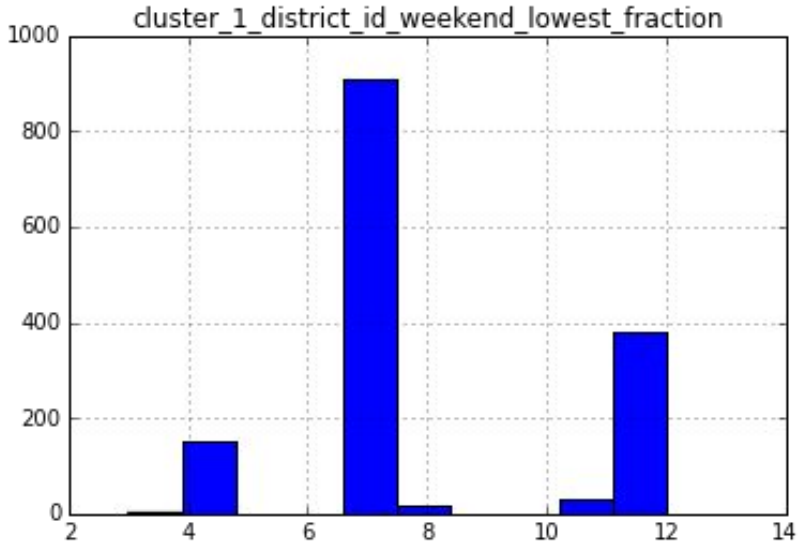

2014_0.0471189149477_weekday_cluster_3_mean+std
std
mean
800
600
400
200
0
time

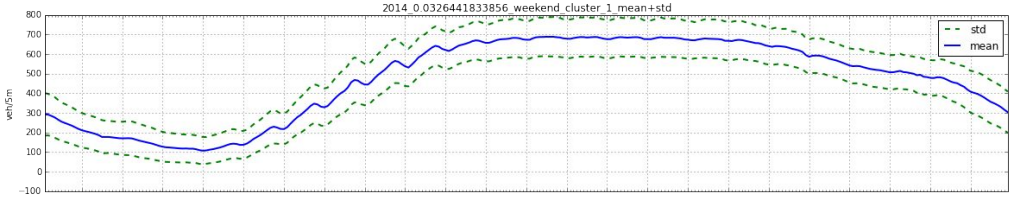

cluster_3_district_id_weekday_lowest_fraction
3000
2500
2000
1500
1000
500
0
2
4
6
8
10
12
14

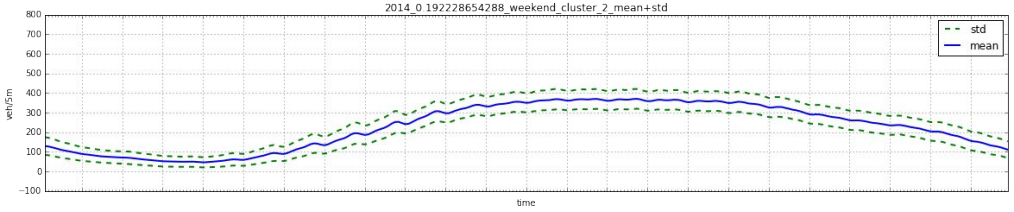

2014_0.120878411292_weekend_cluster_0_mean+std
2014_0.0326441833856_weekend_cluster_1_mean+std
2014_0.192228654288_weekend_cluster_2_mean+std
2014_0.111595844709_weekend_cluster_3_mean+std
std
mean
veh/5m
time
800
700
600
500
400
300
200
100
0
−100

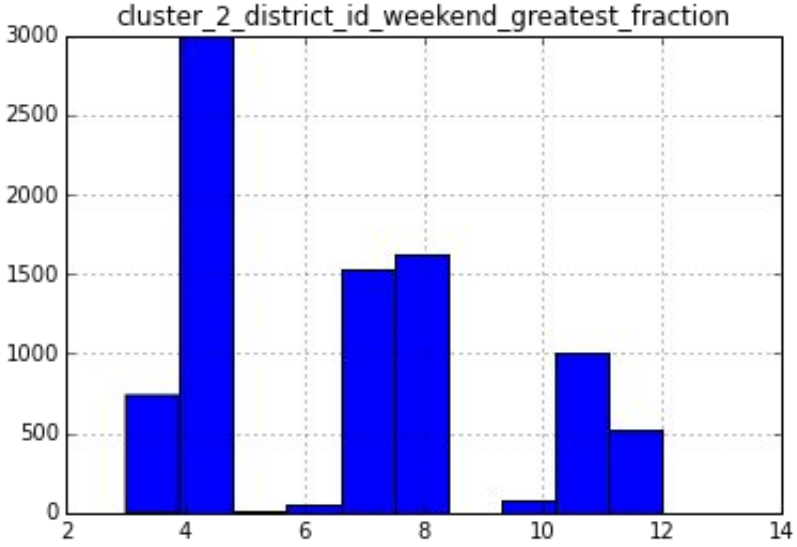

2014_0.161787862653_weekend_cluster_4_mean+std
std
mean
veh/5m
time
2014_0.190704972075_weekend_cluster_5_mean+std
std
mean
veh/5m
time
2014_0.190160071596_weekend_cluster_6_mean+std
std
mean
veh/5m
01H
02H
03H
04H
05H
06H
07H
08H
09H
10H
11H
12H
13H
14H
15H
16H
17H
18H
19H
20H
21H
22H
23H
time


DN
Siskiyou
Modoc
2
Trinity
Shasta
Lassen
Hum
Tehama
Plumas
1
Men
Glenn
Butte
Sierra
Lake
Col
Placer
3
El Dorado
Yolo
Sac
Alp
4
Cal
10
Tuolumne
Mono
CC
SF
Ala
San Mateo
SCl
Mariposa
9
Merced
Madera
Santa Cruz
Fresno
Inyo
Tulare
Mon
Kin
6
5
San
Luis
Obispo
Kern
San Bernardino
Santa
Barbara
Ven
Los
Angeles
8
7
Riverside
Ora
12
San
Diego
11
Imperial


cluster_1_district_id_weekend_lowest_fraction
1000
800
600
400
200
0
2
4
6
8
10
12
14


2014_0.0326441833856_weekend_cluster_1_mean+std
std
mean
veh/5m
time

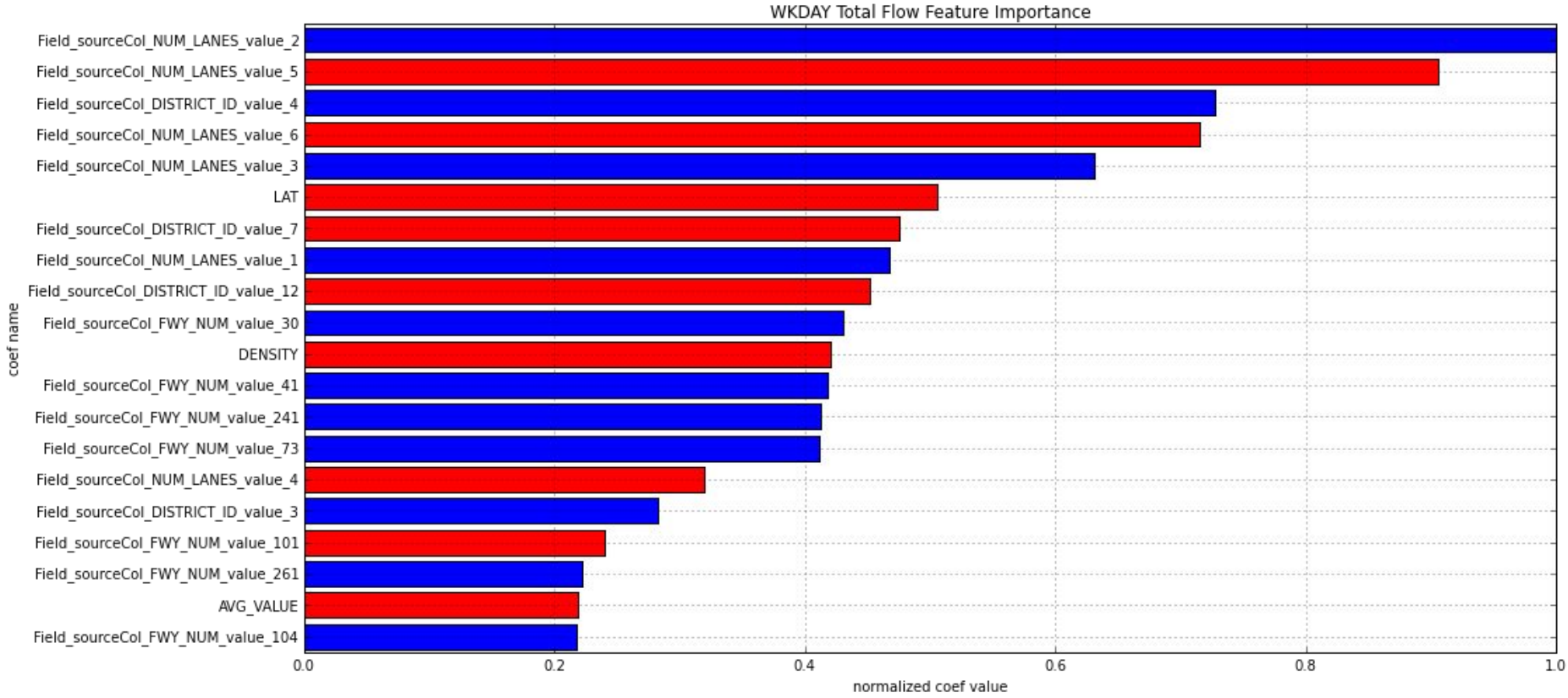

2014_0.192228654288_weekend_cluster_2_mean+std
std
mean
veh/5m
time


cluster_2_district_id_weekend_greatest_fraction
3000
2500
2000
1500
1000
500
0
2
4
6
8
10
12
14

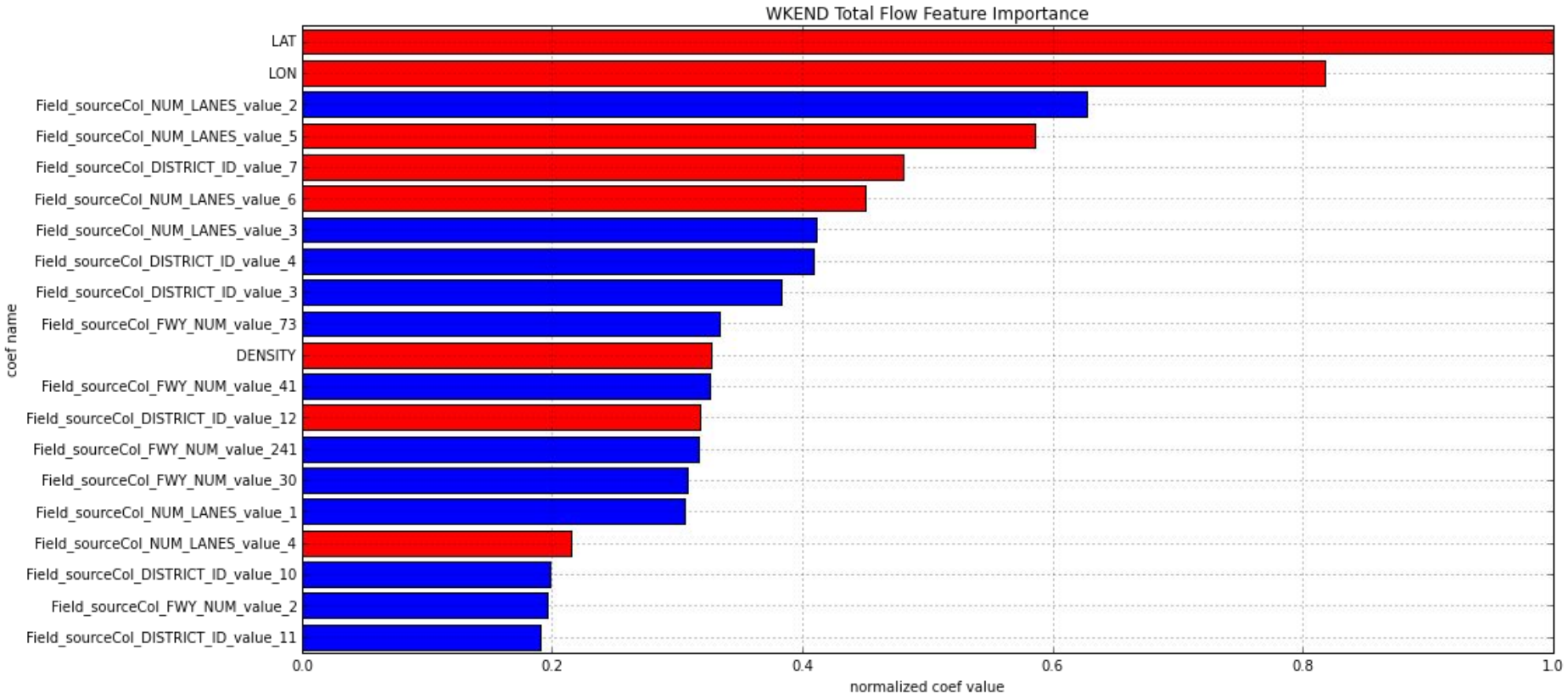
Elastic Net Regression

- Influential Traffic Factors
- Assessment Criteria: None

Weekday Partition



Weekend Partition



Demo

Analysis

- Traffic GIS Visualization
  - Display each PeMS traffic station on a map using its latitude/longitude
  - Load yearly (2008-2015) traffic volume data sets
  - Color-encode traffic stations using its V0 and V1 eigenvector coefficients via a diverging color scheme
  - Filter traffic stations by direction, freeway number, or coefficient value
  - Find traffic stations by ID and zoom to station
  - Reconstruct traffic volume readings of a station for any day

Traffic GIS

Conclusion

Better Insight → Better Solutions

What traffic patterns exist?
What factors have most influence on traffic?

Future Work

- Expand External Data Sources → Influential factors
- Modeling and Analysis of other CalTrans PeMS features
  - Occupancy
  - Speed
- Expansion on GIS visualization → Directional Arrows along freeway: North+South, East+West
- Effect distance of CHP incident has on Traffic Flow
- Outlier Detection
  - KMeans Clustering
  - Mahalanobis Distance

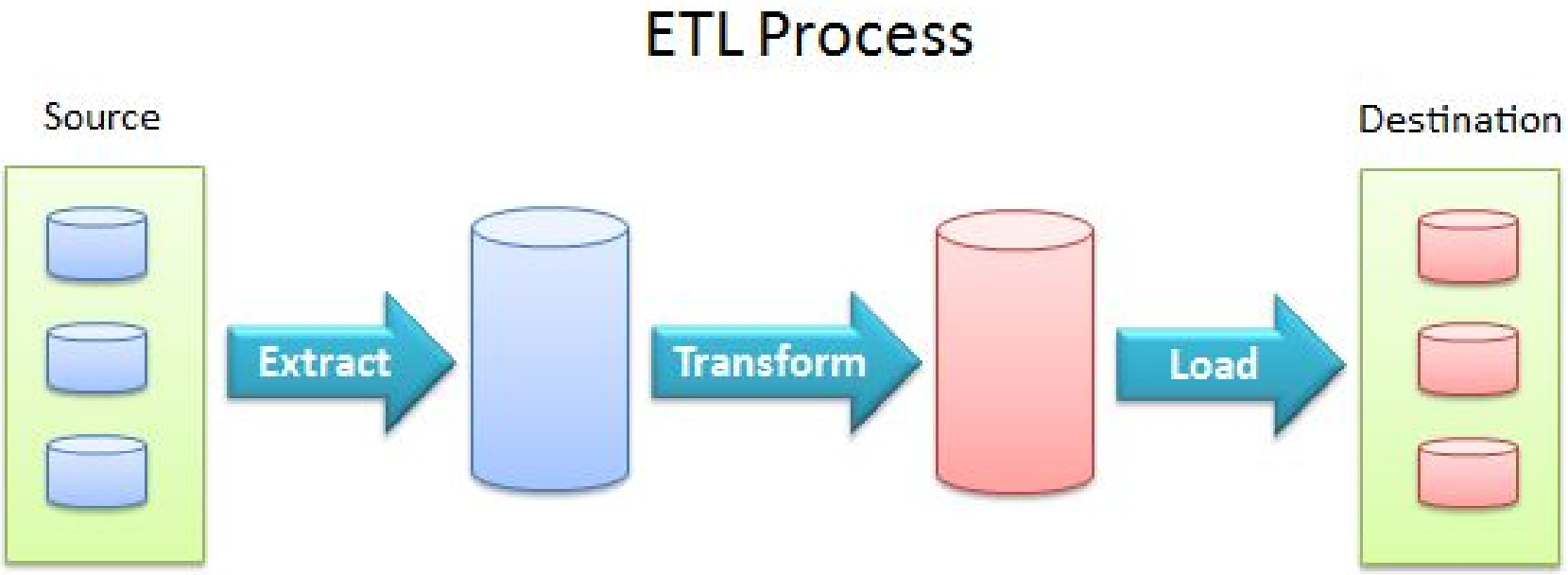
Acknowledgements

- Yoav Freund
- Friends
- Family
  - Girlfriend
  - Fiancee
  - Wife
- Kevin Coakley
- Ilkay Altintas

Questions?

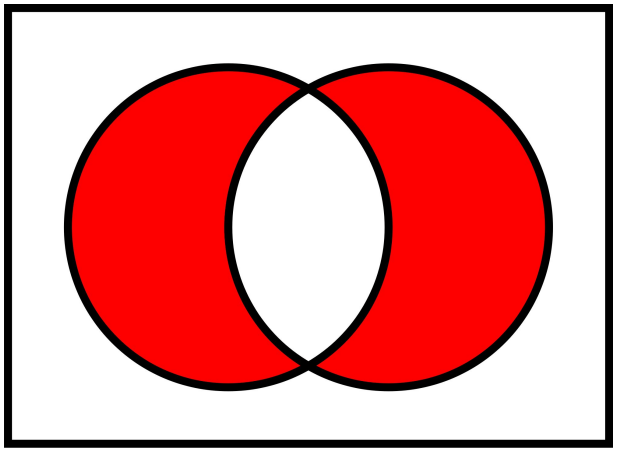
Backup Slides

Data Preparation



Query / Join

- Temporal Data
- Geospatial Data
- Multiway Joins



```
SELECT t.ID, t.Num_Lanes, t.Length, t.Urban, t.Density,
       f.Num,
       CASE f.Direction WHEN 'N' THEN 1 WHEN 'E' THEN 3 WHEN 'S' THEN 2 WHEN 'W' THEN 4 ELSE -1 END,
       z.Avg_Value,
       CASE WHEN chp.ID IS NULL THEN 'F' ELSE 'T' END,
       CAST(chp.CC_CODE AS CHAR(4)), CAST(chp.Description AS CHAR(78)), chp.Duration,
       YearDOYToDate(o.year, o.DOY),
       o.Flow_Coef[1], o.Flow_Coef[2], o.Flow_Coef[3], o.Flow_Coef[4], o.Flow_Coef[5],
       o.Flow_Coef[6], o.Flow_Coef[7], o.Flow_Coef[8], o.Flow_Coef[9], o.Flow_Coef[10]
FROM Observations o
        INNER JOIN Traffic_Station t ON (o.Station_ID=t.ID)
        INNER JOIN ST_Type st ON (t.Type_ID=st.ID AND st.type='ML')
        INNER JOIN Freeways f ON (f.ID=t.Fwy_ID)
        LEFT OUTER JOIN Zillo_Home_Value z ON
            ((EXTRACT(YEAR FROM z.month)=o.year) AND
             EXTRACT(MONTH FROM z.month)=EXTRACT(MONTH FROM YearDOYToDate(o.year, o.DOY)))
        INNER JOIN County_Zip cz ON (t.ZIPCODE=cz.ZIPCODE AND cz.ZIPCODE=z.ZIPCODE)
        LEFT OUTER JOIN CHP_INC chp ON (
            CAST(chp.time AS DATE)=YearDOYToDate(o.year, o.DOY)
            AND chp.Fwy_ID=t.Fwy_ID
            AND ST_Distance(chp.Location, t.Location) < 804.672 -- Half-Mile away
        )
WHERE o.year={y}
ORDER BY 1, 13;
```