Cluster Analysis for I-35 IAJR Using Permanent Traffic Count Station Volume Data

A Case Study

Background

- Microscopic traffic simulation modeling integral part of TxDOT’s Interchange Access Justification Report (IAJR) process
- Process established in order to obtain Federal approval
- Traditional approach assumes a “representative day” as basis for data collection, model development, and analysis

Questions

- What constitutes a "representative day"?
- What are "typical" traffic conditions?
- Should decision making consider only typical traffic conditions on representative day(s)?
- What to do about widely available, more comprehensive sources for time-dynamic data and how should they be integrated into the process?

FHWA Traffic Analysis Toolbox: Volume III

- “TAT3” Definitive reference for development and calibration of simulation models
- Originally published in 2004
- 2019 Update High-Priority Focus Areas
  - Fully Integrate Time-Dynamic Representation of Congestion
  - Require Better Representation of Recurrent and Non-Recurrent Conditions
  - Remove Subjective Calibration Criteria
  - Emphasize Accurate Bottleneck Modeling

What does TAT3 change?

- Expands upon need to consider more data over longer period of time
- Identify representative days for which models can be developed and calibrated
- Underlying assumption: *Expanding window of time and traffic conditions for modeling and analysis yields better decision making*

More Questions

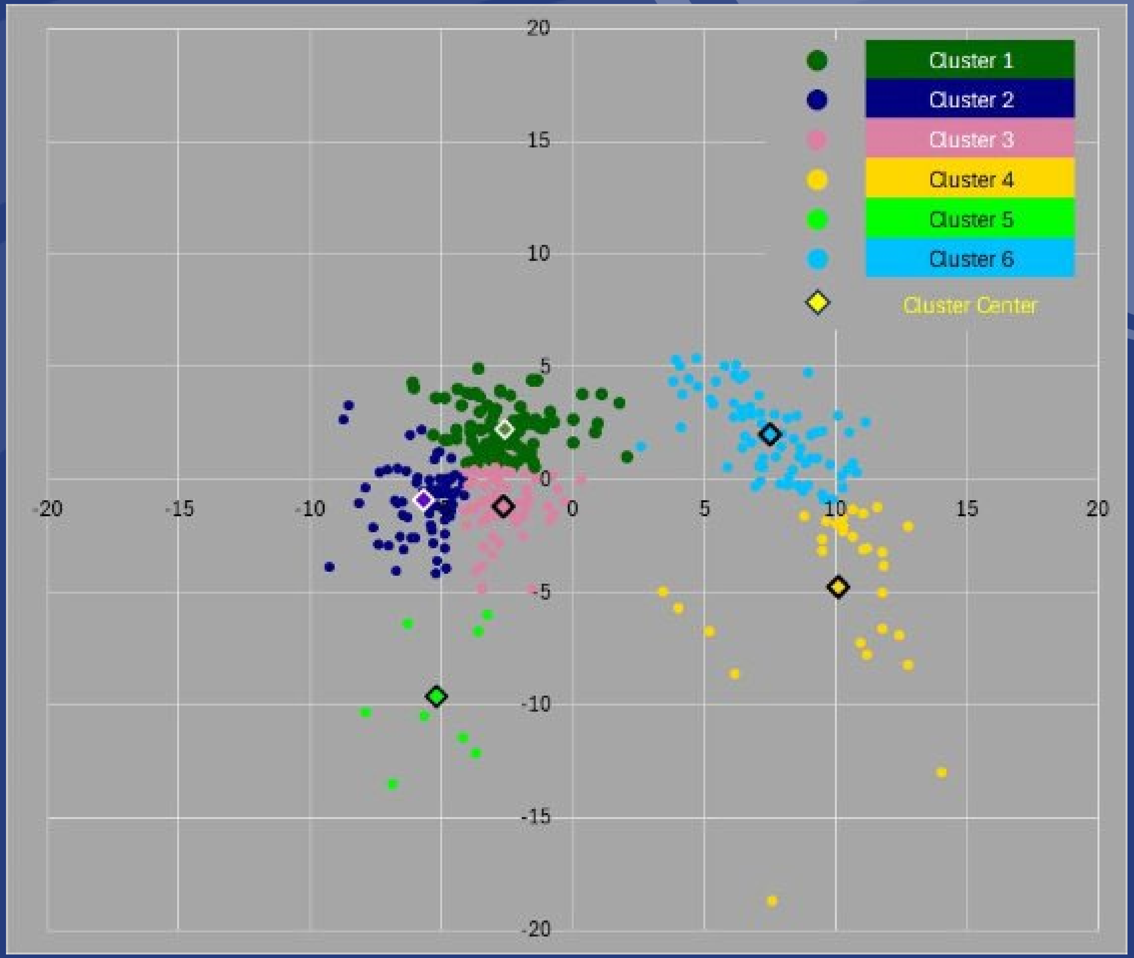
- How to incorporate more data and what data to use?
- How to identify representative days?
- What is a cluster analysis and how is it used?

Cluster Analysis

What is it? Why do it?

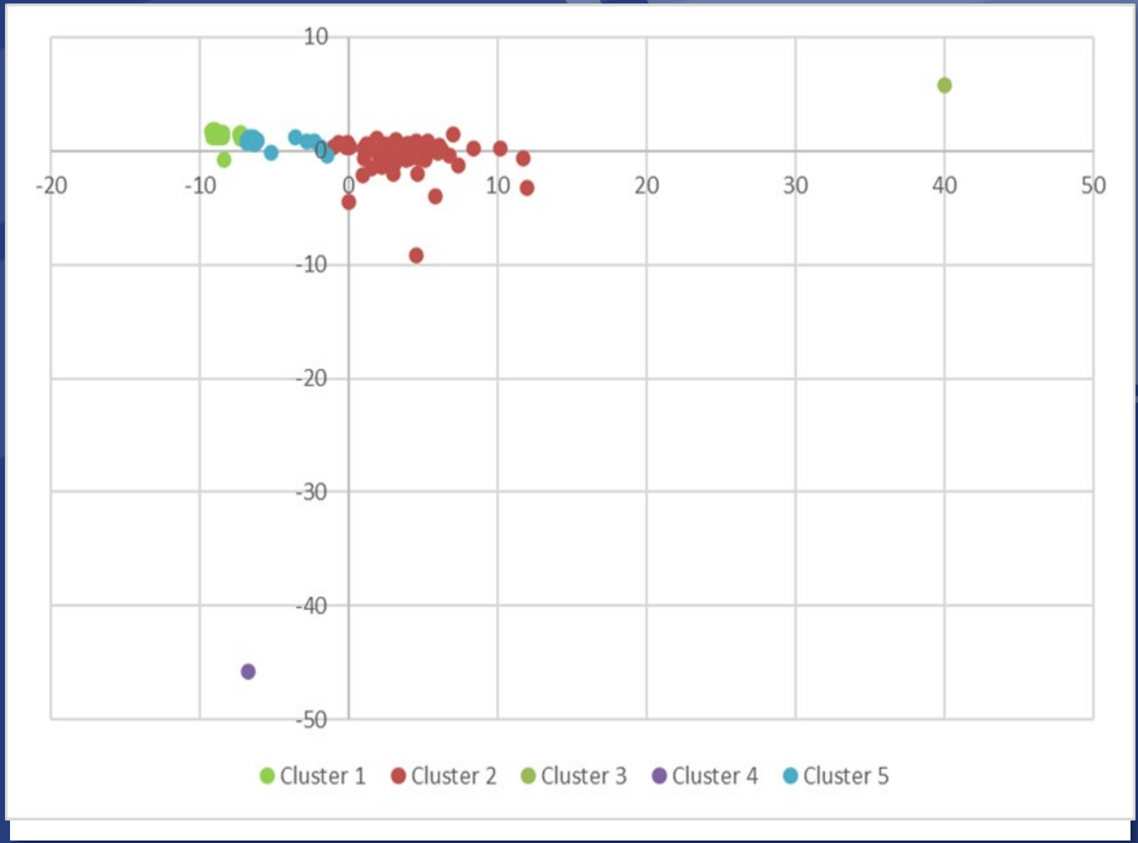
Purpose of a Cluster Analysis

- Grouping of objects (i.e. observations) into clusters so that objects within individual clusters are more similar to each other than to objects on other clusters
- For traffic analyses, these “objects” are data variables (e.g. traffic volumes, speeds)
- Purpose – Identify cluster(s) that are most representative of typical days for which analyses should be performed (and upon which decisions can be made)



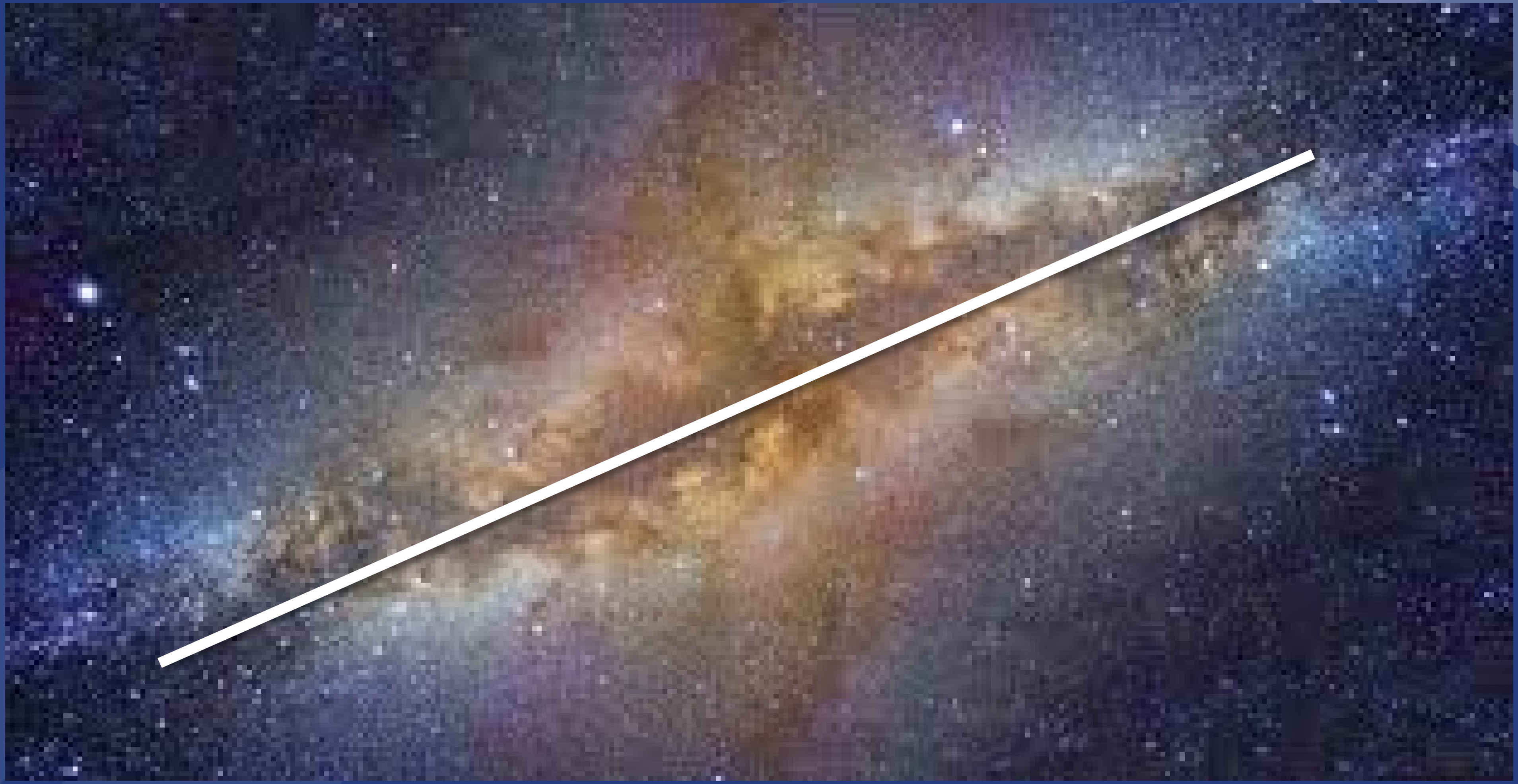
Filtering the Data

- Filter out redundant or low impact attributes
- Highly correlated with key measures of interest but lowly correlated with each other
- Principal Component Analysis (PCA) commonly used
- Each dimension was a new linear combination of original variables weighted differently such that new variables (principal components) were not correlated
- New axes provided best angle to see and evaluate the data



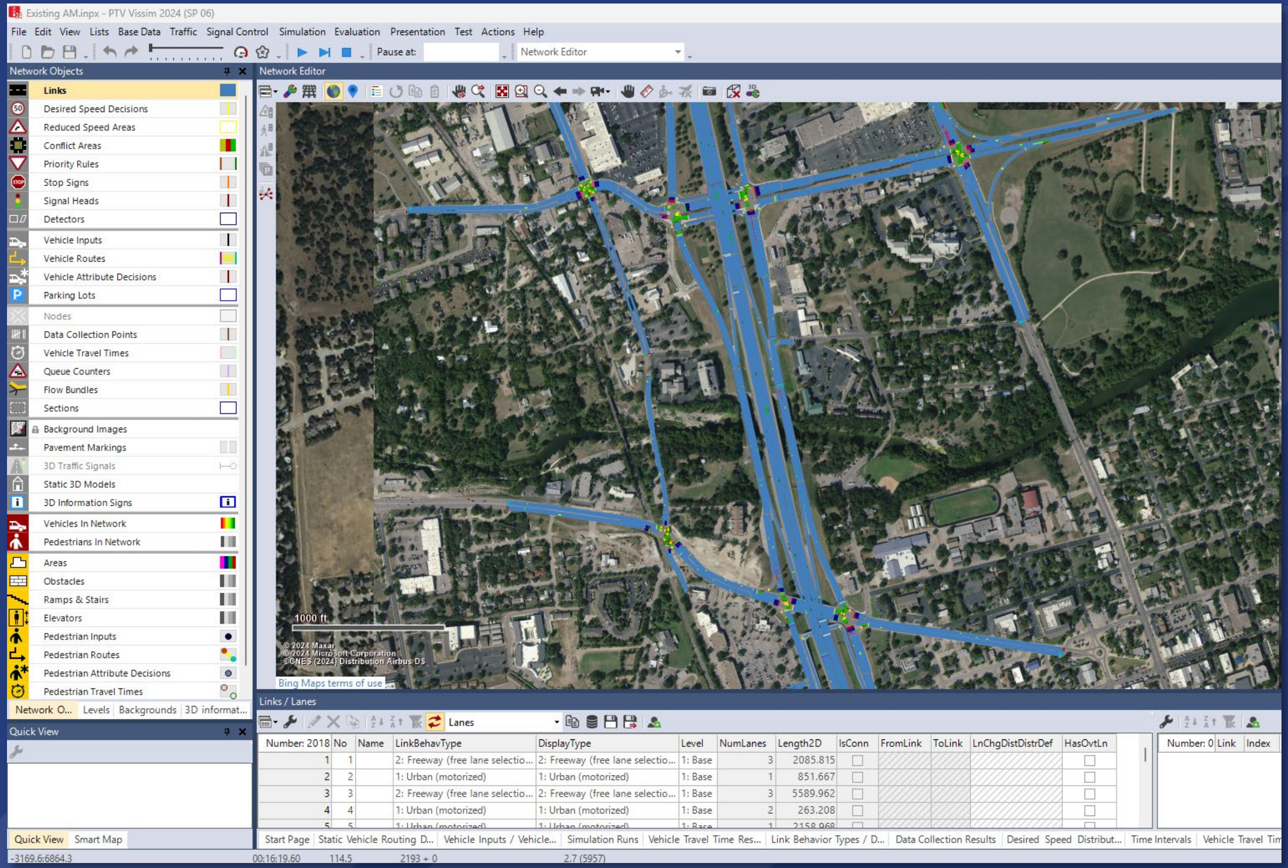
Source: Casey Cheng, published in *Towards Data Science*

Viewing the Milky Way Galaxy



Viewing the Milky Way Galaxy

Simulation Model Data Types Needed



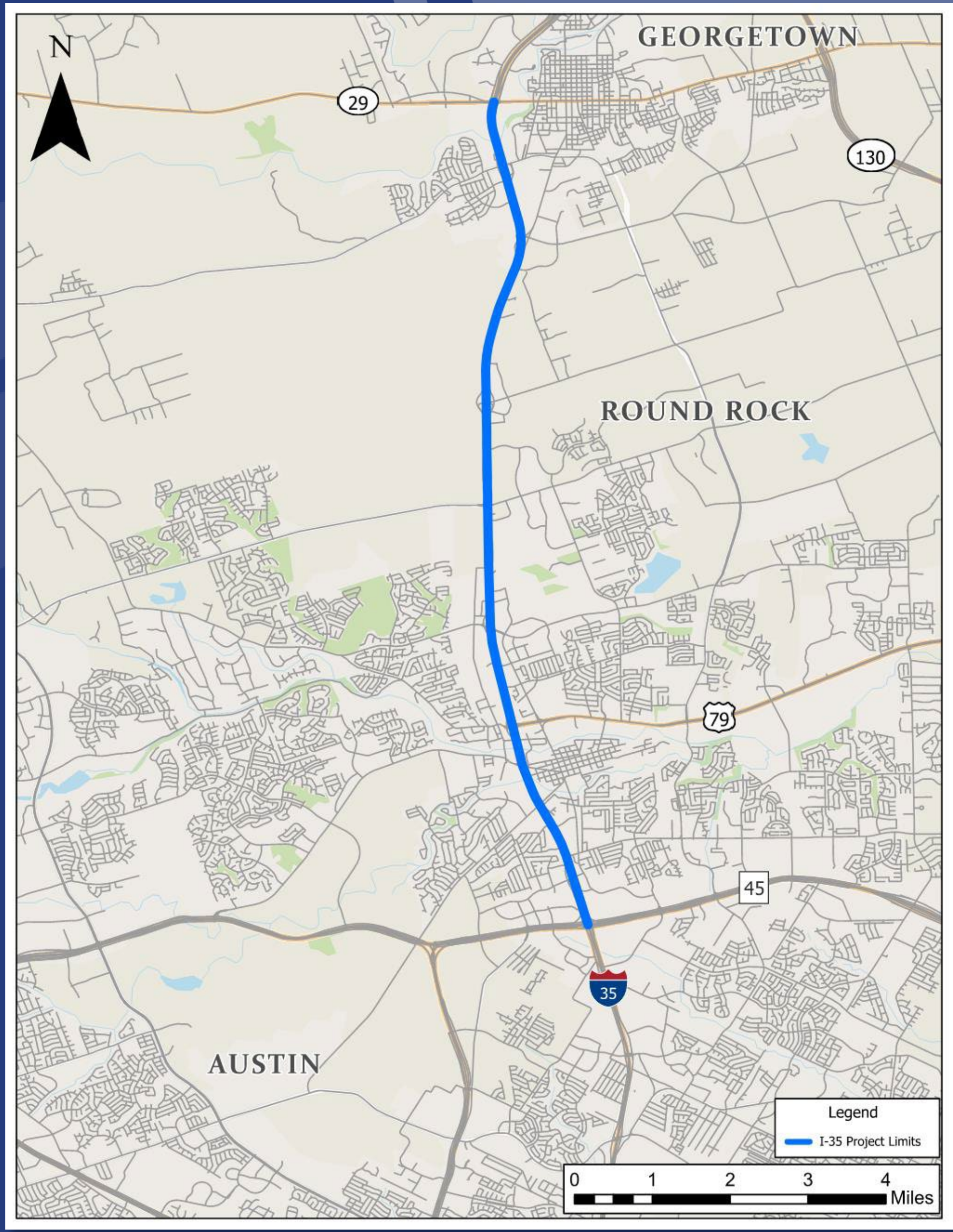
- Roadway geometry (length, # lanes, lane widths, horizontal and vertical alignment, etc.)
- Traffic control (speed limits, signs, signal timing, lane use restrictions)
- Demand volumes
- Vehicle and driver characteristics
- Event data affecting demand – precipitation and temperature, crashes, incidents, etc.

Additional Data (at a Minimum) for Model Calibration (TAT3 Guidance)

- Localized Performance Measure
  - To Capture Bottleneck Dynamics
  - Examples – Bottleneck Throughput or Duration, Density, Queuing
- System Performance Measure
  - Travel Time or Speed Profiles
- May choose additional performance measures to differentiate between alternatives
  - Crash or Incident Data
  - Weather Data (Precipitation, Temperature)
  - OD “Big Data”

Case Study IAJR

- I-35 north of Austin, TX
- From SH 45N to SH 29
- ~10.8 miles
- One permanent count station
- Objective: Provide example of how permanent count station was used to expand field traffic data that were used in the cluster analysis of the simulation model input data



Steps in the Process

- Identify Data Variables & Timeframe
- Collect Data
- Normalize Data
- Filter Data
- Determine Optimal No. Clusters
- Identify Cluster Characteristics
- Identify Representative Day(s)
- Develop Traffic Volume Factors

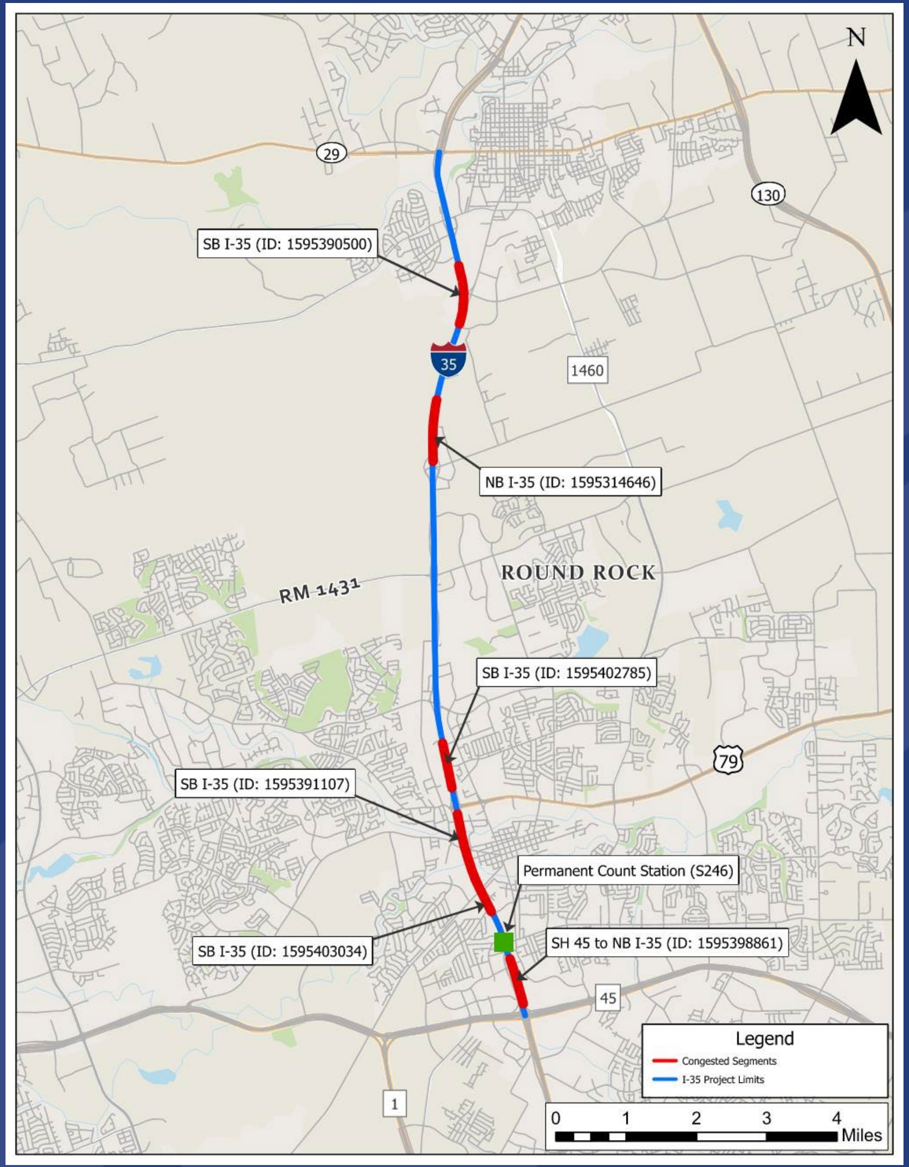
I-35 Case Study Data Elements

- Throughput (volumes) at Bottleneck Downstream Location Ends - Localized Performance Measure
- Travel Times for corridor and Bottleneck Locations (INRIX) – Systemwide Performance Measure
- Weather
  - Daily Precipitation
  - Average Daily Temperature
- Crash Data (Surrogate for Incidents)

Crash Severity Index

Crash Severity	Value
Not Injured	1
Possible Injury	2
Non-Suspected Serious Injury	3
Suspected Serious Injury	4
Death	5
Unknown Injury	1.25

Six Recurring Bottleneck Locations



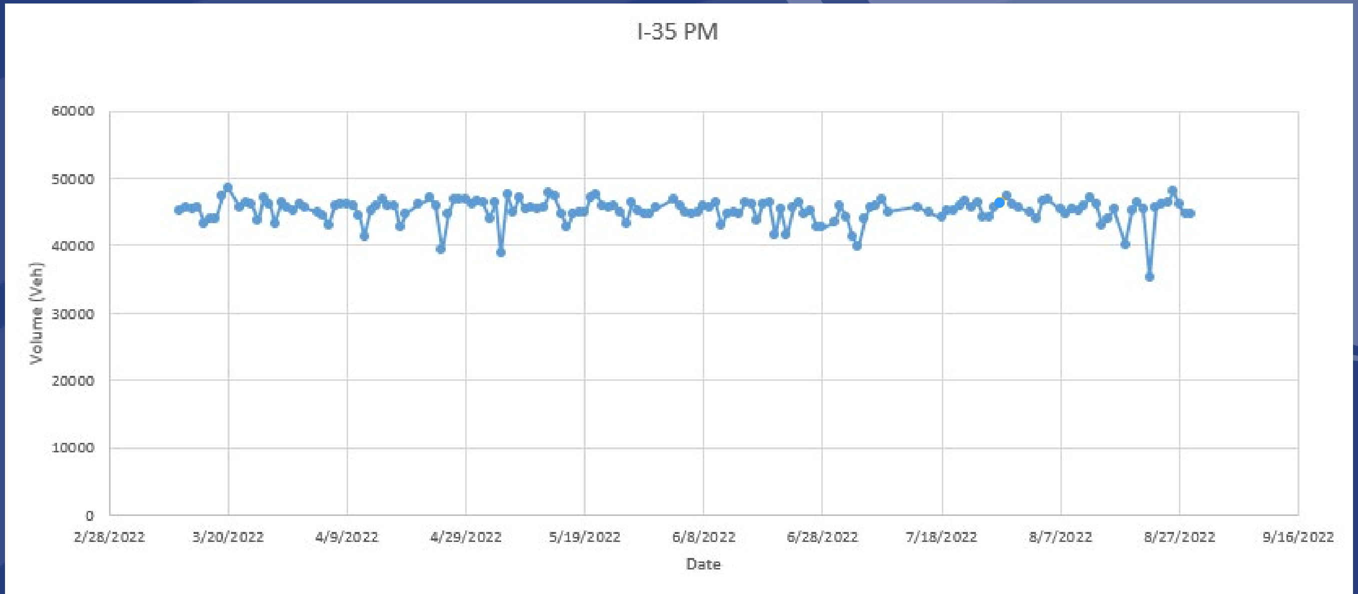
- I-35 Northbound – 3 segments
- I-35 Southbound – 3 segments
- INRIX segment IDs
- Permanent Count Station (S246)

Steps in the Process

- Identify Data Variables & Timeframe
- Collect Data
- Normalize Data
- Filter Data
- Determine Optimal No. Clusters
- Identify Cluster Characteristics
- Identify Representative Day(s)
- Develop Traffic Volume Factors

Data Collection

- Timeframe: March 12, 2022 – August 29, 2022
- 100 days of PCS data within this timeframe
- 38 days of actual counts (mainline, ramps and intersections)
- INRIX travel times
- Crash data from TxDOT Crash Record Information System (CRIS)
- Rain and temperature data from National Weather Service



Steps in the Process

1. Identify Data Variables & Timeframe
2. Collect Data
3. Normalize Data
4. Filter Data
5. Determine Optimal No. Clusters
6. Identify Cluster Characteristics
7. Identify Representative Day(s)
8. Develop Traffic Volume Factors

Normalize the Data

- Varying data types - values and units
- How to make comparable?
- Normalize – transform everything to a uniform, comparable basis
- Case Study: 0.00 – 1.00

<u>Example – Traffic Volumes</u>

Observed: 4,604

Range: 2,887 to 4,909

- $NormalizedValue = \frac{Observed - Minimum}{Maximum - Minimum}$
- $NormalizedValue = \frac{4{,}604 - 2{,}887}{4{,}909 - 2{,}887} = 0.85$

Data Normalization Example

Actual Data

Date	PCS1 Vol AM Peak	PCS2 Vol AMPeak	SL Zone Data	Avg. TTI1	Avg. TTI2	Ave. Temp.	Precip.
6/1/2022	8,093	7,472	560,275	5.2	3.2	73.2	0.00
6/2/2022	8,210	7,959	544,639	5.5	3.8	72.7	0.00
6/3/2022	8,327	9,419	567,499	5.1	3.1	68.7	0.58
6/4/2022	7,972	7,873	452,844	4.4	2.9	67.0	0.00
6/5/2022	6,788	8,269	379,018	4.1	2.9	63.6	0.00
6/6/2022	8,527	7,072	574,716	4.9	3.2	65.4	0.00
6/7/2022	7,210	6,924	561,667	5.0	3.1	69.3	0.00
6/8/2022	8,052	7,174	552,891	4.9	3.1	71.6	0.00
6/9/2022	8,216	8,044	521,091	4.9	3.2	70.5	0.32
6/10/2022	8,594	9,441	573,086	4.9	3.0	66.9	0.00
6/11/2022	8,204	7,493	439,223	4.5	2.9	68.8	0.00
6/12/2022	6,792	9,295	366,781	4.1	2.9	73.8	0.00
6/13/2022	8,483	7,388	560,216	4.9	3.1	77.2	0.00
6/14/2022	7,749	6,961	502,343	5.0	3.2	75.1	0.00
6/15/2022	8,256	7,350	561,812	4.8	3.2	81.3	0.05
6/16/2022	8,492	8,285	565,519	4.9	3.0	74.4	0.00
6/17/2022	8,854	9,484	547,962	4.8	3.0	74.2	0.38
6/18/2022	8,593	8,594	451,939	4.4	2.9	72.5	0.00
6/19/2022	6,869	9,356	373,245	4.2	2.9	68.2	0.00
6/20/2022	8,463	8,434	557,145	4.7	3.0	67.8	0.00
6/23/2022	8,432	8,169	515,114	4.9	3.0	79.3	0.00
6/24/2022	8,997	9,404	474,592	4.9	3.0	73.0	0.00
6/25/2022	8,479	7,903	462,124	4.3	2.9	73.6	0.00
6/26/2022	6,687	9,956	365,553	4.0	2.9	71.7	0.00

Normalized Data

Date	PCS1 AM Peak	PCS2 AMPeak	SL Zone Data	Avg. TTI1	Avg. TTI2	Ave. Temp.	Precip.
6/1/2022	0.820	0.932	0.877	0.555	0.231	0.886	0.000
6/2/2022	0.837	0.919	0.846	0.709	0.388	0.879	0.000
6/3/2022	0.854	0.907	0.892	0.537	0.293	0.823	0.324
6/4/2022	0.803	0.502	0.660	0.201	0.129	0.799	0.000
6/5/2022	0.633	0.315	0.511	0.082	0.092	0.752	0.000
6/6/2022	0.882	0.960	0.907	0.445	0.273	0.777	0.000
6/7/2022	0.693	0.903	0.880	0.479	0.213	0.832	0.000
6/8/2022	0.814	0.909	0.862	0.444	0.231	0.864	0.000
6/9/2022	0.838	0.885	0.798	0.449	0.262	0.849	0.179
6/10/2022	0.892	0.872	0.903	0.453	0.298	0.798	0.000
6/11/2022	0.836	0.499	0.633	0.243	0.126	0.825	0.000
6/12/2022	0.633	0.330	0.486	0.067	0.068	0.895	0.000
6/13/2022	0.876	0.830	0.877	0.412	0.216	0.942	0.000
6/14/2022	0.771	0.783	0.760	0.473	0.242	0.913	0.000
6/15/2022	0.843	0.885	0.880	0.367	0.232	1.000	0.028
6/16/2022	0.877	0.875	0.888	0.443	0.392	0.903	0.000
6/17/2022	0.929	0.876	0.852	0.397	0.361	0.900	0.212
6/18/2022	0.892	0.474	0.658	0.210	0.223	0.877	0.000
6/19/2022	0.644	0.337	0.499	0.123	0.127	0.816	0.000
6/20/2022	0.873	0.813	0.871	0.324	0.220	0.811	0.000
6/23/2022	0.869	0.890	0.786	0.438	0.347	0.972	0.000
6/24/2022	0.950	0.836	0.704	0.427	0.365	0.884	0.000
6/25/2022	0.875	0.456	0.679	0.171	0.181	0.892	0.000
6/26/2022	0.618	0.267	0.484	0.003	0.140	0.865	0.000

Steps in the Process

- Identify Data Variables & Timeframe
- Collect Data
- Normalize Data
- Filter Data
- Determine Optimal No. Clusters
- Identify Cluster Characteristics
- Identify Representative Day(s)
- Develop Traffic Volume Factors

Principal Component Analysis

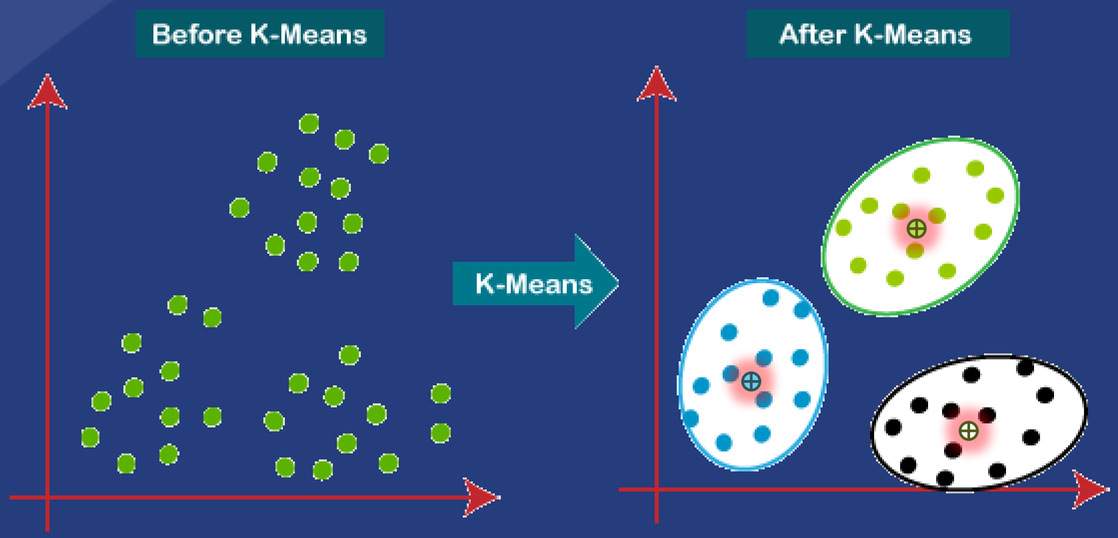
- PCA algorithm applied in WEKA
- Combined into two-dimensions such that the new variables (principal components) are not correlated
- Data dimensionality is reduced while preserving original information

Steps in the Process

- Identify Data Variables & Timeframe
- Collect Data
- Normalize Data
- Filter Data
- Determine Optimal No. Clusters
- Identify Cluster Characteristics
- Identify Representative Day(s)
- Develop Traffic Volume Factors

Cluster Analysis

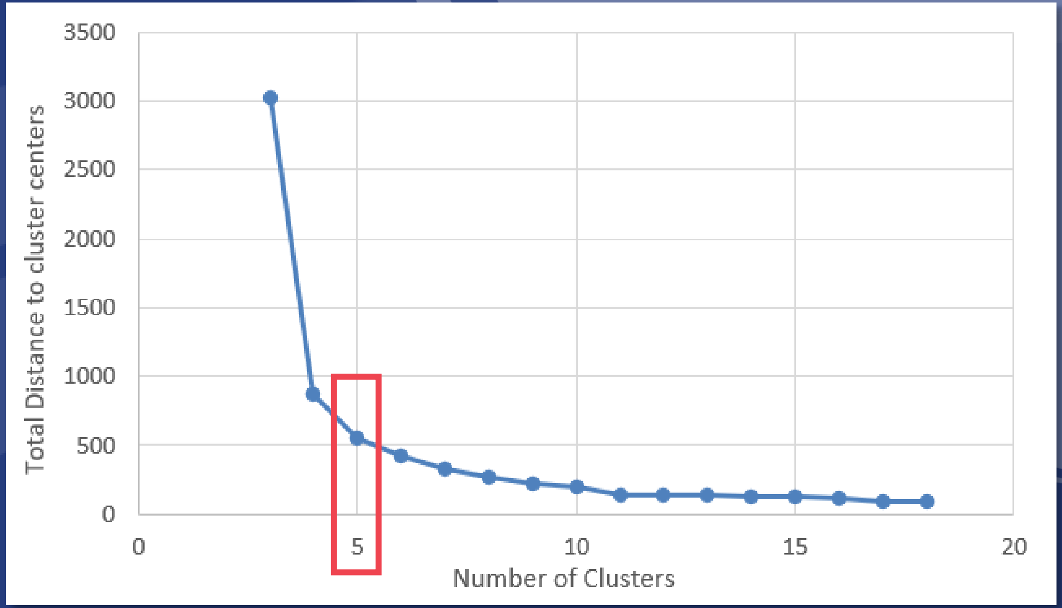
- TAT3 mentions several possible clustering techniques
- K-means is most widely used
- Objective: Partition or separate total number of observations (*n*) into *k* clusters such that each observation belongs to cluster having the closest mean
- Each cluster has its own mean (centroid)

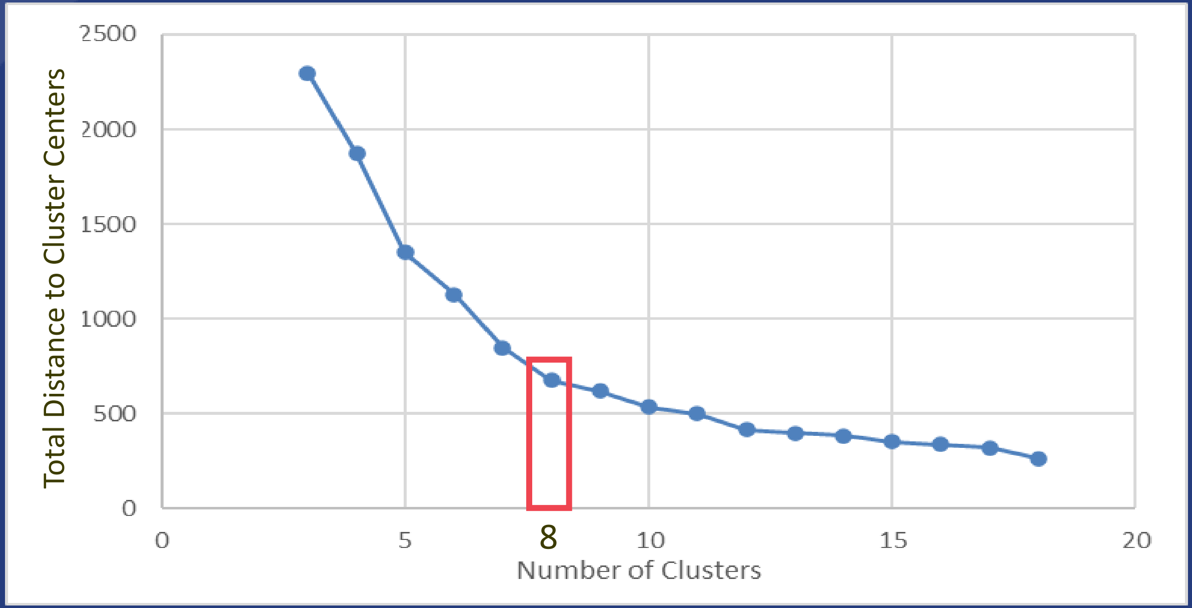


Source: *Analytics Vidhya*

Optimum Number of Clusters

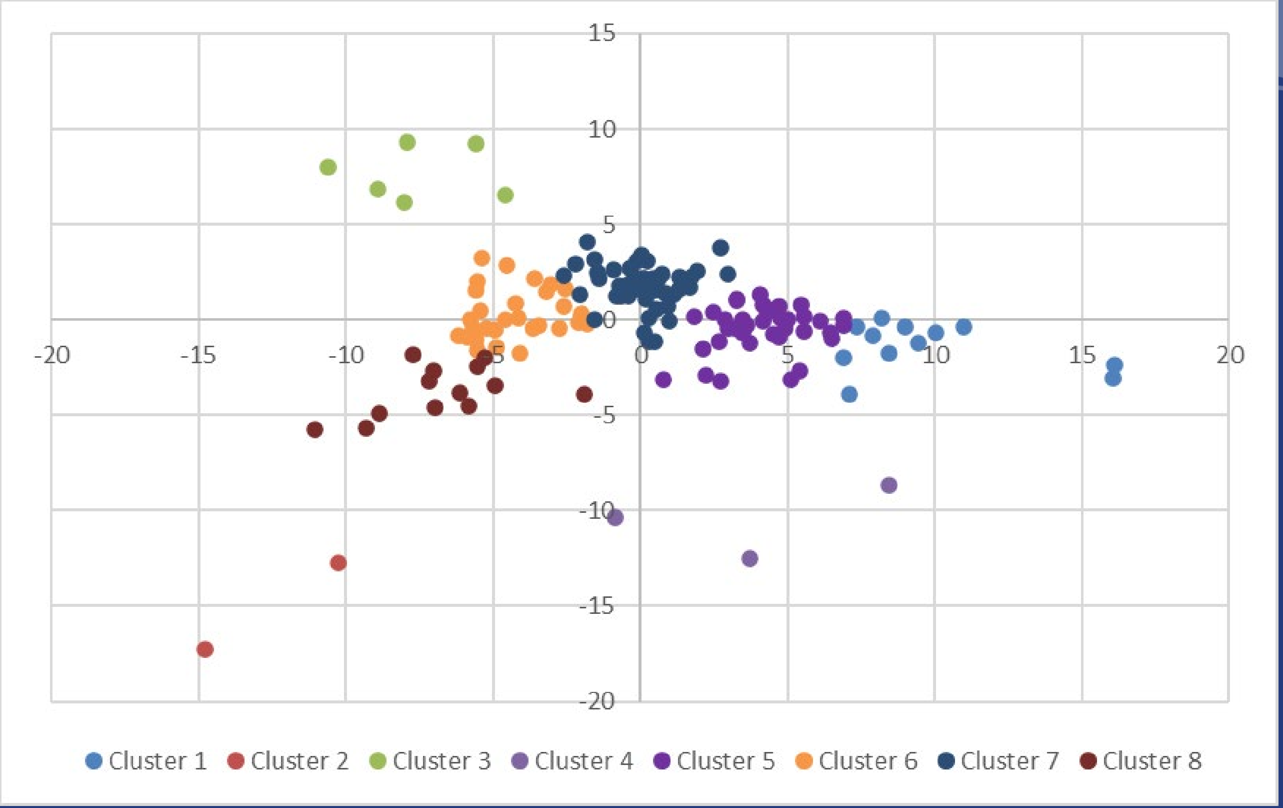
- Too few clusters – greater size, variability in the data set
- Too many clusters – smaller size, too many different scenarios to evaluate
- Optimal – Evaluate reasonable number of scenarios that are most representative of normal conditions that support comprehensive decision making
- Elbow Method – Easily understood and frequently used in k-means



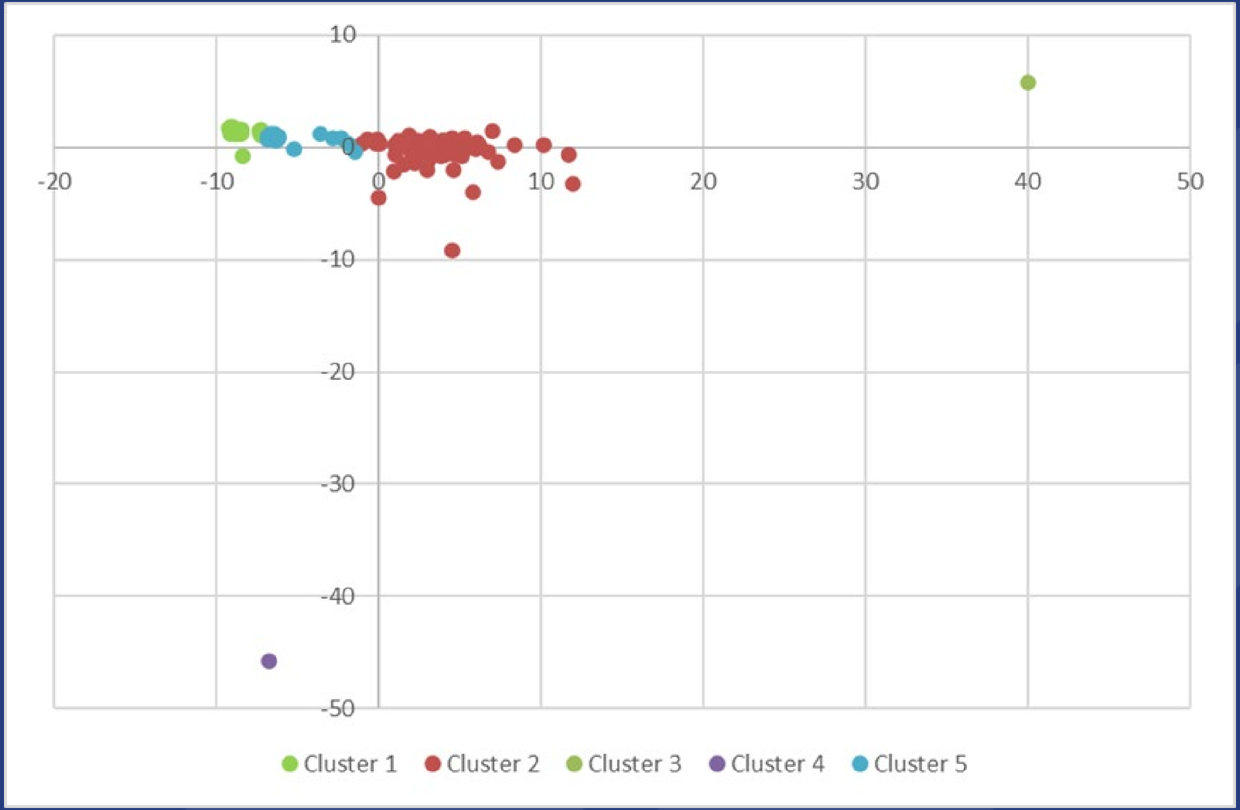


Two-Dimensional Cluster Results

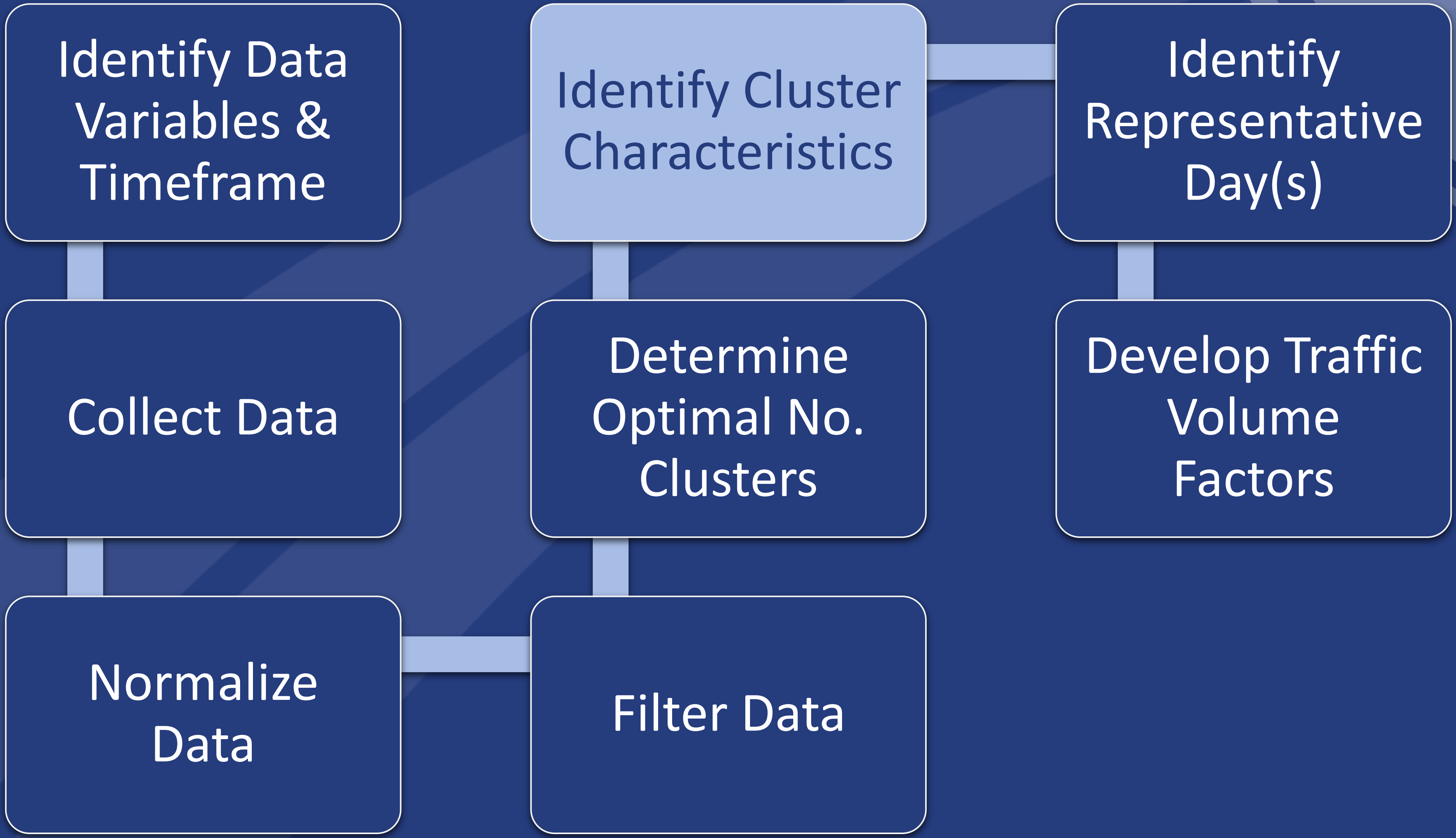
PM Peak



AM Peak



Steps in the Process



Cluster Characteristics– A.M. Peak

Day of the Week	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Monday	2	19	0	0	1
Tuesday	0	22	1	0	0
Wednesday	0	19	0	0	0
Thursday	0	22	0	0	0
Friday	0	19	0	0	3
Saturday	3	0	0	1	20
Sunday	22	0	0	0	0
Total	**27**	**101**	**1**	**1**	**24**

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
		Sunday	Weekday	Outlier #1	Outlier #2	Saturday
Permanent Count Station Volume (veh)	Northbound	10,515	20,879	18,630	15,074	16,202
	Southbound	9,536	21,288	11,359	11,899	15,149
Average Travel Time (sec)	1595398861	30.35	39.21	31.87	29.917	32.98
	1595314646	37.78	43.73	38.48	386.13	37.73
	1595390500	42.00	89.86	312.32	47.80	47.42
	1595391107	23.87	31.64	137.20	25.12	25.51
	1595403034	28.32	73.88	167.58	31.61	30.20
	1595402785	41.04	42.60	41.97	41.49	41.58
Average Crash Factor		0.08	4.40	0	0	1.83
Average Temperature (°F)		77	79	88	87	79
Average Precipitation (in)		0.07	0.06	0	0	0.06

Cluster Characteristics– P.M. Peak

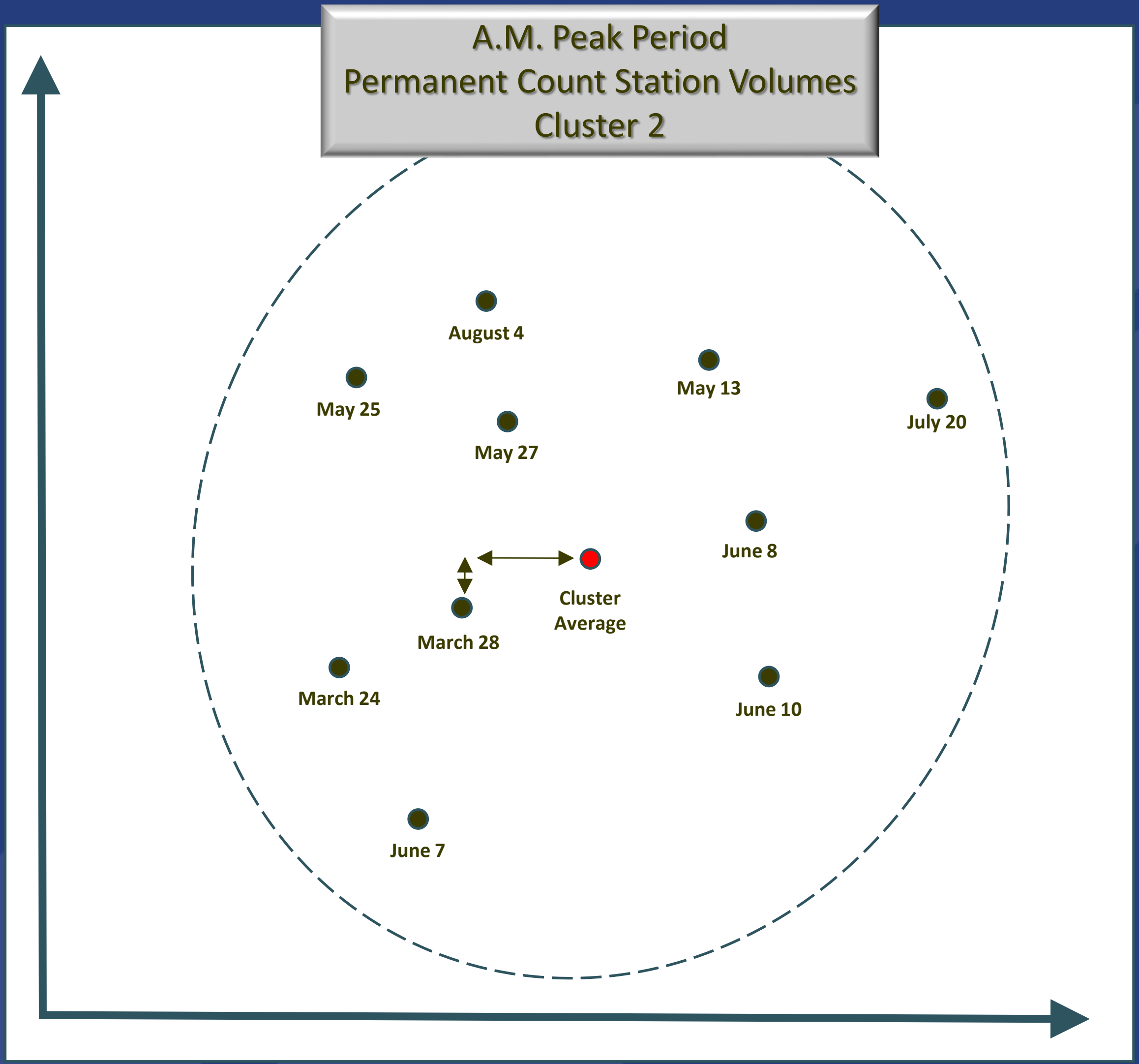
Day of the Week	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Monday	3	0	2	0	1	2	14	0
Tuesday	1	0	1	0	3	3	14	1
Wednesday	0	0	1	0	1	3	12	2
Thursday	0	0	2	1	0	11	7	1
Friday	0	2	0	0	0	11	0	9
Saturday	0	0	0	0	20	0	4	0
Sunday	8	0	0	2	12	0	0	0
Total	**12**	**2**	**6**	**3**	**37**	**30**	**51**	**13**

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
		Lighter Sundays	Outlier #1	Low Vol High TT Weekdays	Outlier #2	Weekend	Thursday/ Friday	Monday-Wednesday	Weekday High CF-Higher TT
PCS Volume (veh)	Northbound	23,373	23,831	22,457	23,368	23,954	24,210	24,075	24,255
	Southbound	21,177	21,245	17,542	21,468	21,832	21,344	21,042	21,409
Average Travel Time (sec)	1595398861	38.94	97.88	102.02	47.14	44.62	101.74	87.33	108.04
	1595314646	39.27	51.89	45.99	68.02	40.22	40.89	41.07	48.33
	1595390500	78.10	147.66	133.52	102.28	103.86	124.66	110.16	135.05
	1595391107	30.68	46.05	43.05	33.13	33.52	37.51	34.88	39.99
	1595403034	51.23	134.57	118.28	72.18	79.77	113.98	89.63	123.82
	1595402785	42.11	106.09	47.71	116.10	45.83	47.16	45.36	53.03
Crash Factor		0.75	3	2.16	5	4.81	3.55	2.99	9.84
Average Temperature (°F)		81.37	67	82.5	85	79.55	77.45	78.87	74.34
Average Precipitation (in)		0.14	0	0.023	0.0033	0.10	0.049	0.024	0.016

Steps in the Process

- Identify Data Variables & Timeframe
- Collect Data
- Normalize Data
- Filter Data
- Determine Optimal No. Clusters
- Identify Cluster Characteristics
- Identify Representative Day(s)
- Develop Traffic Volume Factors

Identify Representative Day

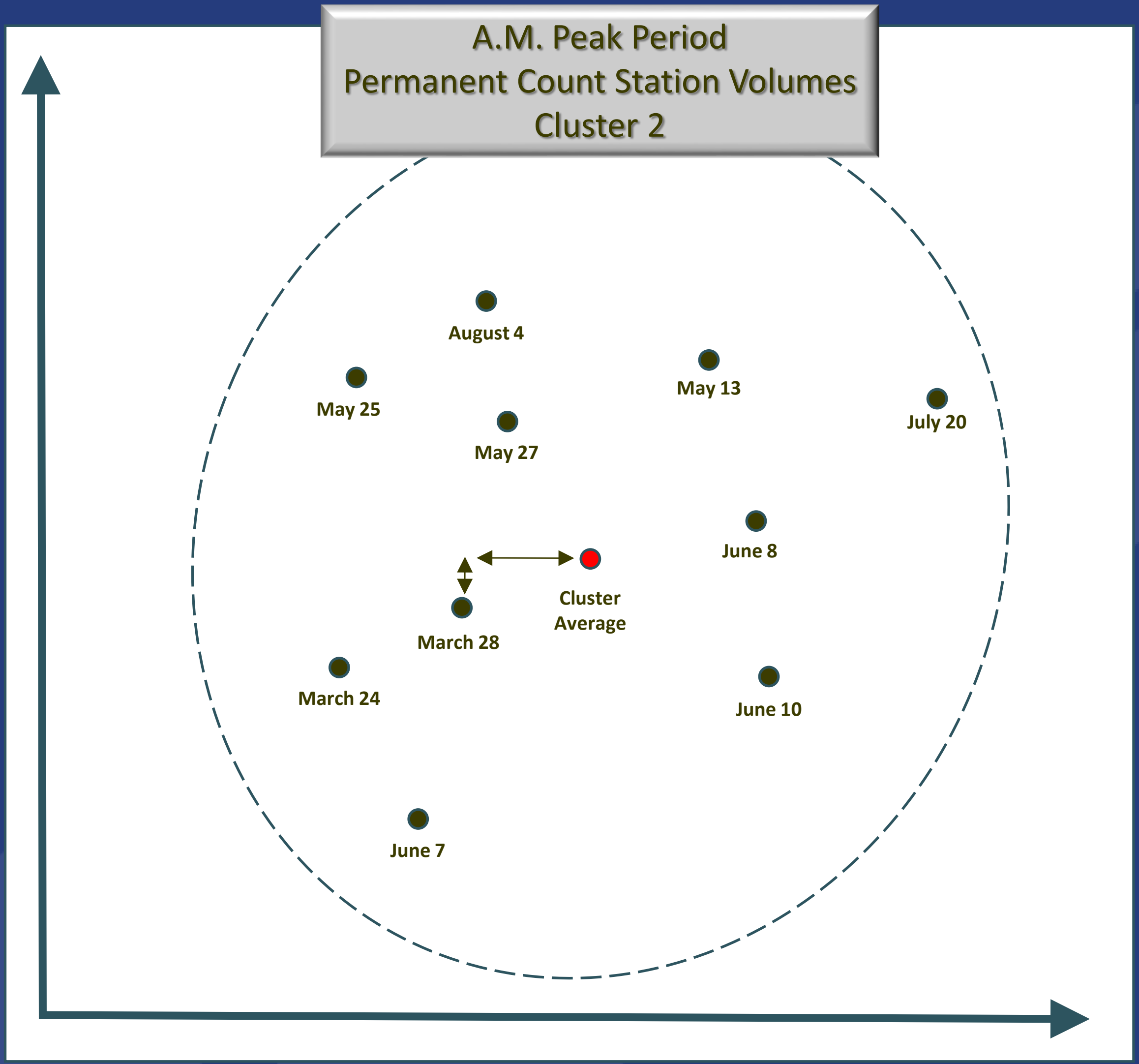


1. Determine average value for each input variable in cluster
2. Calculate difference from cluster average, expressed as percentage of the mean

$$d_{m,i} = \frac{\sqrt{(m_{avg} - m_i)^2}}{m_{avg}}$$

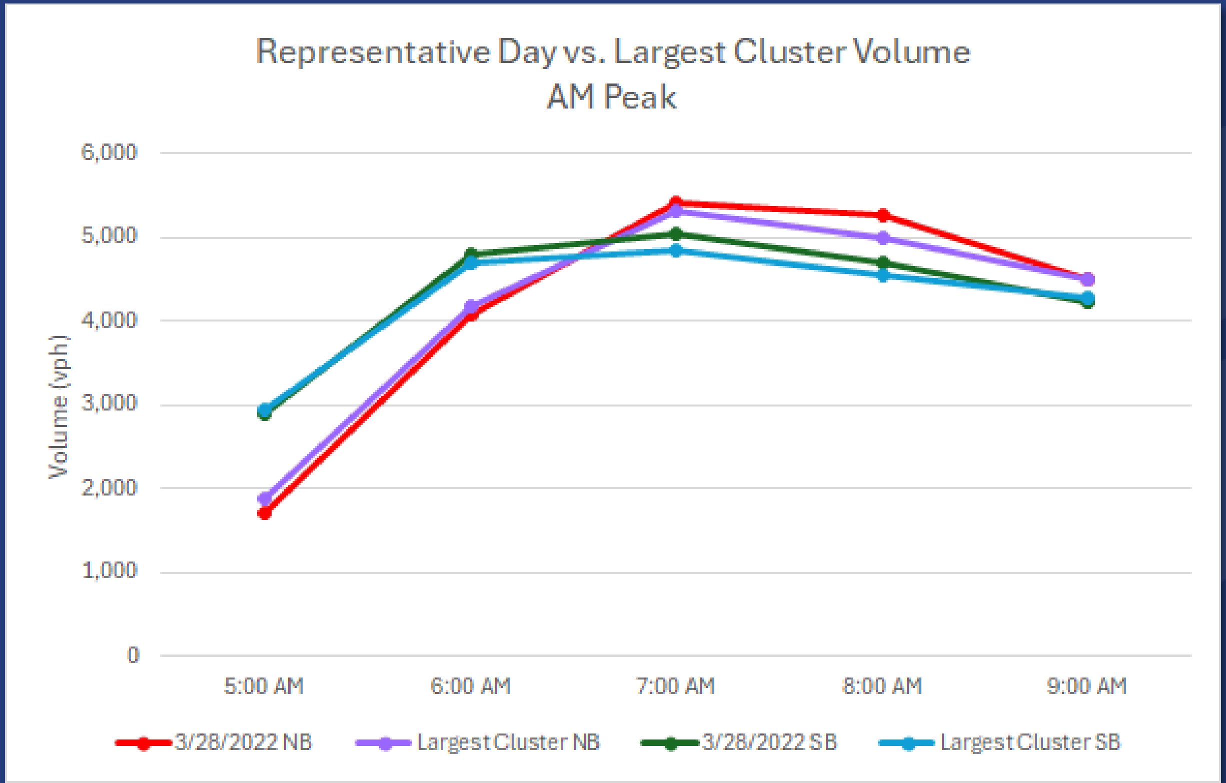
3. Sum distance to mean

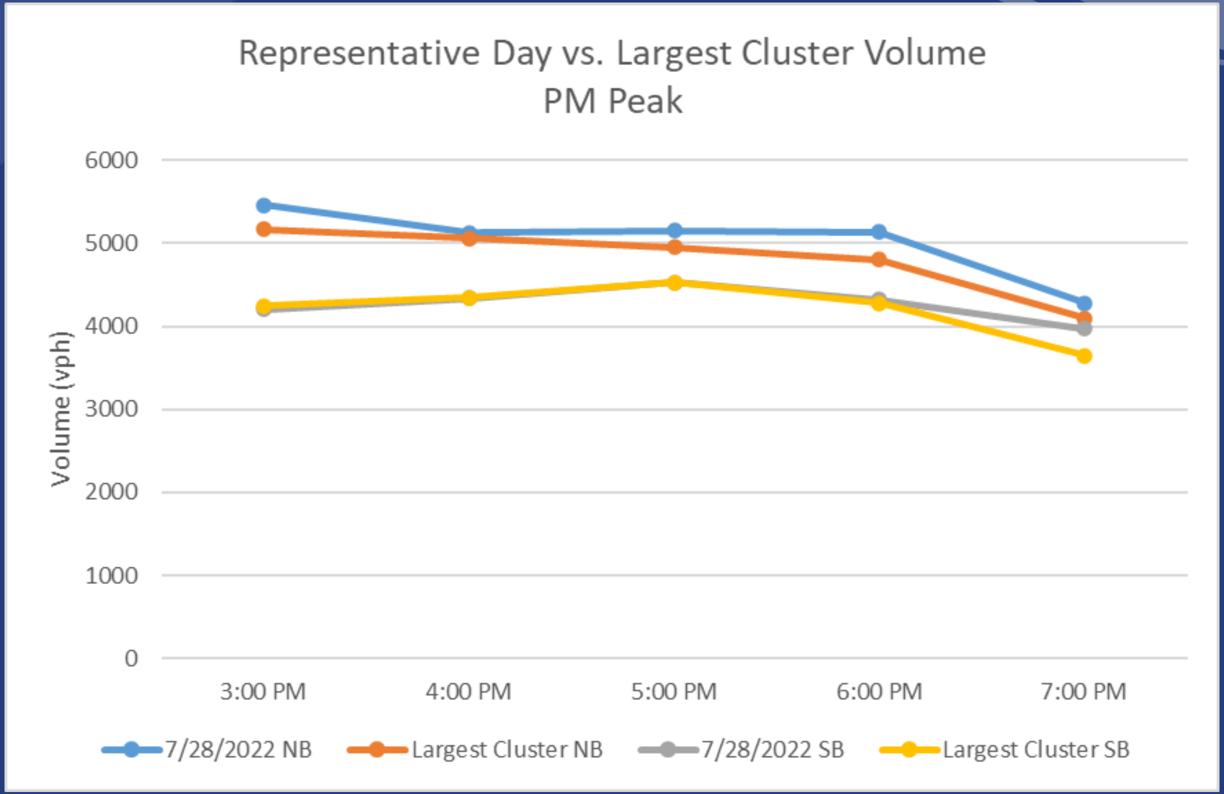
A.M. Peak Most Representative Day



Date	Sum of Distances
March 28th	0.47
May 27th	1.15
June 8th	1.15
May 25th	1.71
August 4th	1.74
March 24th	2.21
June 7th	2.42
May 13th	2.64
July 20th	2.99
June 10th	3.31

Representative Day vs. Largest Cluster Volume





Steps in the Process

1. Identify Data Variables & Timeframe
2. Collect Data
3. Normalize Data
4. Filter Data
5. Determine Optimal No. Clusters
6. Identify Cluster Characteristics
7. Identify Representative Day(s)
8. Develop Traffic Volume Factors

Application to Simulation Models

When the representative days are not same as the days for which traffic counts were collected, how are the count data used?

$$Representatative\ Day\ Factor = \frac{PCS\ Vol_{RepDayCluster}}{PCS\ Vol_{CountDay}}$$

PCS: Permanent Count Station – within project limits or nearby

Factors applied to actual counts (segment, turning movements) used as input to existing conditions simulation models

Count Date	AM Factor	PM Factor
May 3rd	1.028	1.055
May 4th	1.007	1.005
May 5th	1.011	1.196
May 17th	0.986	1.043
May 19th	0.991	1.037
May 26th	1.013	1.078
June 7th	0.986	1.036

Summary

1. Identify data needs, types, sources
2. Assemble and prepare data
3. Normalize data
4. Reduce dimensionality (Principal Component Analysis)
5. Determine optimal number of clusters
6. Identify Cluster Characteristics
7. Identify representative day(s)
8. Apply representative day factors to simulation model inputs

ATG

DCCM

Questions?

Tom Creasey, P.E., Ph.D.
ATG | DCCM
tcreasey@dccm.com

Adam Stacy, P.E.
ATG | DCCM
astacy@dccm.com