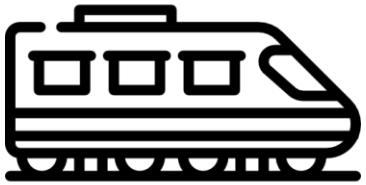


Lucia Scheele & Laura Vetter

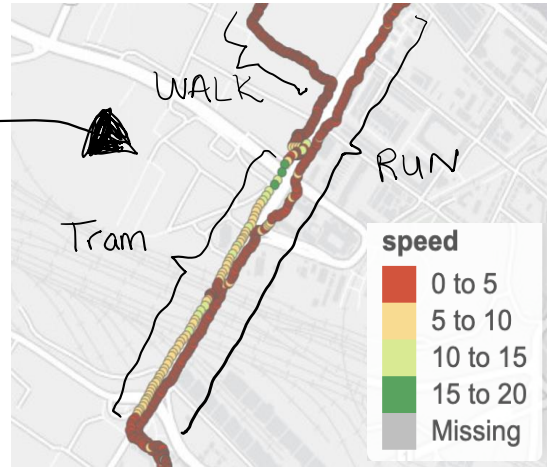
Patterns & Trends in Environmental data FS 2024

Travel Mode Detection with k-means clustering

Data and research question

Strava trajectories (randomly recorded)

! Mixed-movement types in one trajectory

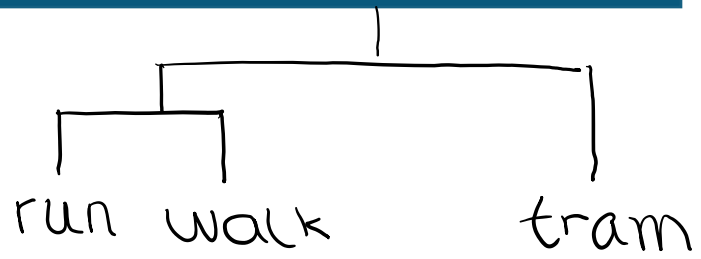


elevation	timestamp	ActivityName	ActivityType	year	geometry
410.4	2024-04-02 14:05:15	Zugfahrt nach Hause	trainride	2024	POINT (2692957 1232239)
410.4	2024-04-02 14:05:16	Zugfahrt nach Hause	trainride	2024	POINT (2692958 1232232)
410.4	2024-04-02 14:05:17	Zugfahrt nach Hause	trainride	2024	POINT (2692954 1232235)

Which travel modes can be differentiated in a collection of trajectories using k-means clustering?

Best possible outcome:

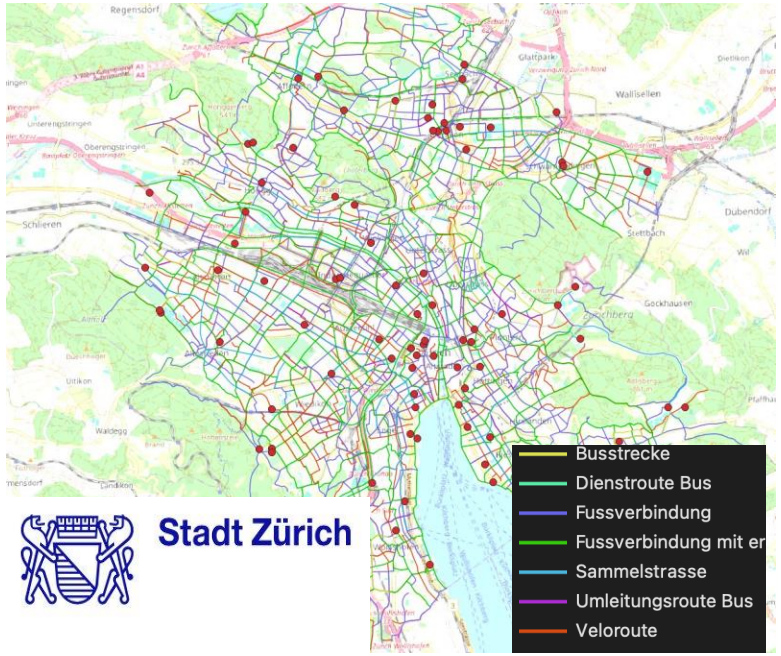
→ **Clusters** contain only one movement type



Contextual data

Kommunaler Richtplan Verkehr

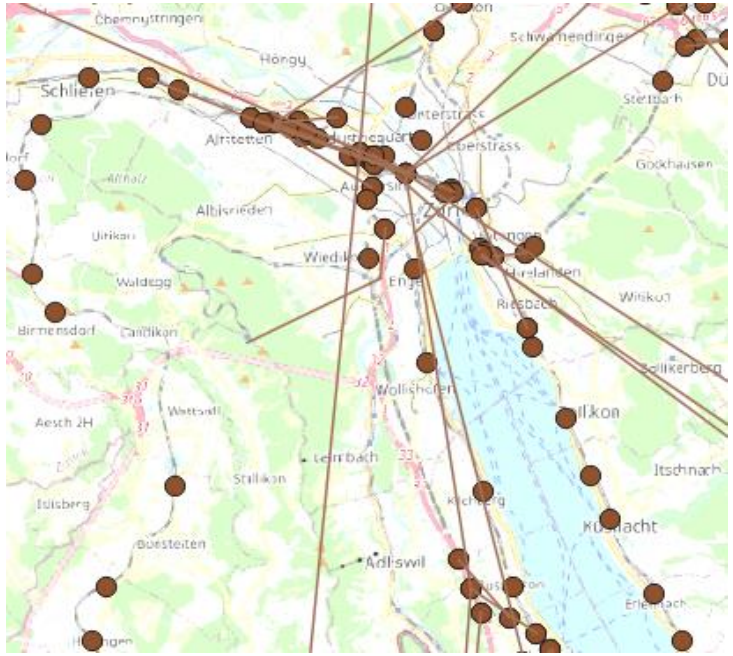
Information on Street type of Zurich



https://data.stadt-zuerich.ch/dataset/geo_kommunaler_richtplan_verkehr

Betriebspunkte

Location of train stations in Switzerland



https://data.sbb.ch/explore/dataset/linie-mit-betriebspunkten/information/?location=12,47.39138,8.52385&basemap=00c4d6

Research plan

Preprocessing

Mixed type trajectories → Segmentation → Clean **segments** →

Enrich **segements** with:

- Movement parameters
- Environmental data

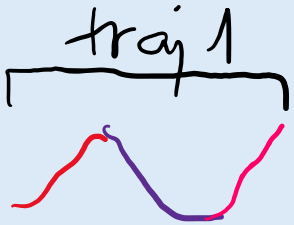
→ Enriched **segments** Table

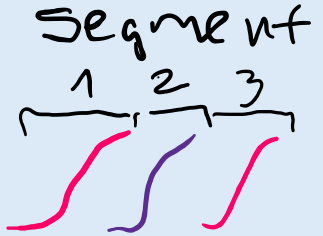
Analysis

Enriched Data Table

K-means clustering algorithm

Cluster categories



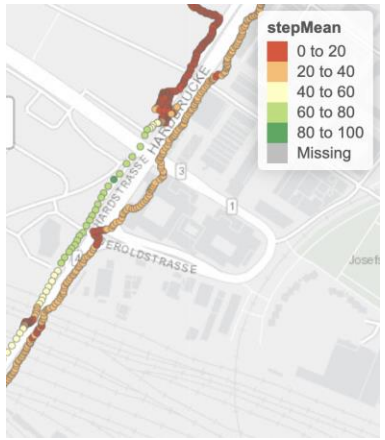


Verification

Comparison

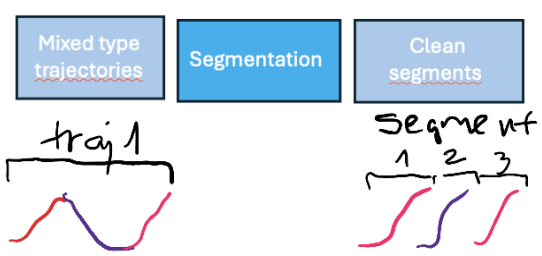
- Visual knowledge
- Manual classification in GIS

Results: Segmentation

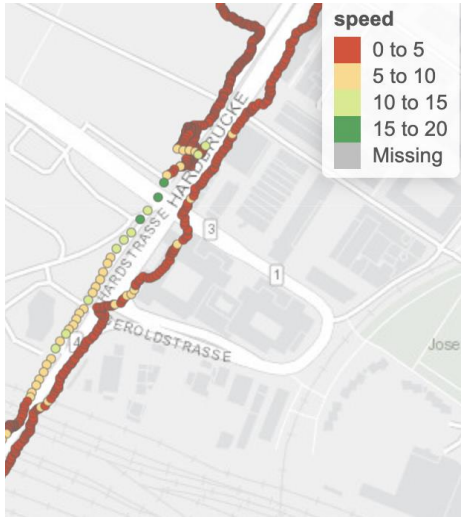



```r
mixed_Laura  <- mixed_Laura  |>
  mutate(category = case_when(
    stepMean < 20 ~ "walking",
    stepMean >= 20 & stepMean < 40 ~ "running",
    stepMean >= 40 ~ "tram"
  ))
```

Which parameter splits the mixed trajectory at the point where the movement type changes?



time_diff	speed	speed_kmh	acceleration	avg_acc_10s
1	1.6946088	6.100592	-5.300063e-01	0.2210812594
1	1.2286238	4.423046	-4.659850e-01	-0.1569585866
1	2.0737675	7.465563	8.451437e-01	-0.2627122120
1	2.6529501	9.550620	5.791825e-01	-0.0981261405
1	1.2169836	4.381141	-1.435967e+00	0.1229935197
1	3.5462777	12.766600	2.329294e+00	-0.0552682745
1	1.4832470	5.339689	-2.063031e+00	0.2286238290
1	0.3751003	1.350361	-1.108147e+00	0.0715153672



Plan B: Use fix information

~~Mixed type trajectories~~ ~~Segmentation~~ ~~Clean segments~~

Preprocessing | Analysis | Verification

Table of trajectory fixes (= single points)

Enrich data

Movement parameters

Environmental data

Enriched fixes

K-means clustering algorithm

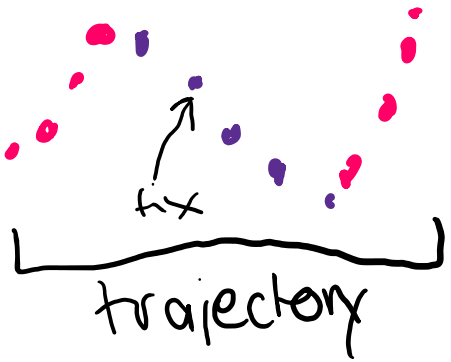
Comparison

Fixes with cluster categories

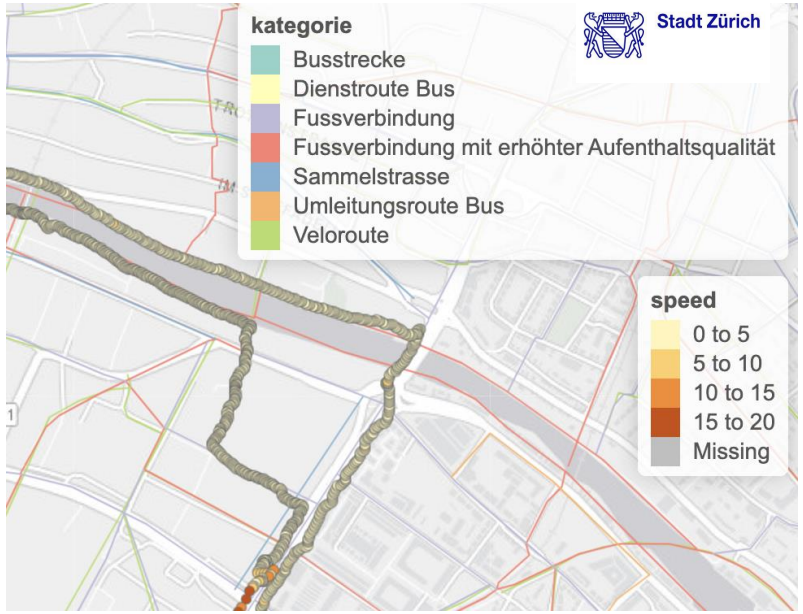
vs.

Visual knowledge

Manual classification in GIS



Results: Join with contextual data



No actual intersect. Neighbourhood analysis:

Enrich data

Movement parameters

Environmental data

```
joined <- st_join(mixed_Laura, lines_select,
                  join = st_is_within_distance,
                       dist = 0.5, # Distanz in Metern
                       left = TRUE)# damit auch werte erhalten
bleiben, die nicht within distance sind
```

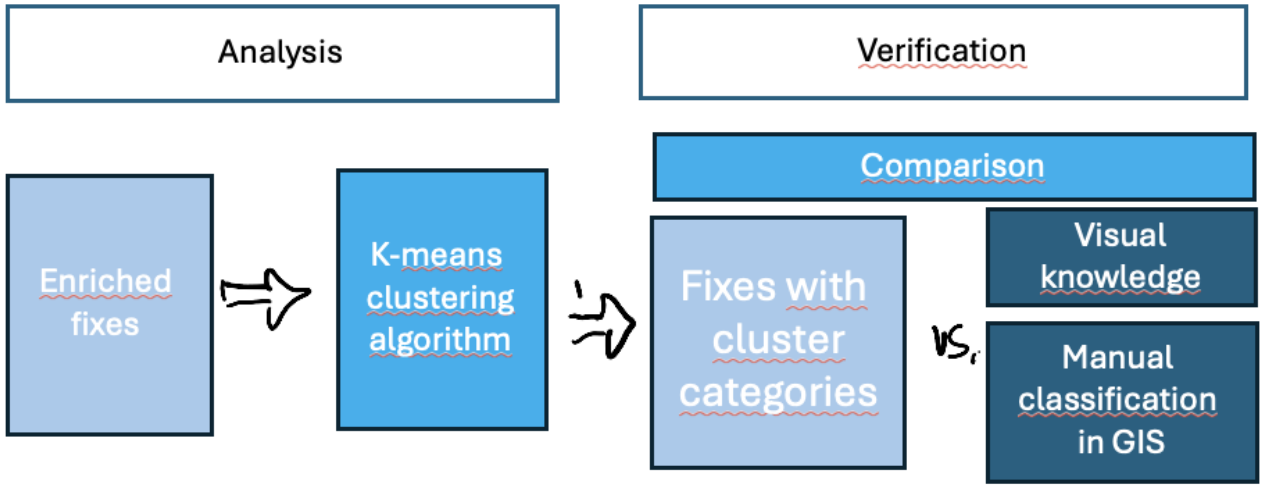
kategorie	teilplan
Veloroute	Velo
Veloroute	Velo
Veloroute	Velo
NA	*NA*
NA	*NA*

Works! But only 230 fixes have description, out of 3800 (for 1 trajectory)

Work in Progress! :)

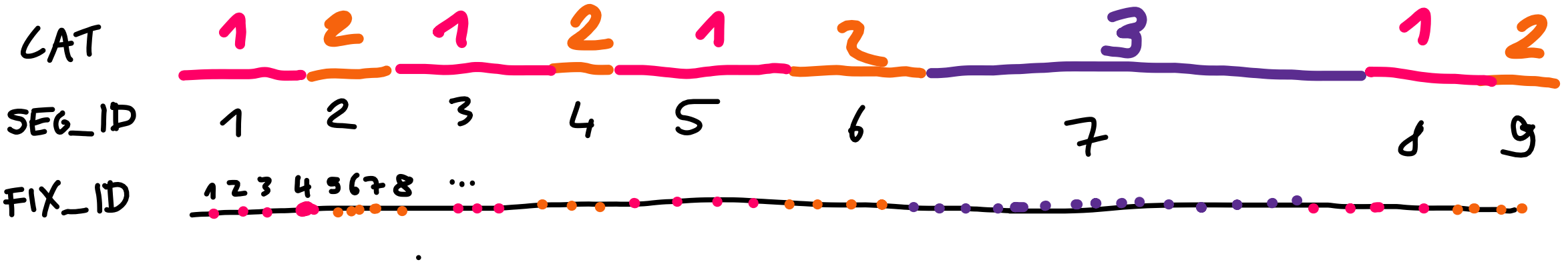
Further Steps

- Enriching with Infrastructure Data (not Intersection but Neighbourhood analysis)
- K means application **km <- kmeans(df, centers = 4, nstart = 25)**
- Comparison



Feedback & Discussion

Segments vs. fixes – what is more appropriate?



SEG_ID	GEOM	Avg_speed	CAT
1	Line(xxx)	5	walking
2	Line(xxx)	10	running
3	Line(xxx)	25	train

Fix_ID	GEOM	Avg_speed	CAT
1	POINT(xxx)	(slide?)	walking
2	POINT(xxx)	(slide?)	running
3	POINT(xxx)	(slide?)	train

Sources

- Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Vachon, E., (2018). Com- bining Bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data. Joint European Conference on Machine Learning and Knowledge Discovery in Databases 569–584.
- Sadeghian, P., Zhao, X., Golshan, A., & Håkansson, J. (2022). A stepwise methodology for transport mode detection in GPS tracking data. Travel Behaviour and Society, 26, 159-167.
- Shamoun-Baranes, J., Bom, R., van Loon, E. E., Ens, B. J., Oosterbeek, K., & Bouten, W. (2012). From sensor data to animal behaviour: an oystercatcher example. *PloS one*, *7*(5), e37997.
- Laube, P., & Purves, R. S. (2011). How fast is a cow? Cross-scale analysis of movement data. Transactions in GIS, 15(3), 401-418.

Further questions:

- Join of table from trajectory and the infrastructure data shows empty colums. Why? Already answered. But: possibility to replace NAs with something meaningful, that k means doesnt take into account?

id1	kategorie	teilplan
NA	*NA*	*NA*
NA	*NA*	*NA*
NA	*NA*	*NA*
NA	*NA*	*NA*
NA	*NA*	*NA*
NA	*NA*	*NA*
NA	*NA*	*NA*

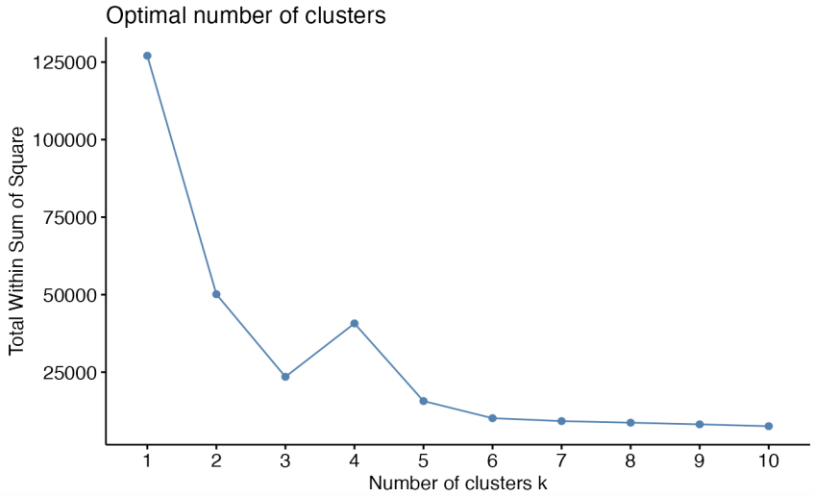
- Involving decision tree for manual analysis?

- How to treat this k means result? What is this?

km	list [9] (S3: kmeans)	List of length 9
cluster	integer [19300]	2 2 2 2 2 2 ...
centers	double [4 x 1]	1.25e+06 3.63e+00 2.68e+06 2.68e+06
totss	double [1]	2.18446e+16
withinss	double [4]	1.56e+09 2.76e+05 1.74e+08 1.42e+08
tot.withinss	double [1]	1872333455
betweenss	double [1]	2.18446e+16
size	integer [4]	3860 11580 1362 2498
iter	integer [1]	2
ifault	integer [1]	0

- How to calculate sinuosity? Trajs doesnt work?

K means



1. Find optimal number of clusters with `fviz_nbclust(df, kmeans, method = "wss")`

Usually, it is where the curve "knicks" in this case at k=3

2. create a df that has only the variables that should be considered (like velocity, sinuosity, Strassenbelag etc)

```
library(factoextra)
library(cluster)
```

3. apply k means:

`km <- kmeans(df, centers = 4, nstart = 25)`

N start is objects per cluster(?)

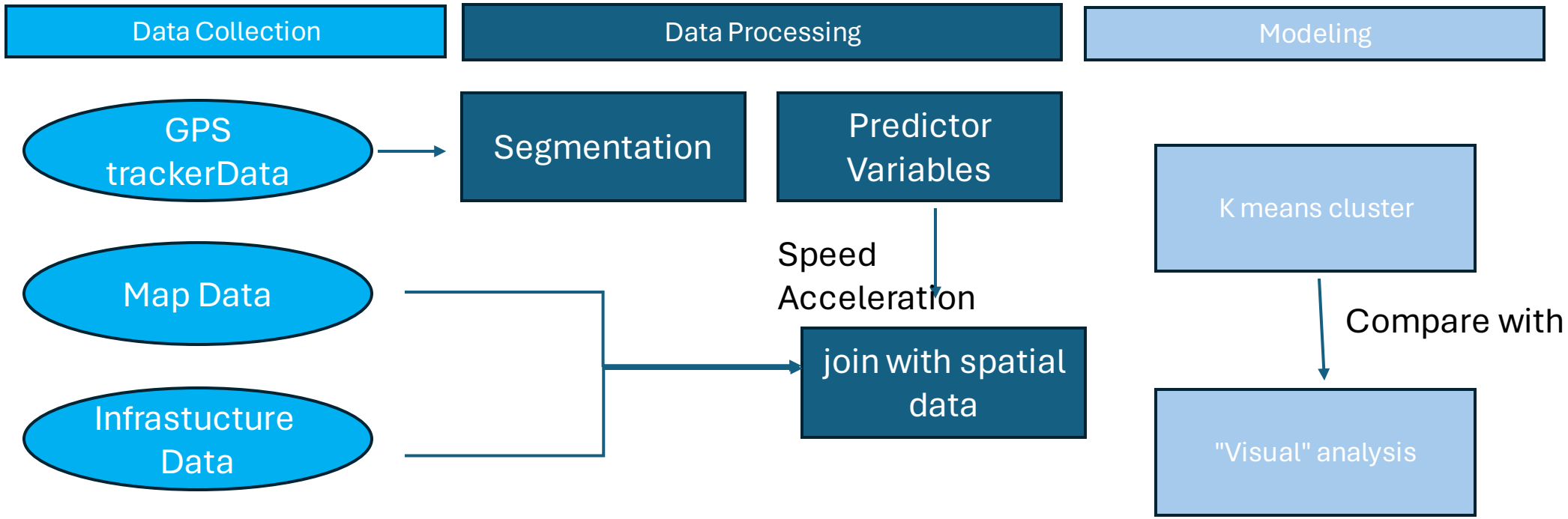
But: cannot put NAs in k means, therefore the category infrastructure cannot be included...

km	list [9] (S3: kmeans)	List of length 9
cluster	integer [19300]	2 2 2 2 2 2 ...
centers	double [4 x 1]	1.25e+06 3.63e+00 2.68e+06 2.68e+06
totss	double [1]	2.18446e+16
withinss	double [4]	1.56e+09 2.76e+05 1.74e+08 1.42e+08
tot.withinss	double [1]	1872333455
betweenss	double [1]	2.18446e+16
size	integer [4]	3860 11580 1362 2498
iter	integer [1]	2
ifault	integer [1]	0

I think for k means we need a data frame that actually contains the segments, not the points.

- https://www.statology.org/k-means-clustering-in-r/

Travel Mode Detection Methods



Travel Mode Detection Methods

Segmentation of trajectories in to segments of 20 points (lag-lead)

Calculation of step mean for 20 points.

Filtering of Outliers (missing signal) with
Filter timelag>1=FALSE

Calculation of predictor variables (velocity, acceleration(?), sinuosity?..?

Segmentation analysis & Context aware Movement analysis

Categorization according to ve locity of segment

Join data with OSM map – verify results

Addition of Strassenbelag to every point (or segment?)

Categorization according to Strassenbelag

K means clustering

Give K means the variables velocity & strassenbelag & the number of classes (e.g. 7)

Comparison

Are the classes encountered in the manual segmentation & k means the same?