Institute of Interactive Systems and Data Science -
Graz University of Technology

Big Data Analysis for Road Accident Risk Prediction in Graz

Michael Jantscher

Supervisors:
Dipl.-Ing. Dr.techn. Roman Kern

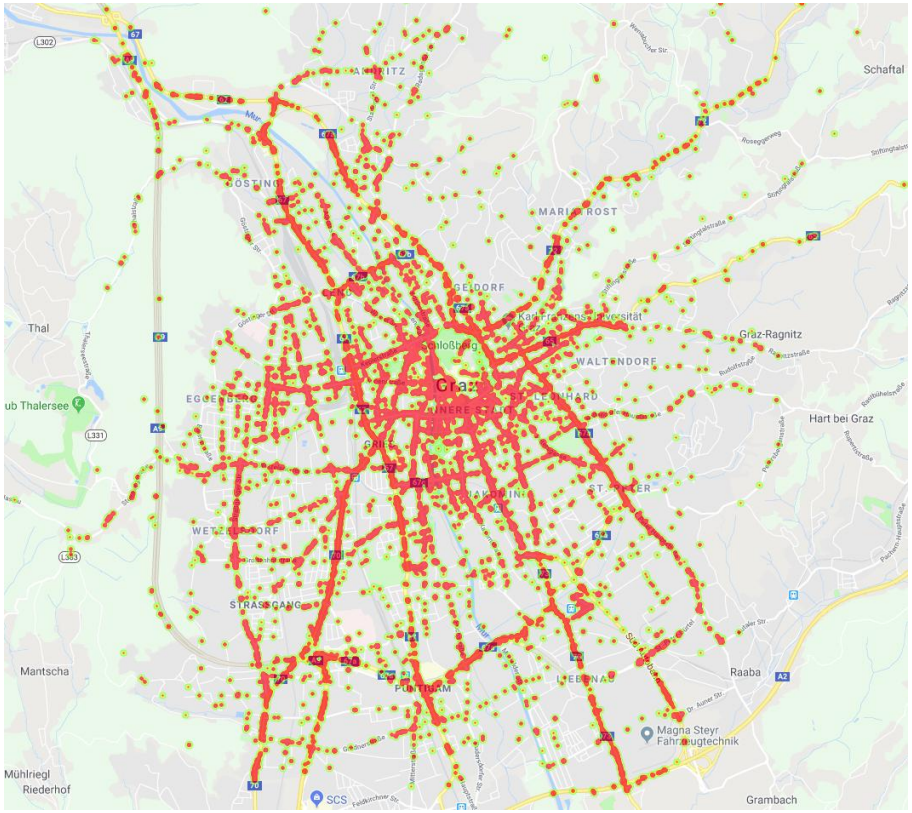
Graz, 19th March 2020

Overview

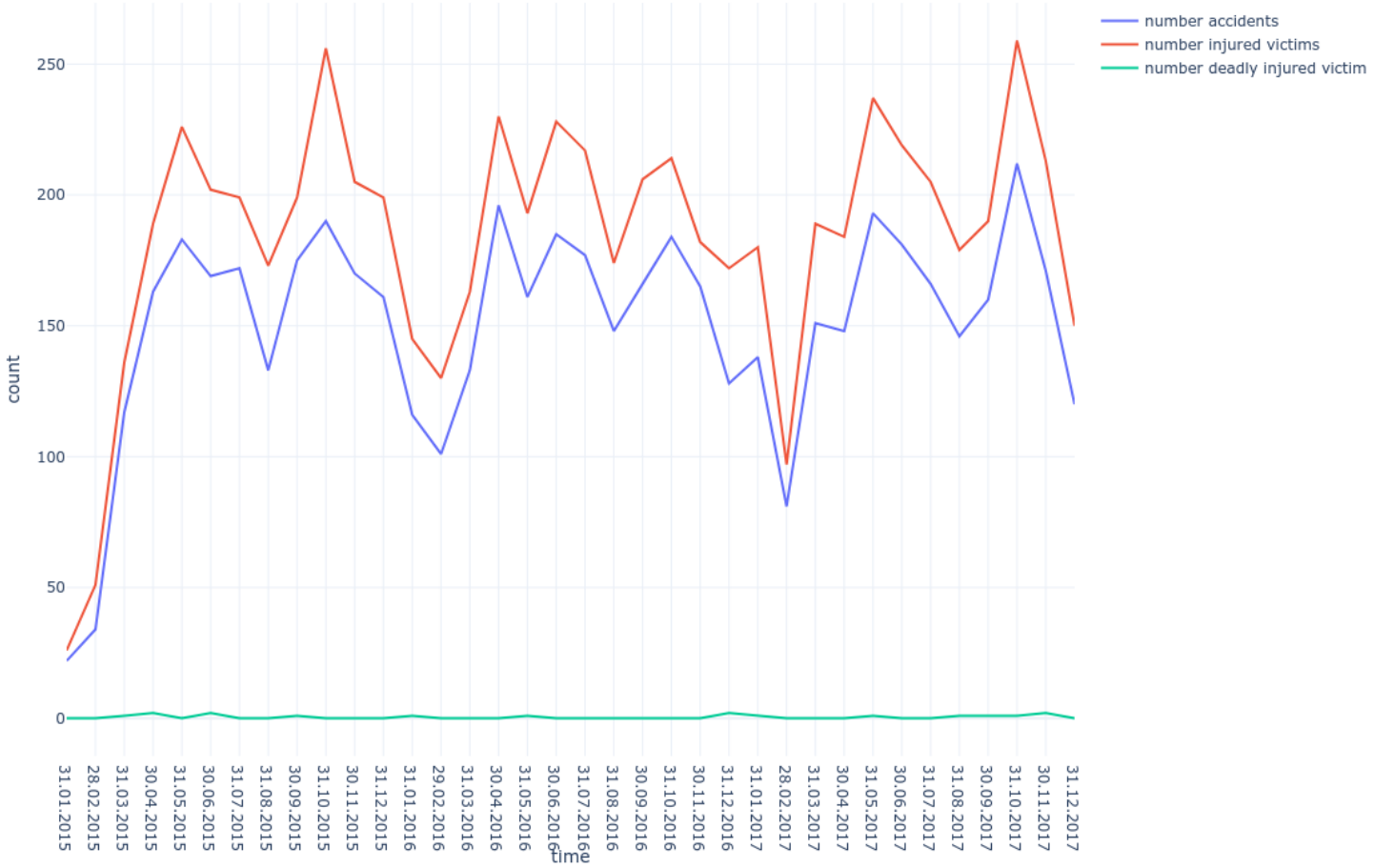
- Previous work
  - Key factors contributing to road accidents
  - Case Study research on temporal and spatial data
  - Accident severity analysis based on the Austrian crash data set
- Goal
  - Exploratory data analysis and statistical tests
  - Missing value imputation of traffic flow data
  - City wide traffic accident likelihood estimation

Statistics

- Tracked by Austrian police officers
- 5416 accidents between 2015 – 2017
- Constant accident rate



Statistics


number accidents
number injured victims
number deadly injured victim
count
time
0
50
100
150
200
250
31.01.2015
28.02.2015
31.03.2015
30.04.2015
31.05.2015
30.06.2015
31.07.2015
31.08.2015
30.09.2015
31.10.2015
30.11.2015
31.12.2015
31.01.2016
29.02.2016
31.03.2016
30.04.2016
31.05.2016
30.06.2016
31.07.2016
31.08.2016
30.09.2016
31.10.2016
30.11.2016
31.12.2016
31.01.2017
28.02.2017
31.03.2017
30.04.2017
31.05.2017
30.06.2017
31.07.2017
31.08.2017
30.09.2017
31.10.2017
30.11.2017
31.12.2017

Datasets

- Vehicle crash data
- Road Network Graphs
  - OpenStreetMap (OSM) [1]
  - Graphenintegrations-Plattform (GIP) [2]
- Population specific data
- Weather data
- Traffic flow

[1] OpenStreetMap https://wiki.openstreetmap.org (Accessed on: 2020-03-08)
[2] GIP http://gip.gv.at (Accessed on: 2020-03-08)

Vehicle Crash Data

- 5416 records between 2015 and 2017
- Attributes:
  - Occurrence location (GPS + Region information)
  - Occurrence time
  - Car specific data
  - Street specific data
  - Weather conditions
  - Injury severity
  - ...

Road Network Graphs

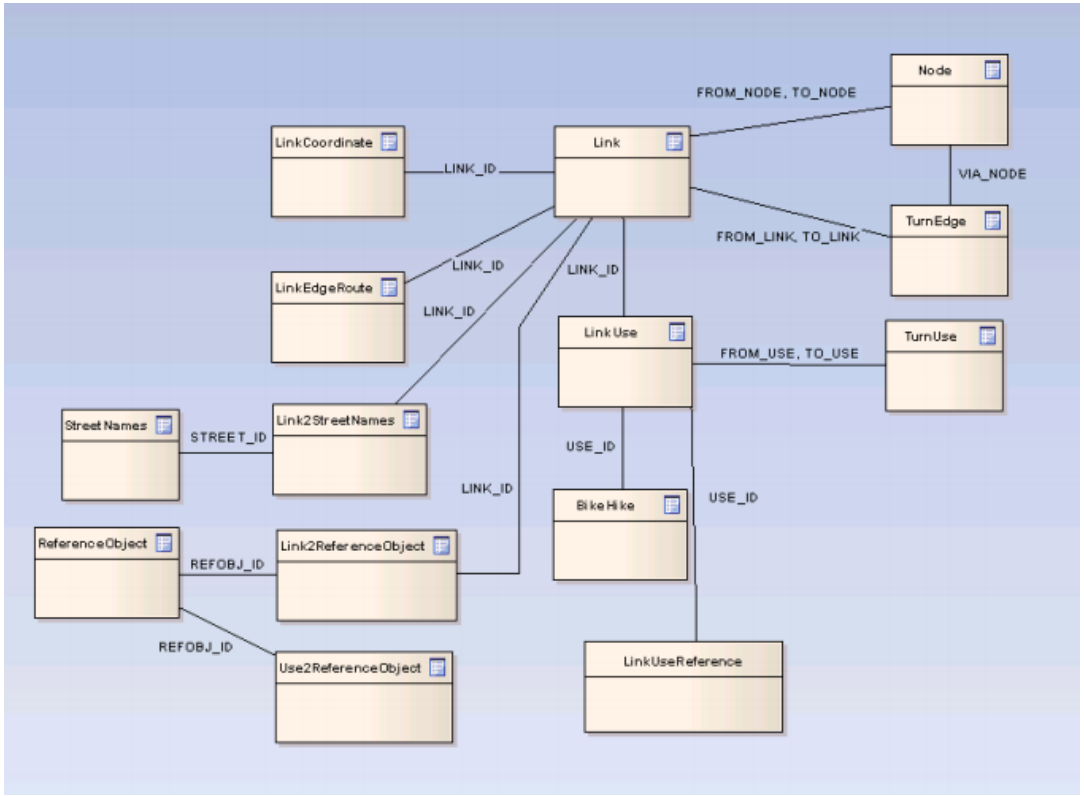
- OpenStreetMap (OSM)
  - OSMNX [3] download of drivable roads in Graz
  - Routable graph

Feature	Description
osmid	unique street vertex identifier
lanes	representing number of lanes
length	representing the street segments length
maxspeed	maximum speed value
name	street name
oneway	True if there is an access restriction for this street segment

[3] Boeing, G. 2017. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." Computers, Environment and Urban Systems 65, 126-139.

Road Network Graphs

- Graphenintegrations-Plattform (GIP)



Source: http://www.gip.gv.at/assets/downloads/1912_dokumentation_gipat_ogd.pdf

Road Network Graphs

- Feature Engineering
  - Closeness Centrality of road links [4]

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

  - Road Curvature

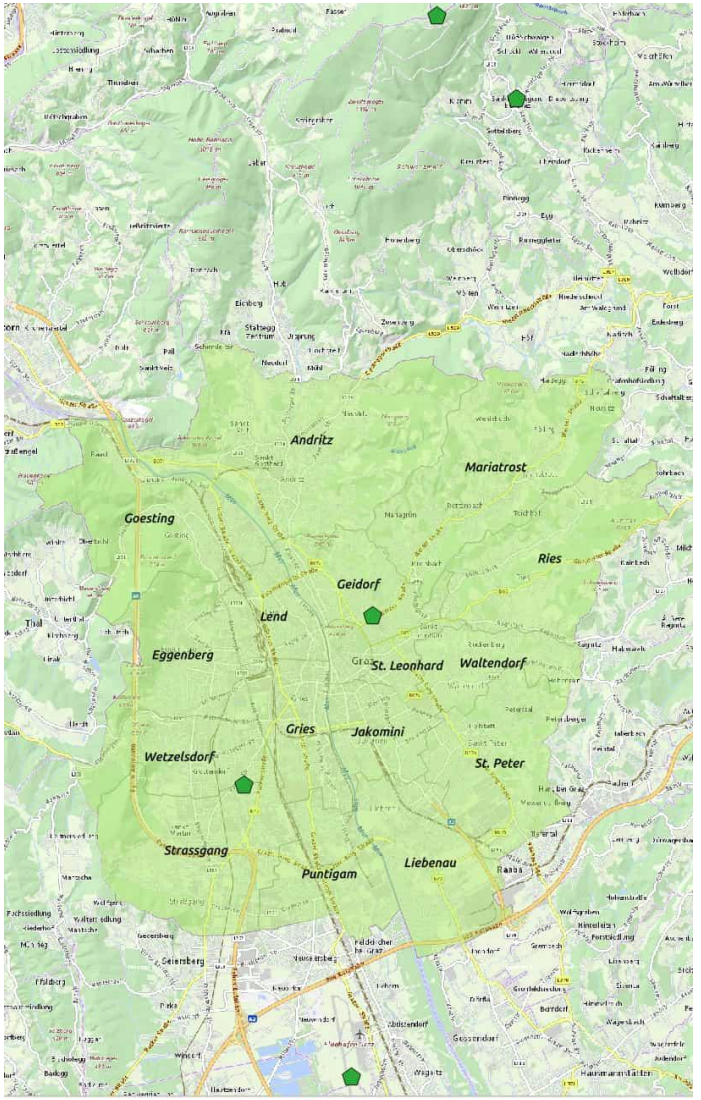
$$curve_{link} = \frac{length(link)}{d(link_{start}, link_{end})}$$

  - Road Slope
  - Junction plateau definition

[4] Linton C. Freeman: Centrality in networks: I. Conceptual clarification. Social Networks 1:215-239, 1979. http://leonidzhukov.ru/hse/2013/socialnetworks/papers/freeman79-centrality.pdf

Weather Data

- ZAMG weather stations
- Temperature and Rainfall
- Match weather data with road links
  - Inverse distance weighting



Population Specific Data

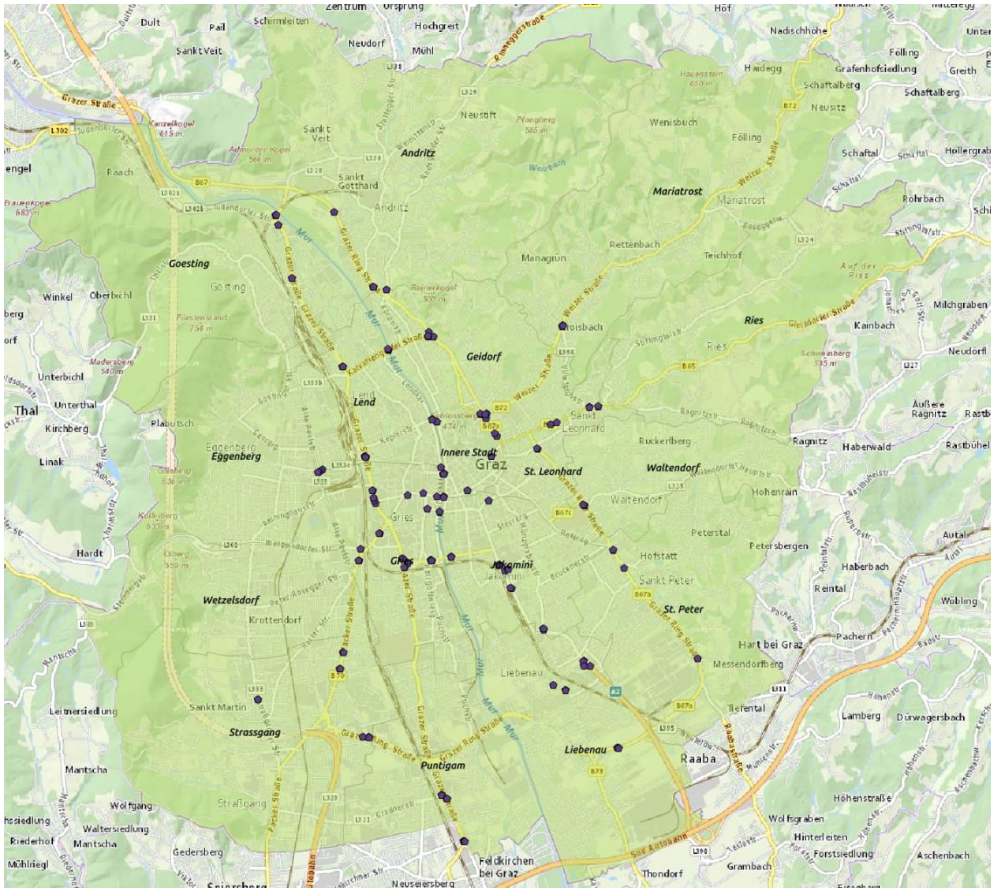
- Open Government Data Austria [5]
  - Population by district and age export
- Population density by district

$$density(district) = \frac{population(district)}{area(district)}$$

[5] Open Data Austria https://www.data.gv.at/ (Accessed on: 2020-03-08)

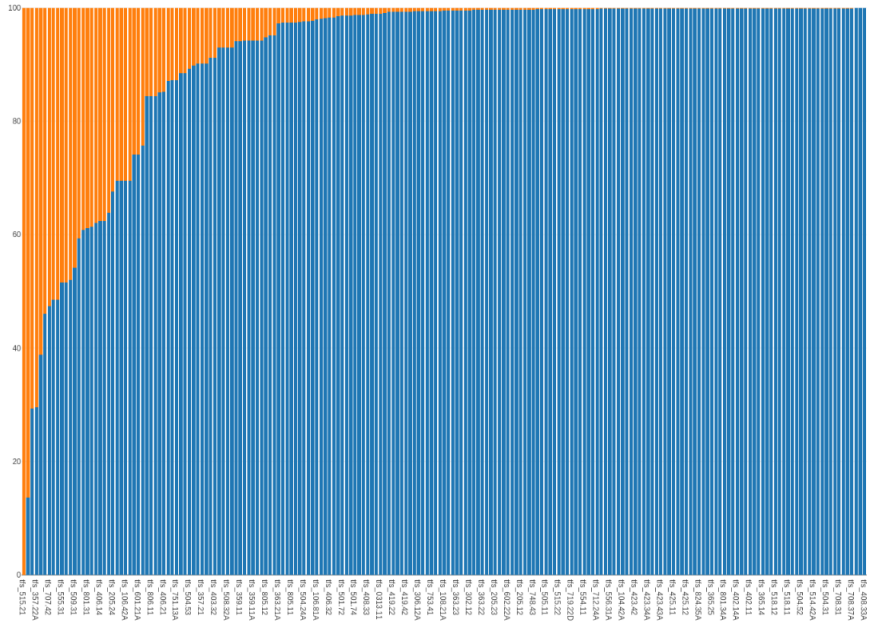
Traffic Flow

- Department of Roads Graz



Traffic Flow Analysis

- Only 15% Missing Values (MV) for more than 170 stations
- Missing Value series
  - 61% MV series are lower than 4 samples
  - Peeks at 26, 84 and 96 consecutive MV
- Univariate vs Multivariate Imputation



Missing Value Imputation

- Split data set per year
- Multiple Imputation by Chained Equation [6] (MICE)
  - Imputation phase
    - Bayes Regression
  - Analysis phase
    - Calculate statistics like mean and variance
  - Pooling phase
    - Calculates the overall estimation of the imputed values

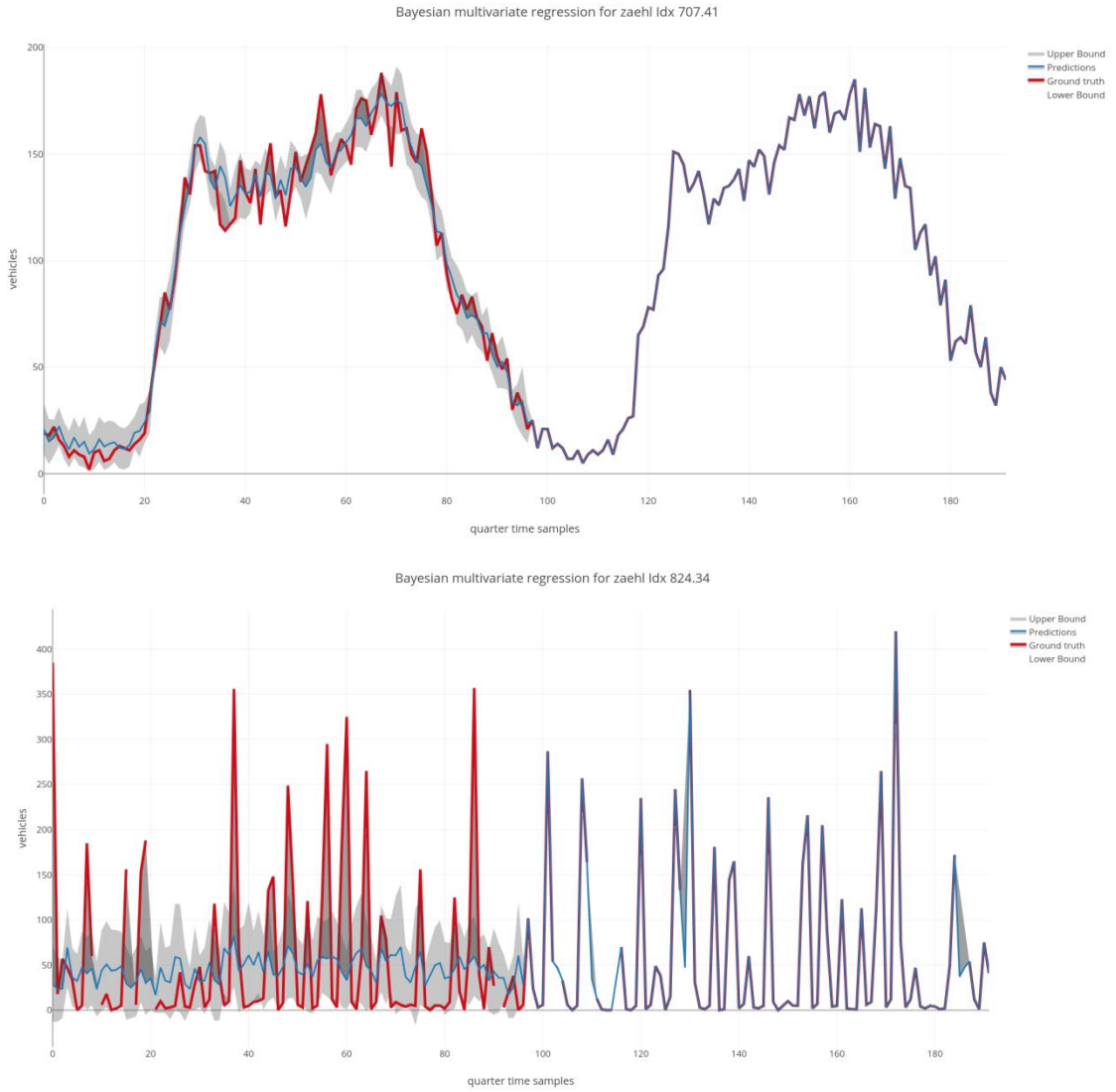
[6] Buuren, S van and Karin Groothuis-Oudshoorn (2010). “mice: Multivariate imputation by chained equations in R.” In: Journal of statistical software, pp. 1–68

Validation

- Validation on each of the three models
- Randomly remove a given percentage of non missing values
- RMSE as validation score
- Stable RMSE by different missing value rates

hyper parameter	model		
	2015	2016	2017
number_nearest_features	20	20	20
max_imputation_cycles	10	10	10
diff_variance_threshold	-	-	-
nr_multiple_imputation	10	10	10

Validation


Bayesian multivariate regression for zaehl Idx 707.41
Upper Bound
Predictions
Ground truth
Lower Bound
vehicles
quarter time samples
Bayesian multivariate regression for zaehl Idx 824.34
Upper Bound
Predictions
Ground truth
Lower Bound
vehicles
quarter time samples

Accident Prediction

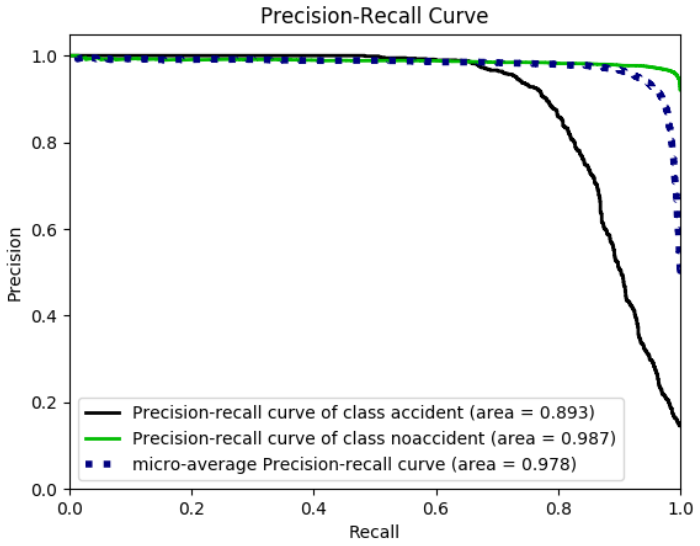
- Imbalanced classification problem
- Negative samples
  - Minority oversampling with matching rules [7]
- With and without sparse, pointwise traffic flow measurements

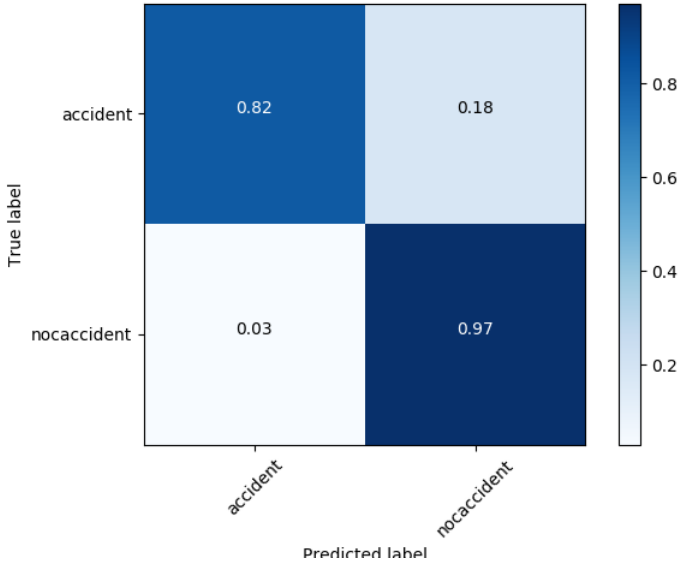
feature name	feature type	details	data source
accident_date	categorical	01.01.2015 - 31.12.2017	Crash data set
accident_time	categorical	24 categories	
is_holiday	categorical	2 categories	
road_condition	categorical	5 categories	
is_junction	categorical	2 categories	GIP graph
max_speed	numerical	0-80 km/h	
average_speed	numerical	0-100 km/h	
number_lanes	numerical	0-6 lanes	
road_width	numerical	0-22 meters	
road_surface	categorical	2 categories	
centrality	numerical	0 - 1	
curvature	numerical	1 - 2	
slope	numerical	0% - 18.56%	
district	categorical	17 categories	Open Data Graz
population_density	numerical	663 - 10423 resident per km^2	
temperature	numerical	−14.3°C - +33.3°C	ZAMG data
precipitation	numerical	0 - 15.1 mm	

[7] Ke, Jintao et al. (2019). “PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data.” In: Trans-portmetrica A: transport science 15.2, pp. 872–895

Accident Prediction

- Gradient Boosting Classifier [8] (XGBoost)
  - Without traffic flow measurements
  - Random Grid Search
    - Hyper parameter search based on the F1 score





[8] XGBoost https://xgboost.readthedocs.io/en/latest/ (Accessed on: 2020-03-08)

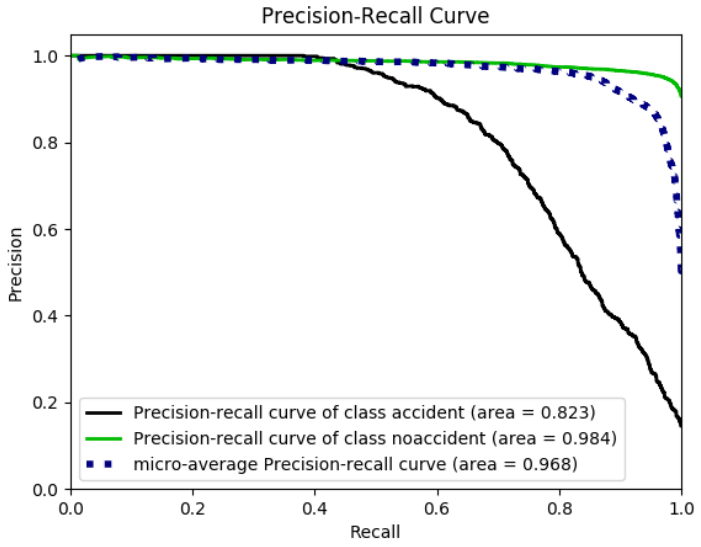
Accident Prediction

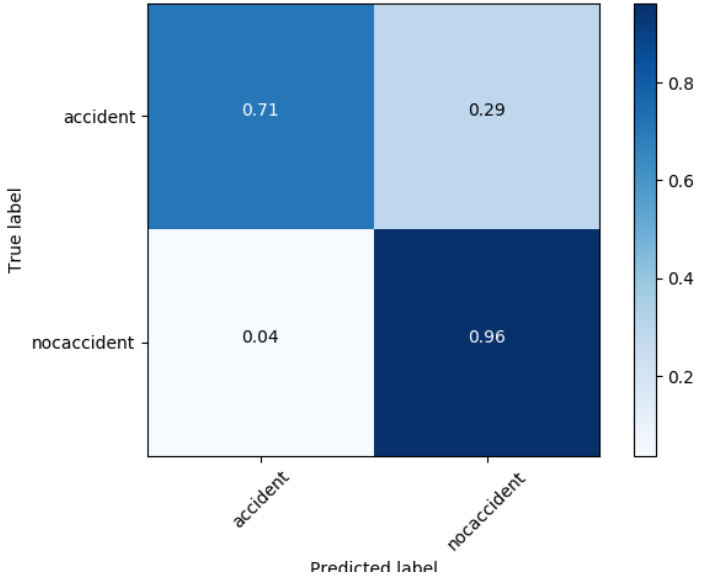
- Feature Importance
  - Gain importance metric
  - Permutation importance / Ablation study
  - F1 score: 0.82

permuted feature	F1 score
datetime	0.73
centrality	0.59
curvature	0.63
slope	0.64
width	0.66
number of lanes	0.69
population density	0.71
district	0.78
precipitation	0.81
temperature	0.82

Accident Prediction

- Gradient Boosting Classifier (XGBoost)
  - With pointwise traffic flow measurements





Conclusion

- Data processing
  - Temporal and spatial data sources
  - Feature Engineering and Map Matching
  - Exploratory data analysis
- Missing value imputation
  - MICE
  - Quality of imputed values depend on flow pattern
- Crash likelihood prediction
  - Negative sampling
  - XGBoost classification
  - Pointwise traffic flow values
- Future Work
  - City wide traffic flow estimation
  - Additional data sources

Institute of Interactive Systems and Data Science -
Graz University of Technology

Big Data Analysis for Road Accident Risk Prediction in Graz

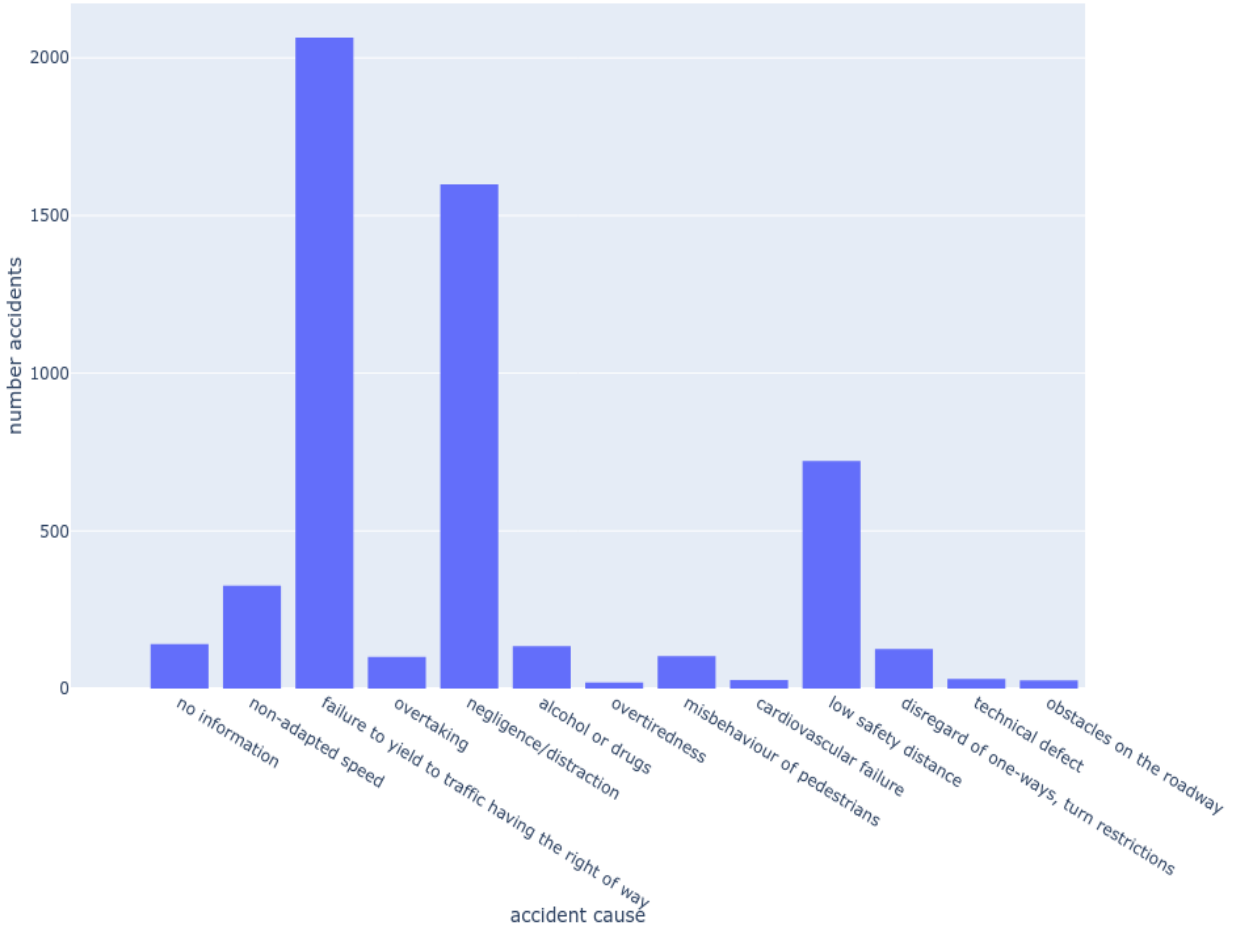
Michael Jantscher

Supervisors:
Dipl.-Ing. Dr.techn. Roman Kern

Graz, 19th March 2020

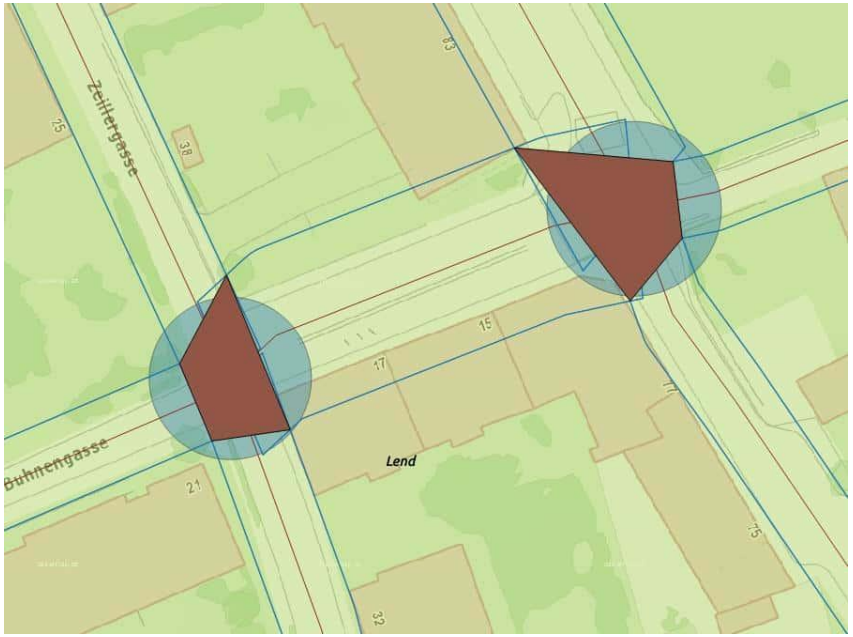
Backup Material

Statistics

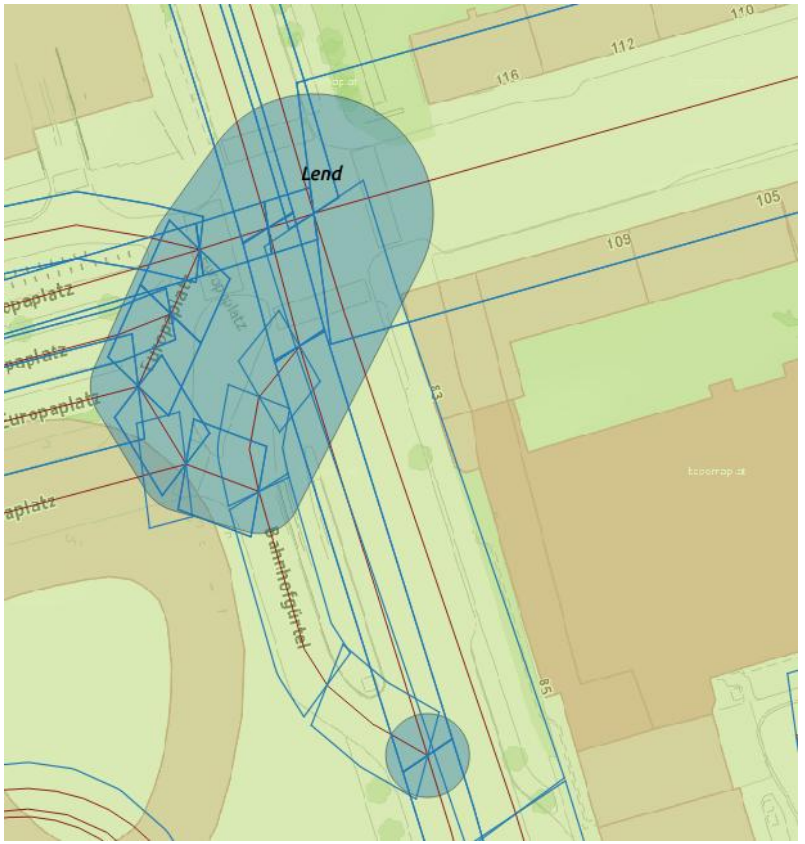

number accidents
2000
1500
1000
500
0
no information
non-adapted speed
failure to yield to traffic having the right of way
overtaking
negligence/distraction
alcohol or drugs
overtiredness
misbehaviour of pedestrians
cardiovascular failure
low safety distance
disregard of one-ways, turn restrictions
technical defect
obstacles on the roadway
accident cause

Junction definition

- Gemäß § 2 Abs 1 Z 17 StVO ist eine Kreuzung eine Stelle, auf der eine Straße eine andere überschneidet oder in sie einmündet, gleichgültig in welchem Winkel. Die Schnittpunkte der gedachten Straßenbaulinien bilden dabei die Eckpunkte des Kreuzungsbereichs und die gedachten Verlängerungen der Straßenbaulinien grenzen den Kreuzungsbereich ab

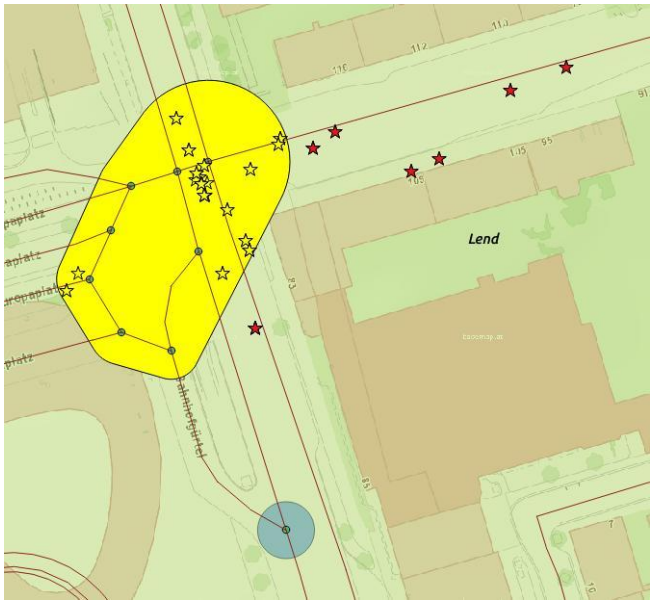


Junction definition



Accident hotspots

GIP ID	road name	number accidents
201005262	Joanneumring	24
201007193	Annenstraße	18
201007177	Keplerstraße	17
201000754	Kärntnerstraße	15
101571709	Lazarettgürtel	14
101374559	Conrad-von-Hötzendorf-Straße	14



Inverse distance weighting

- Spatial interpolation method for high variable data sets

$$u(x) = \sum_{k=1}^{N} \frac{w(x, x_k)}{\sum_{j=1}^{N} w(x, x_j)} u_k$$

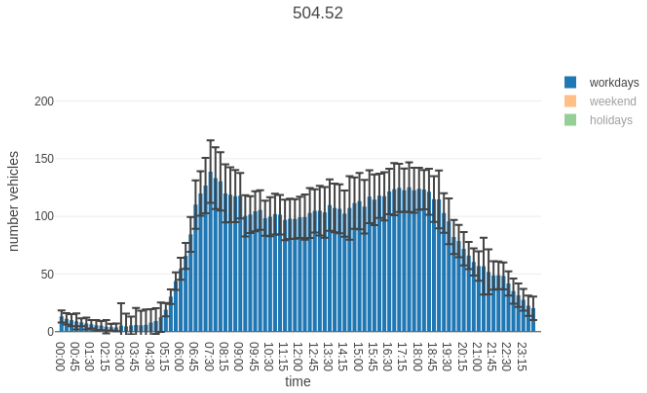
Where $u(x)$ is the interpolation value of a point x based on known samples $u_k = u(x_k)$ for $k = 1, 2, ..., N$.
$w(x, x_k)$ represents the weighting function

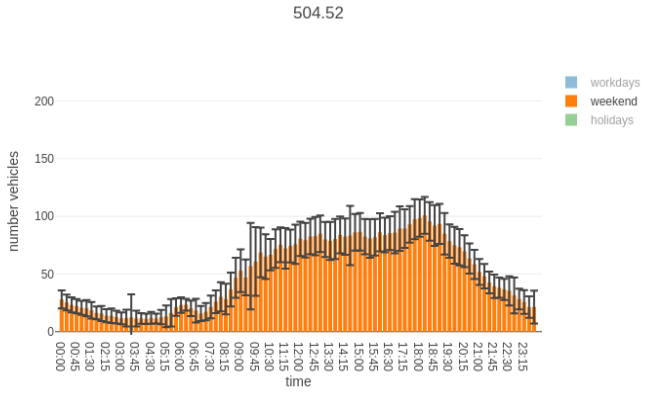
$$w(x, x_k) = \frac{1}{d(x, x_k)^p} \tag{3.2}$$

Where d stands for a given distance metric, typically the Euclidean distance, between x and x_k.

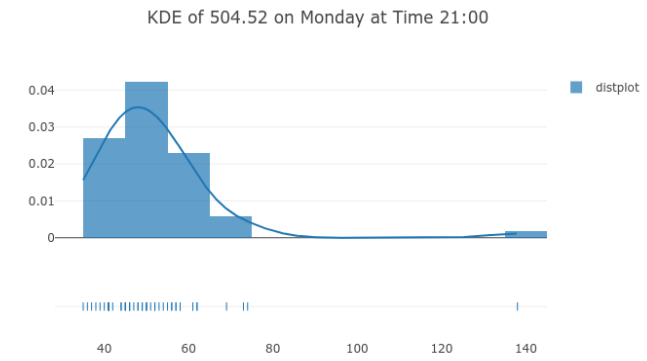
Traffic Flow Analysis

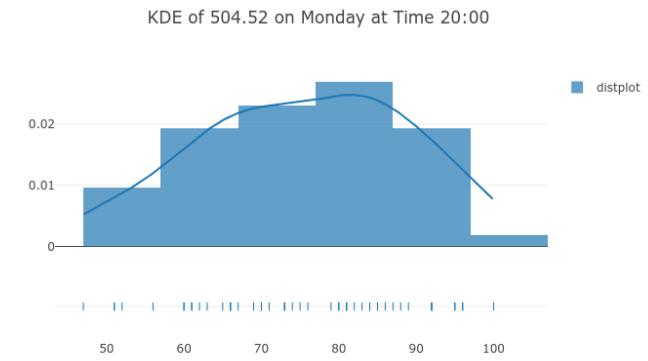
- Workday and weekend pattern





- Distribution over different daily timestamps





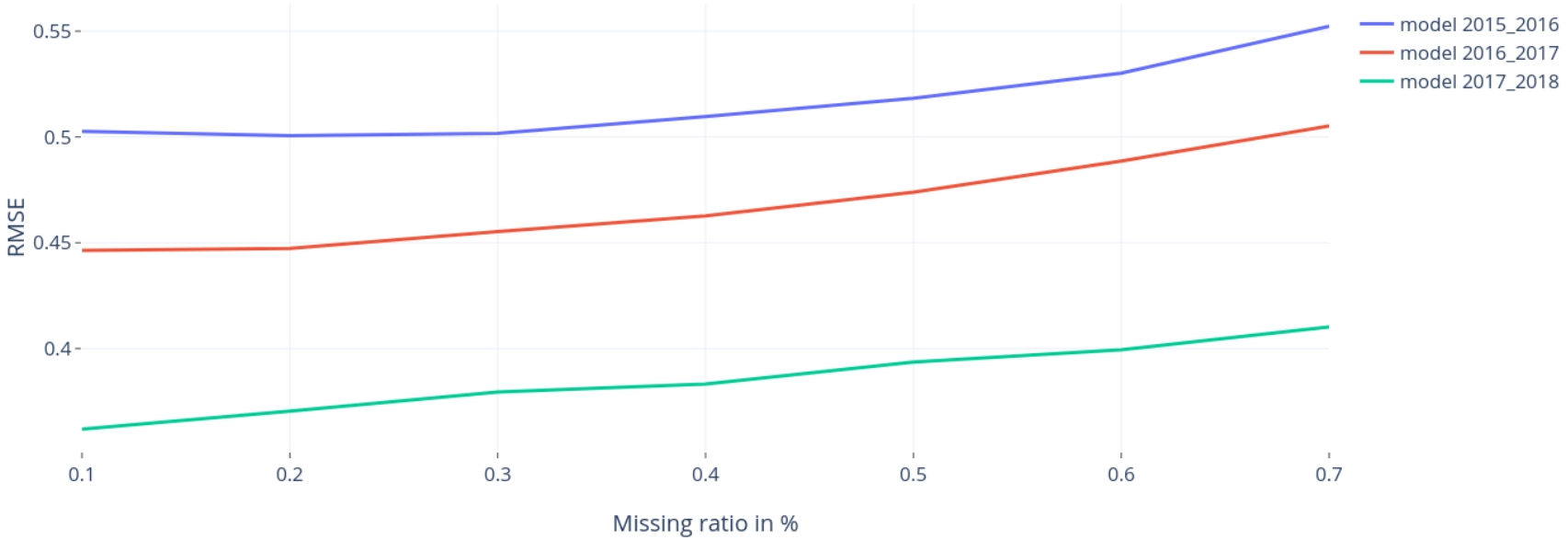
MICE Imputation

- In the first imputation step, all missing values are masked so that for further turns the imputed values still can be recognized and the mean over X_j^{obs} is imputed. These imputation can be thought of as place holders.
- In the second step, the place holder mean imputations for X_j is set back to missing.
- In the third step X_j^{mis} are regressed on X_{-j} via the $BayesRidgeRegression$ approach. Not all X_{-j} features are taken for imputation but just the n strongest correlated feature vectors. The Pearson correlation coefficient is used for estimating these n vectors.
- In the fourth step the missing values X_j^{mis} are replaced with the regressed values.
- In step five step two to step four is repeated for all X_j^{miss} with $(j = 1, ..., p)$
- In step six the imputation process as stated in step 5 is repeated for a given number of cycles with the imputations being updated at each cycle.

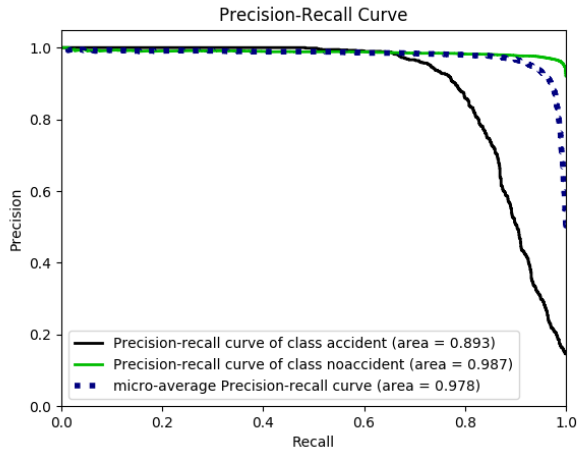
MICE Imputation

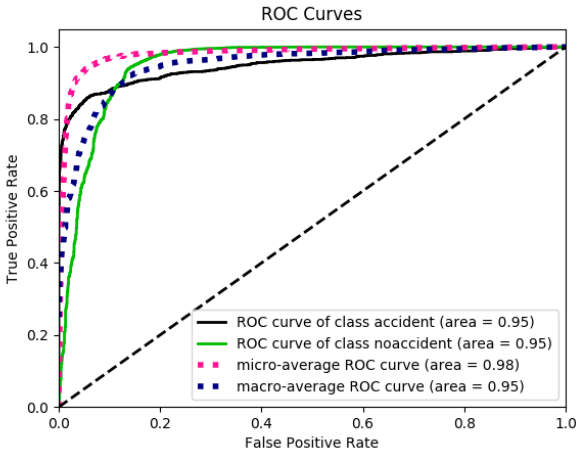
- Estimated Value $\theta_{MI} = \frac{1}{M}\sum_{i=1}^{M}\theta_i$
- Within Variance $Var_{within} = \frac{\sum_{i=1}^{M} SE_i^2}{M}$
- Between Variance $Var_{between} = \frac{\sum_{i=1}^{M}(\theta_i - \theta_{MI})^2}{M-1}$
- Total Variance $Var_{MI} = Var_{within} + (1 + \frac{1}{M})Var_{between}$

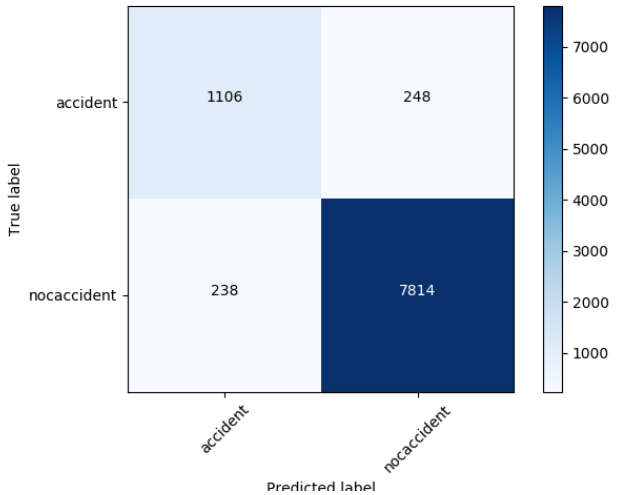
Validation

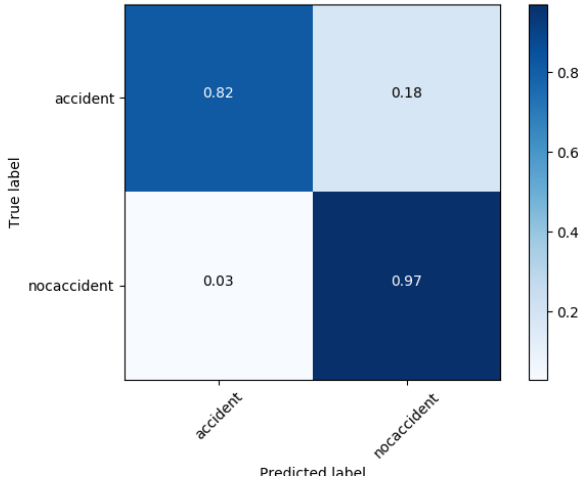

0.55
0.5
0.45
0.4
RMSE
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Missing ratio in %
model 2015_2016
model 2016_2017
model 2017_2018

Accident Prediction









Accident Prediction

- Hyper parameters

hyper-parameter	estimated value
objective	binary:logistic
learning_rate	0.1
max_depth	10
min_child_weight	2
subsample	1
colsample_bytree	0.7
n_estimators	350

Accident Prediction

test set	threshold for class non-accident	precision	recall	accuracy	f1-score
sample over three years	>50%	0.98	0.61	0.94	0.76
	>85%	0.82	0.81	0.95	0.82
year 2017	>50%	0.90	0.63	0.94	0.74
	>85%	0.62	0.82	0.90	0.71

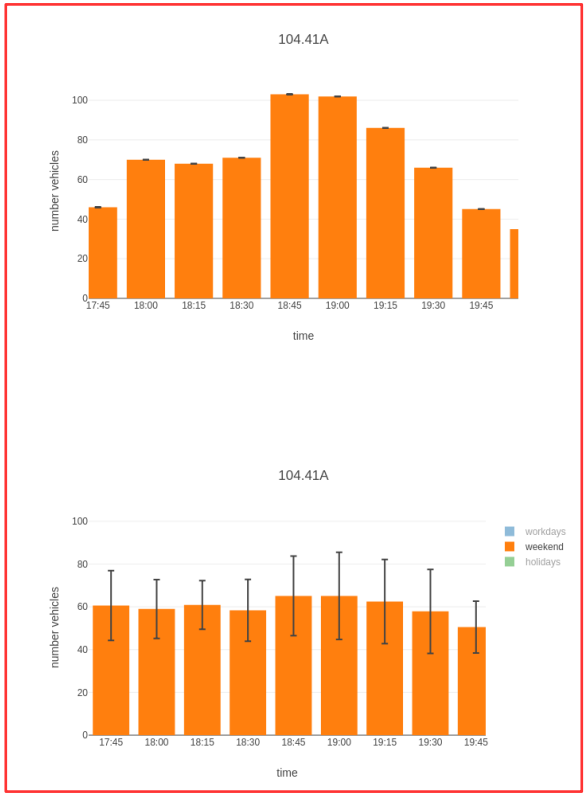
Result without pointwise traffic flow measurements

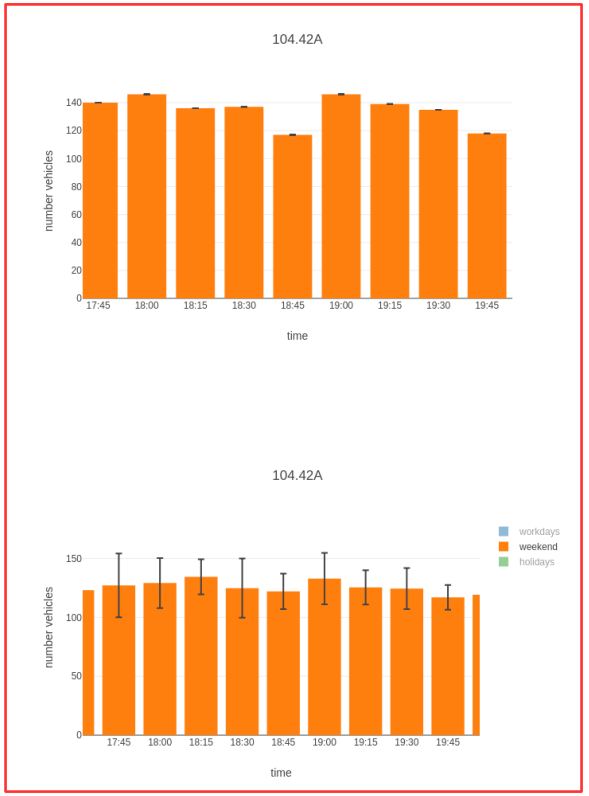
test set	prediction threshold for class non-accident	precision	recall	accuracy	f1-score
sample over three years	>50%	1.0	0.39	0.91	0.53
	>94%	0.76	0.71	0.93	0.74
year 2017	>50%	0.91	0.41	0.91	0.57
	>94%	0.36	0.84	0.76	0.51

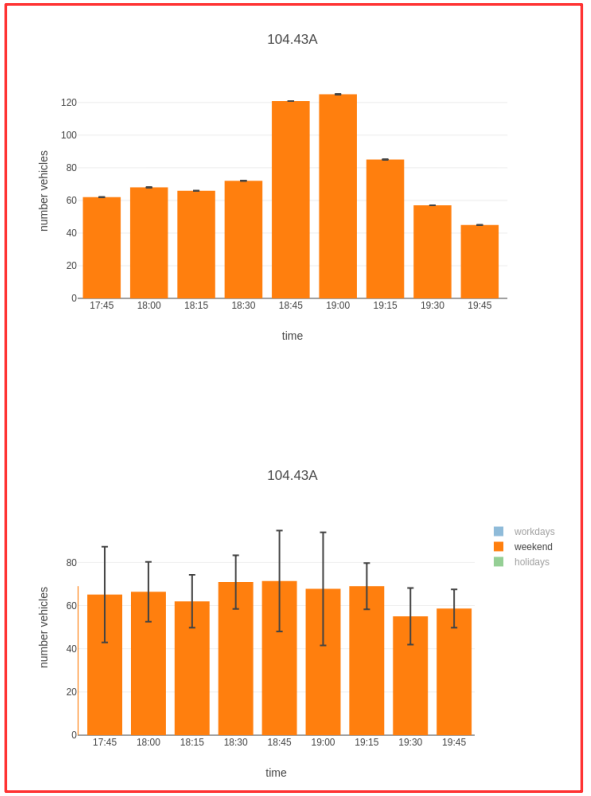
Result with pointwise traffic flow measurements

Accident and Traffic Flow

- Joanneumring
  - One-way street with 3 lanes
- Accident
  - 7th May 2017 at 06:45 p.m.

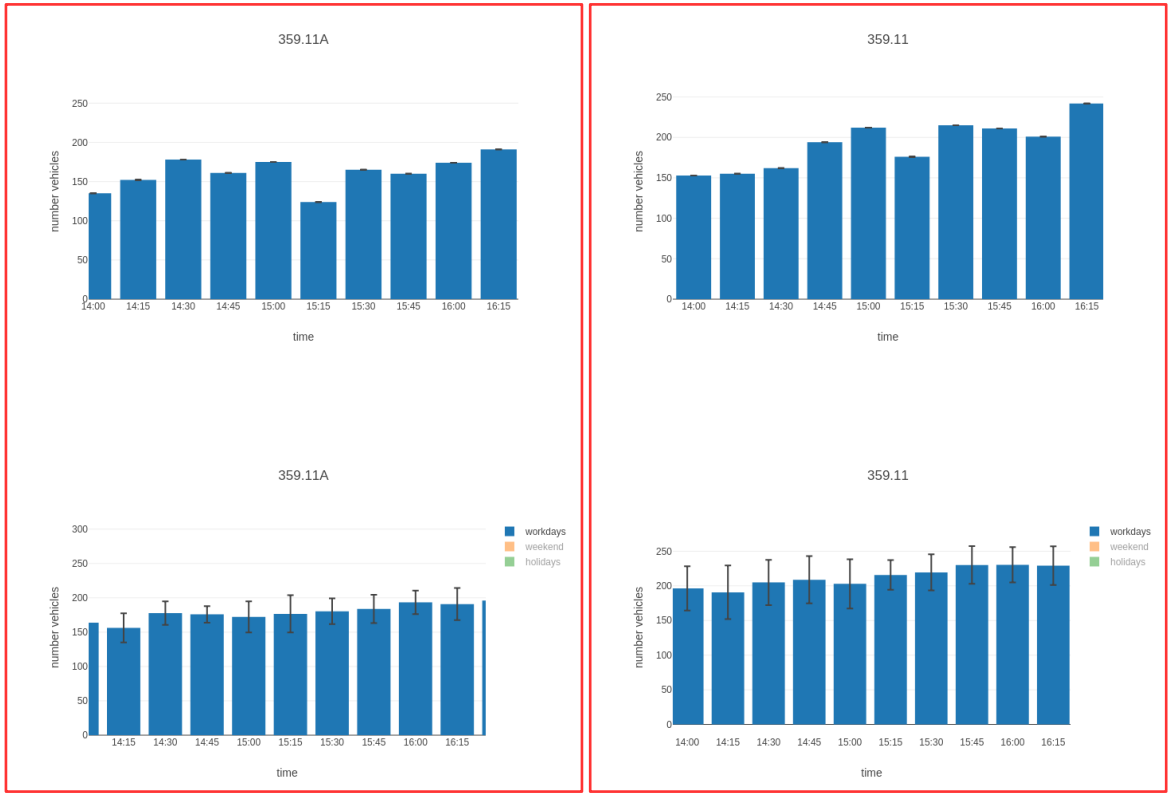






Accident and Traffic Flow

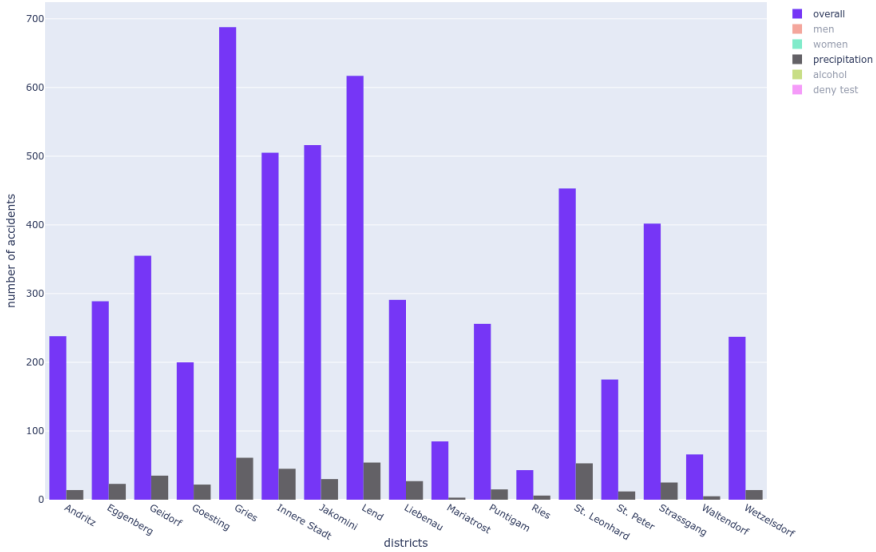
- Weinzöttlstraße
  - One-way street with 2 lanes
- Accident
  - 27th May 2017 at 03:00 p.m.



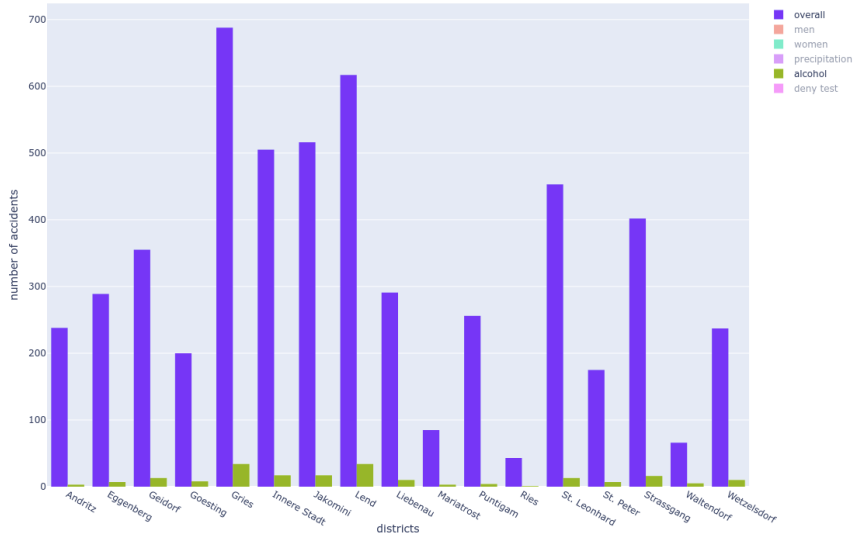
Accident statistics

- Beginning rainfall / snowfall
  - Aggregated in 1 hour intervals
  - Prior hour no precipitation measured
- Statistics
  - 710 accidents between 2015 – 2017
  - 184 accidents by beginning precipitation
  - P(start precipitation) = 2.45%
  - P(start precipitation|accident) = 3.4%

Accident statistics

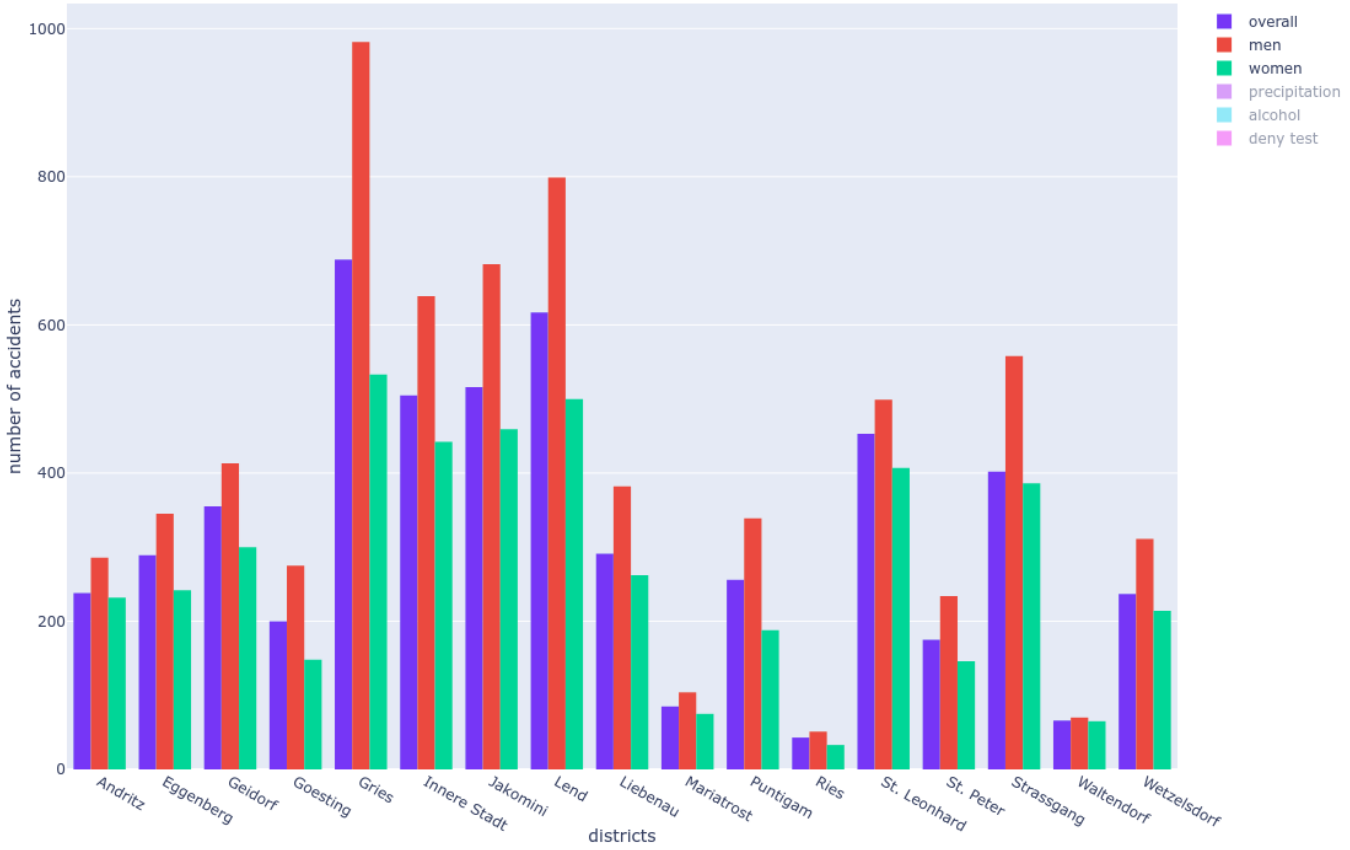


Accidents under precipitation



Accidents under alcohol influence

Accident statistics



Number of accident participants per gender