Introduction to Data Science Project

Accident Analysis in Zurich

David Linder, Marcelo Looser

Content

- Introduction

- Data preparation

- Exploratory analysis

- Classification

- Kernel density estimation

- Questions

Introduction

- Questions that we tried to answer:

    - Which features in the data set are significant when accidents occur, correlations?
    - Can we classify certain features in dependence of a subset of other features?

    - Can we find the underlying probability distribution of the samples?
    - and what do they tell us?

Data preparation

- The given data sets are quiet messy.

- The goal is to clean up all the data sets and save them in destinct tables.

- Link the data sets by certain features.

- With additional merging in a separate table.

Data preparation, accidents in Zurich

- Drop unimportant columns

AccidentUID	AccidentType	AccidentType_de	AccidentType_fr	AccidentType_it	AccidentType_en
A2D2677533867004E0430A865E337004	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
9FD6441F802C20A6E0430A865E3320A6	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
9FDA0DC4856A6094E0430A865E336094	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
A3B66E42396E6000E0430A865E336000	at5	Ãœberqueren der Fahrbahn	accident en traversant une route	Incidente nell'attraversare la carreggiata	Accident when crossing the lane(s)
9FDA0DBE8CCE9096E0430A865E339096	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
9FDA0DC484276094E0430A865E336094	at3	Abbiegeunfall	accident en quittant une route	Incidente nello svoltare	Accident when turning left or right
9FC221265A3BE0F0E0430A865E33E0F0	at2	Auffahrunfall	accident par tamponnement	Incidente di tamponamento	Accident with rear-end collision
9FC221265BADE0F0E0430A865E33E0F0	at1	Ãœberholunfall oder Fahrstreifenwechsel	accident lors d'un dépassement ou lors d'un changement de voie de circulation	Incidente di sorpasso o al cambiamento di corsia	Accident when overtaking or changing lanes
9FD6BC1FBAD4D0FEE0430A865E33D0FE	at7	Parkierunfall	accident en parquant	Incidente nel parcheggiare	Accident when parking

AccidentSeverityCategory	AccidentSeverityCategory_de	AccidentSeverityCategory_fr	AccidentSeverityCategory_it	AccidentSeverityCategory_en	AccidentInvolvingPedestrian	AccidentInvolvingBicycle	AccidentInvolvingMotorcycle	RoadType	RoadType_de
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as3	Unfall mit Leichtverletzten	accident avec blessés légers	Incidente con feriti leggeri	Accident with light injuries	FALSCH	WAHR	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt439	andere
as3	Unfall mit Leichtverletzten	accident avec blessés légers	Incidente con feriti leggeri	Accident with light injuries	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse

Data preparation, accident data

- Change strings to integer, and change false/true statements to 0/1

	A	B	C	D	E	F
1	AccidentUID	AccidentType	AccidentType_de	AccidentType_fr	AccidentType_it	AccidentType_en
2	A2D2677533867004E0430A865E337004	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
3	9FD6441F802C20A6E0430A865E3320A6	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
4	9FDA0DC4856A6094E0430A865E336094	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
5	A3B66E42396E6000E0430A865E336000	at5	Überqueren der Fahrbahn	accident en traversant une route	Incidente nell'attraversare la carreggiata	Accident when crossing the lane(s)
6	9FDA0DBE8CCE9096E0430A865E339096	at0	Schleuder- oder Selbstunfall	dérapage ou perte de maîtrise	Incidente di sbandamento o per colpa propria	Accident with skidding or self-accident
7	9FDA0DC484276094E0430A865E336094	at3	Abbiegeunfall	accident en quittant une route	Incidente nello svoltare	Accident when turning left or right
8	9FC221265A3BE0F0E0430A865E33E0F0	at2	Auffahrunfall	accident par tamponnement	Incidente di tamponamento	Accident with rear-end collision
9	9FC221265BADE0F0E0430A865E33E0F0	at1	Überholunfall oder Fahrstreifenwechsel	accident lors d'un dépassement ou lors d'un changement de voie de circulation	Incidente di sorpasso o al cambiamento di corsia	Accident when overtaking or changing lanes
10	9FD6BC1FBAD4D0FEE0430A865E33D0FE	at7	Parkierunfall	accident en parquant	Incidente nel parcheggiare	Accident when parking

	G	H	I	J	K	L	M	N	O	P
1	AccidentSeverityCategory	AccidentSeverityCategory_de	AccidentSeverityCategory_fr	AccidentSeverityCategory_it	AccidentSeverityCategory_en	AccidentInvolvingPedestrian	AccidentInvolvingBicycle	AccidentInvolvingMotorcycle	RoadType	RoadType_de
2	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
3	as3	Unfall mit Leichtverletzten	accident avec blessés légers	Incidente con feriti leggeri	Accident with light injuries	FALSCH	WAHR	FALSCH	rt433	Nebenstrasse
4	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt439	andere
5	as3	Unfall mit Leichtverletzten	accident avec blessés légers	Incidente con feriti leggeri	Accident with light injuries	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
6	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
7	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
8	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
9	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse
10	as4	Unfall mit Sachschaden	accident avec dommages matériels	Incidente con danni materiali	Accident with property damage	FALSCH	FALSCH	FALSCH	rt433	Nebenstrasse

Data preparation, accident data

- Merge date objects to one column called Date

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	RoadType_fr	RoadType_it	RoadType_en	AccidentLocation_CHLV95_E	AccidentLocation_CHLV95_N	CantonCode	MunicipalityCode	AccidentYear	AccidentMonth	AccidentMonth_de	AccidentMonth_fr	AccidentMonth_it	AccidentMonth_en
2	route secondaire	Strada secondaria	Minor road	2684605	1245194	ZH	261	2011	1	Januar	janvier	Gennaio	January
3	route secondaire	Strada secondaria	Minor road	2682382	1246980	ZH	261	2011	1	Januar	janvier	Gennaio	January
4	autre	Altro	Other	2682791	1247749	ZH	261	2011	1	Januar	janvier	Gennaio	January
5	route secondaire	Strada secondaria	Minor road	2681199	1247102	ZH	261	2011	1	Januar	janvier	Gennaio	January
6	route secondaire	Strada secondaria	Minor road	2682479	1250690	ZH	261	2011	1	Januar	janvier	Gennaio	January
7	route secondaire	Strada secondaria	Minor road	2683365	1253681	ZH	261	2011	1	Januar	janvier	Gennaio	January
8	route secondaire	Strada secondaria	Minor road	2681841	1249487	ZH	261	2011	1	Januar	janvier	Gennaio	January
9	route secondaire	Strada secondaria	Minor road	2683299	1247929	ZH	261	2011	1	Januar	janvier	Gennaio	January
10	route secondaire	Strada secondaria	Minor road	2682866	1247664	ZH	261	2011	1	Januar	janvier	Gennaio	January

	AD	AE	AF	AG	AH	AI	AJ
1	AccidentWeekDay	AccidentWeekDay_de	AccidentWeekDay_fr	AccidentWeekDay_it	AccidentWeekDay_en	AccidentHour	AccidentHour_text
2	aw406	Samstag	samedi	Sabato	Saturday	0	00h-01h
3	aw406	Samstag	samedi	Sabato	Saturday	1	01h-02h
4	aw406	Samstag	samedi	Sabato	Saturday	2	02h-03h
5	aw406	Samstag	samedi	Sabato	Saturday	2	02h-03h
6	aw406	Samstag	samedi	Sabato	Saturday	3	03h-04h
7	aw406	Samstag	samedi	Sabato	Saturday	4	04h-05h
8	aw406	Samstag	samedi	Sabato	Saturday	4	04h-05h
9	aw406	Samstag	samedi	Sabato	Saturday	5	05h-06h
10	aw406	Samstag	samedi	Sabato	Saturday	13	13h-14h

Data preparation, meteo data

- Drop unimportant rows and columns
- Take the average of the remaining data, with respect to n measurement locations.

	A	B	C	D	E	F	G
1	Datum	Standort	Parameter	Intervall	Einheit	Wert	Status
2	01.01.2011 00:00	Zch_Schimmelstrasse	Hr	h1	%Hr	88.11	bereinigt
3	01.01.2011 00:00	Zch_Schimmelstrasse	RainDur	h1	min	0	bereinigt
4	01.01.2011 00:00	Zch_Schimmelstrasse	T	h1	Â°C	2.79	bereinigt
5	01.01.2011 00:00	Zch_Schimmelstrasse	WD	h1	Â°	329.12	bereinigt
6	01.01.2011 00:00	Zch_Schimmelstrasse	WVv	h1	m/s	0.76	bereinigt
7	01.01.2011 00:00	Zch_Schimmelstrasse	p	h1	hPa	974.22	bereinigt
8	01.01.2011 00:00	Zch_Stampfenbachstrasse	Hr	h1	%Hr	92	bereinigt
9	01.01.2011 00:00	Zch_Stampfenbachstrasse	RainDur	h1	min	0	bereinigt
10	01.01.2011 00:00	Zch_Stampfenbachstrasse	T	h1	Â°C	1.6	bereinigt

Data preparation, counts of cars on the road

- Drop unimportant columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	MSID	MSName	ZSID	ZSName	Achse	HNr	Hoehe	EKoord	NKoord	Richtung	Knun	Kname	An	D1I	D2ID	D3ID	D4ID	MessungDatZeit	LieferDat	AnzF	AnzFahrzeuge
2	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 00:00	03.02.2021		Fehlend
3	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 01:00	03.02.2021	256	Gemessen
4	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 02:00	03.02.2021	186	Gemessen
5	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 03:00	03.02.2021	142	Gemessen
6	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 04:00	03.02.2021	116	Gemessen
7	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 05:00	03.02.2021	67	Gemessen
8	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 06:00	03.02.2021	73	Gemessen
9	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 07:00	03.02.2021	43	Gemessen
10	Z001M001	Unbekannt	Z001	Seestrasse (Strandbad Wollishofen)	Seestrasse	451	Unbekannt	2683009.89	1243936.2	auswärts	789	Badanstalt Wollishofen	1	2	Unbekannt	Unbekannt	Unbekannt	01.01.2012 08:00	03.02.2021	52	Gemessen

Data preparation, counts of pedestrians and bicycles on the road

- Drop unimportant columns. (Note: nans were not dropped)
- Add counts together on both sides of the lane to get the total.
- The measurements were taken every 15 min; get the total for one hour.

|	A	B	C	D	E	F	G	H	I
1	FK_ZAEHLER	FK_STANDORT	DATUM	VELO_IN	VELO_OUT	FUSS_IN	FUSS_OUT	OST	NORD
2	ECO09113499	6	01.01.2011 00:00	1	0			2682873	1245891
3	ECO09113506	12	01.01.2011 00:00	0	0			2681385	1247736
4	ECO09113501	8	01.01.2011 00:00	0	0			2683573	1248545
5	Y2G12102806	54	01.01.2011 00:00	5	1			2684006	1246566
6	ECO09022739	5	01.01.2011 00:00	2	0			2682933	1248821
7	ECO09113507	13	01.01.2011 00:00	0				2682683	1250570
8	Y0410090357	52	01.01.2011 00:00	0	0	0	0	2678956	1250443
9	Y0412032046	53	01.01.2011 00:00	0	0	0	0	2679028	1250674
10	ECO09113502	9	01.01.2011 00:00	0	0			2684578	1251967

Data preparation, merging

- The average temperature and rain durations were added to every accident according to the date.

- The count data sets were merged with respect to the location and the date.

- To do this, we located the accidents within a radius of 200m.

Data preparation, merging

- The average temperature and rain dur-
  ations were added to every accident
  according to the date.

- The count data sets were merged with
  respect to the location and the date.

- To do this, we located the accidents
  within a radius of 200m.
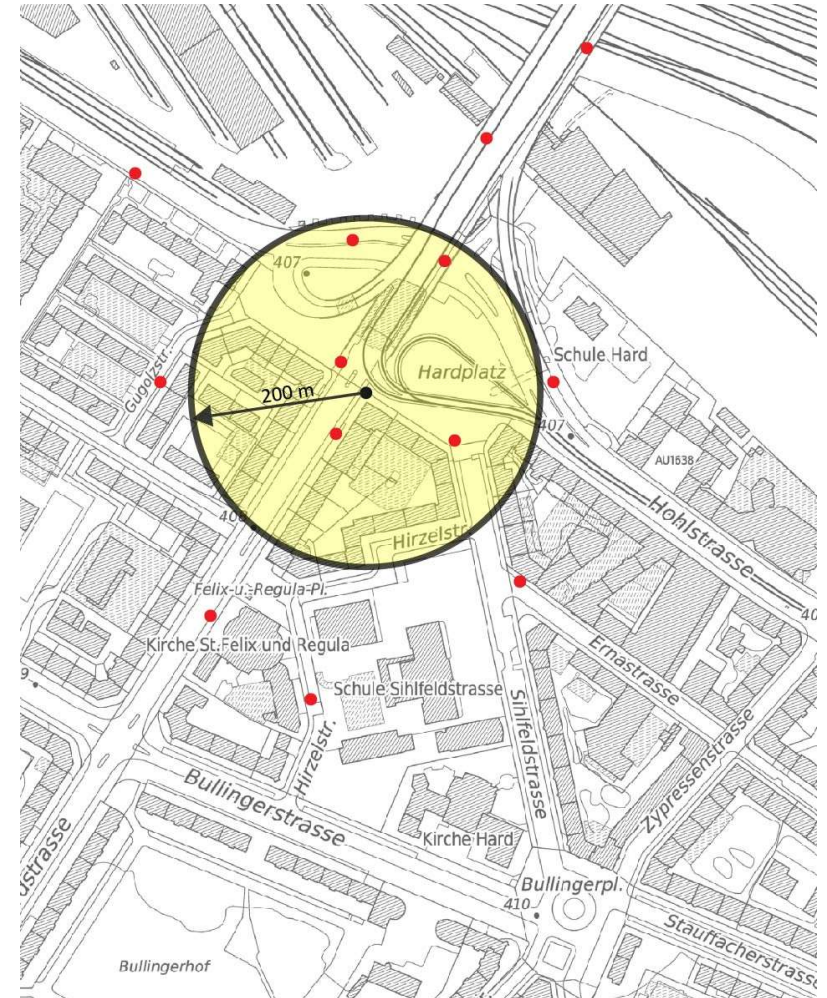
Data preparation

- The merged and cleand data sets are now ready to undergo some analysis.

Date	AccidentType	Severity	Pedestrian	Bicycle	Motorcycle	RoadType	LocationE	LocationN	Temperature	RainDur	SumBiker	SumPedestrian	SumCars
01.01.2011 00:00	0	4	0	0	0	3	2684605	1245194	2.2	0			
01.01.2011 01:00	0	3	0	1	0	3	2682382	1246980	2.3	0	6		
01.01.2011 02:00	0	4	0	0	0	9	2682791	1247749	2.3	0			
01.01.2011 02:00	5	3	0	0	0	3	2681199	1247102	2.3	0			
01.01.2011 03:00	0	4	0	0	0	3	2682479	1250690	2.5	0			
01.01.2011 04:00	3	4	0	0	0	3	2683365	1253681	2.7	0			
01.01.2011 04:00	2	4	0	0	0	3	2681841	1249487	2.7	0			
01.01.2011 05:00	1	4	0	0	0	3	2683299	1247929	2.8	0			
01.01.2011 13:00	7	4	0	0	0	3	2682866	1247664	3.2	0			

Exploratory analysis, technicalities

- To get a grasp of the used data, we used the following modules/methods:

  - Quick matplotlib plots.
  - Seaborn pair plot, to visualize immediate relations between the features.
  - Linear regression via scipy.
  - Kernel density estimation with defaut parameters in seaborn, scipy and sklearn.

- All with the help of pandas DataFrame class.
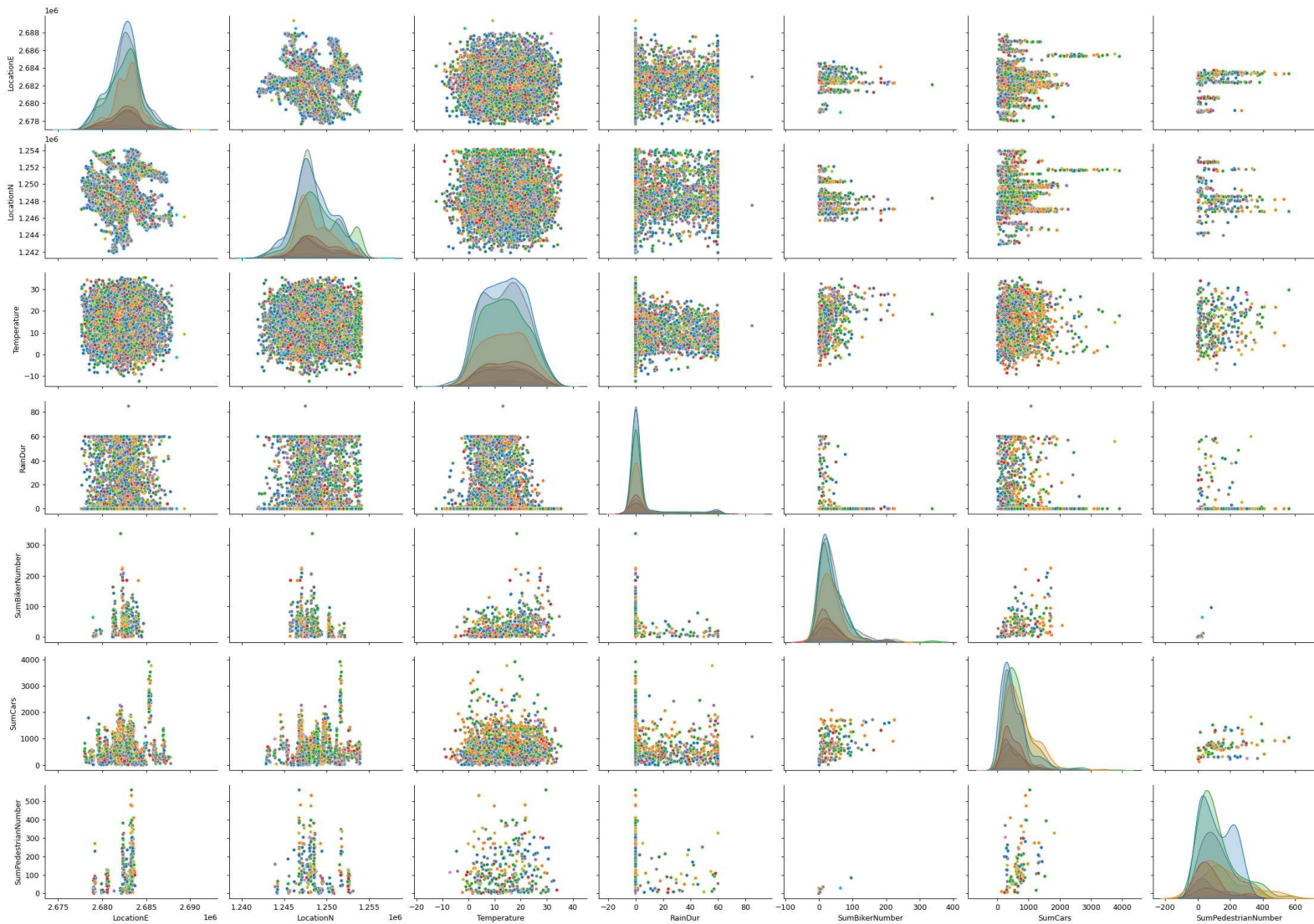
Exploratory analysis

0: Accident with skidding or self-accident

1: Accident when overtaking or changing lanes

2: Accident with rear-end collision

3: Accident when turning left or right

4: Accident when turning-into main road

5: Accident when crossing the lane(s)

6: Accident with head-on collision

7: Accident when parking

8: Accident involving pedestrian(s)

9: Accident involving animal(s)

LocationE LocationN Temperature RainDur SumBicycle SumCar SumPedestrian

SumPedestrian SumCar SumBicycle RainDur Temperature LocationN LocationE

Exploratory analysis

- There is no clear clustering visible.

- We se some trends in the scatter plot, but only trivial ones.

Exploratory analysis

Severity classes:
1: Accident with fatalities
2: Accident with severe injuries
3: Accident with light injuries
4: Accident with property damage

- There is no clear cluster-ing visible.

- The class 4 seems to be dominant.
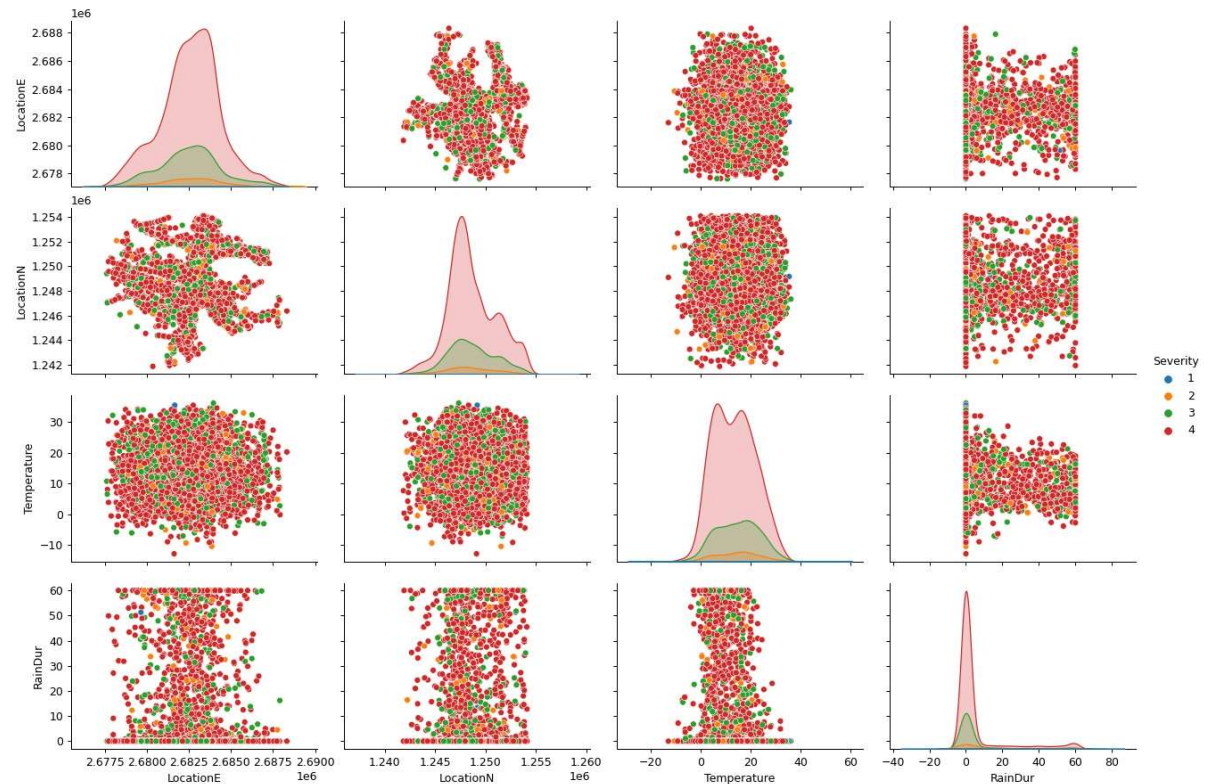
Exploratory analysis

- We tried to find some relations via linear regression of a few non ordinal features.
- But we have not pursued it further.
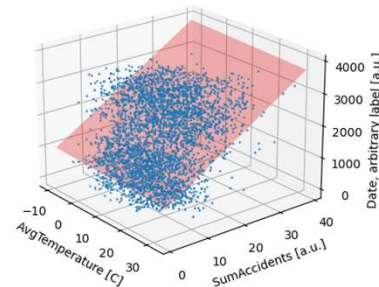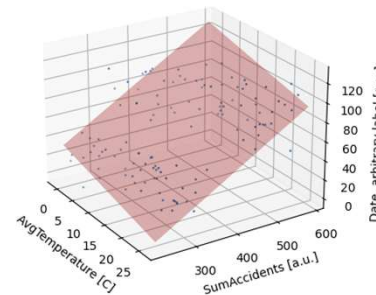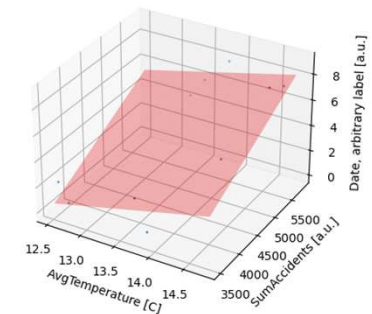
Exploratory analysis

- To see the distribution of the data points in addition with a probability density estimation, we have ploted them on map of Zurich.
- We used seaborn, scipy and sklearn to do this. Note the once on the right are ploted via seaborn and scipy.
- For the exploratory part we only used the default parameters.
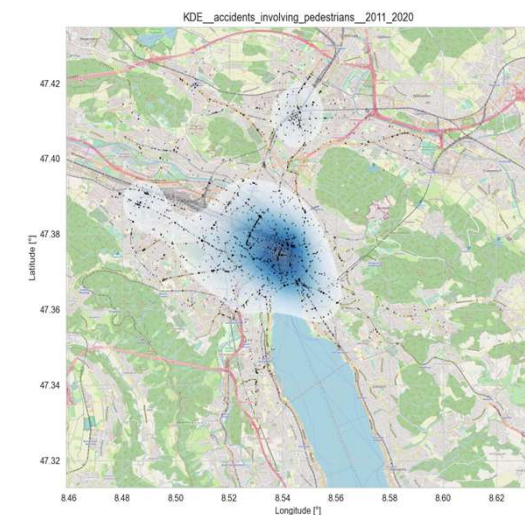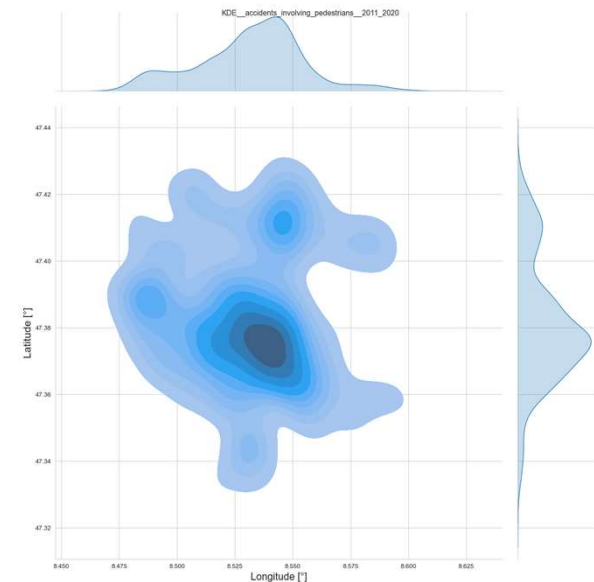
Exploratory analysis

- We decided to use the sklearn implementation. Since we can use the implemented 2-fold cross validation to estimate the best bandwidth.
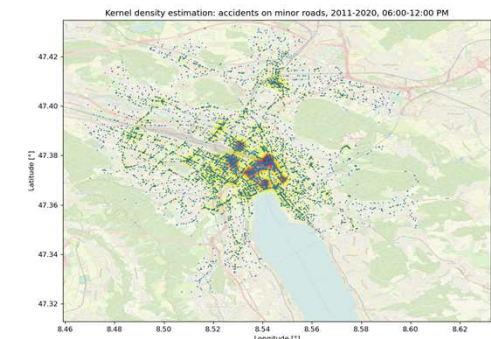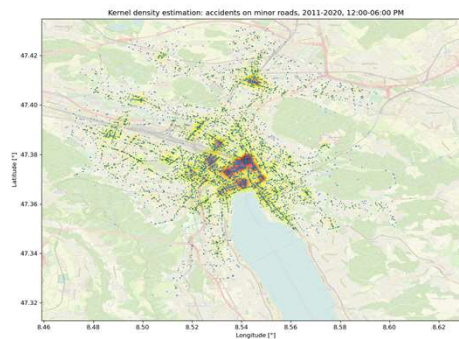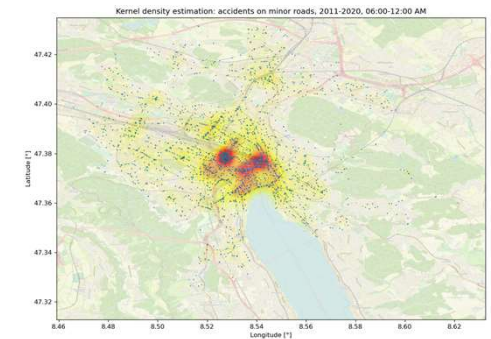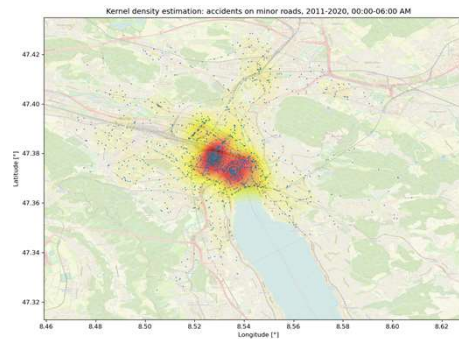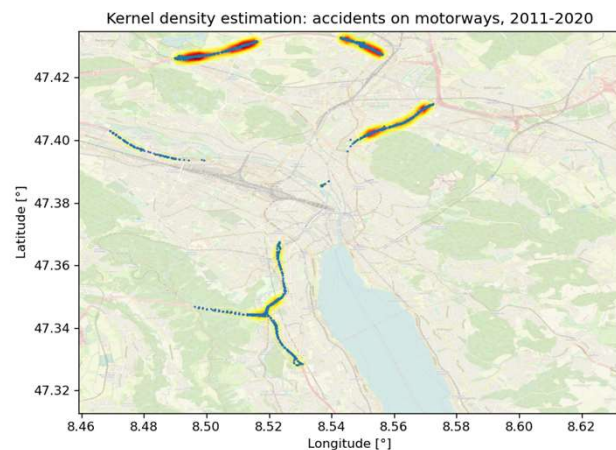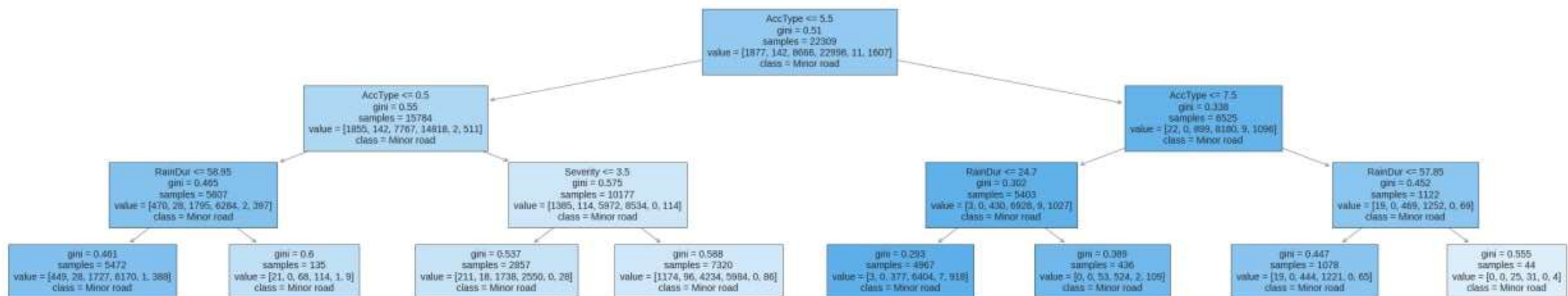
Classifications

- Here we tried to predict three distinct features:
  - Severity
  - Accident type
  - Road type

- This has been done via three different methods:
  - Decision tree
  - Random forest
  - Sequential model in Tensorflow

Classifications

- This is a decision tree with a maximal depth of 3.
- We also tried it with a depth of 5 and no maximal depth.

Classifications

- Most important features used for
  sequential model:

['AccType', 'Severity', 'Pedestrian', 'Bicycle']

```python
model = tf.keras.Sequential([
    feature_layer,
    layers.Dense(128, activation='relu'),
    layers.Dense(128, activation='relu'),
    layers.Dropout(.3),
    layers.Dense(num_units, activation='softmax')
])
optimizer = optimizers.Adam(learning_rate=0.01)
model.compile(optimizer=optimizer,
              loss=tf.keras.losses.SparseCategoricalCrossentropy(),
              metrics=['accuracy'])
```
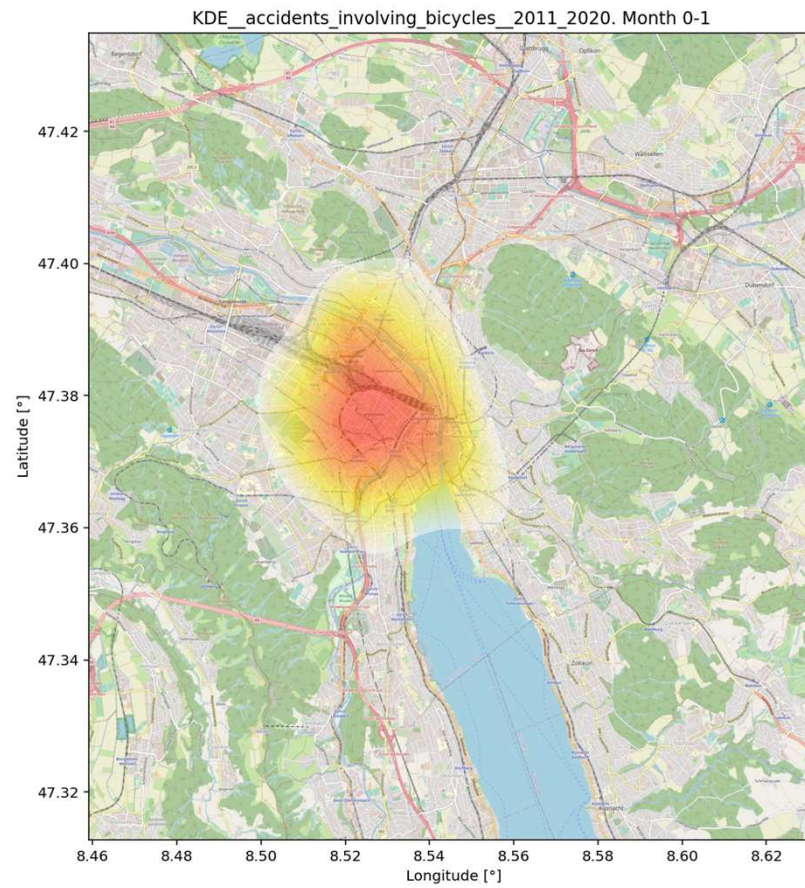
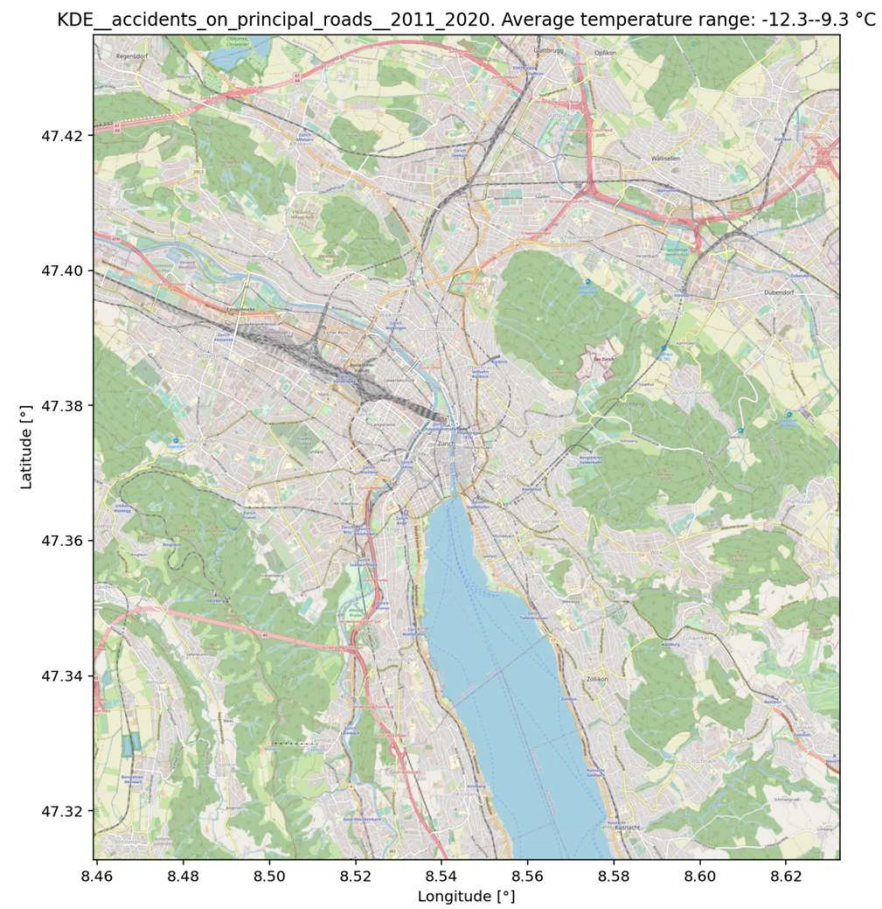| Decision tree (sklearn)			Random forest (sklearn)			Sequential (Tensorflow)	
Target:	RoadType		Target:	RoadType		Target:	RoadType
Accuracy:	0.576		Accuracy:	0.601		accuracy:	0.646
importance			importance			Sequential (Tensorflow)	
Temperature	0.530		Temperature	0.619		Target:	RoadType
RainDur	0.221		RainDur	0.180		accuracy:	0.6517
AccType	0.175		AccType	0.162			
Severity	0.033		Severity	0.019			
Motorcycle	0.020		Bicycle	0.010			
Bicycle	0.016		Motorcycle	0.005			
Pedestrian	0.005		Pedestrian	0.004			

Kernel density estimation

- As mentioned before, 2-fold cross validation to find an estimation of the bandwidth of the kernel density.

- Further, gaussian kernels were used.

- To display them in a nice manner, we used a simple linear interpolation scheme between the calculated frames.

Kernel density estimation

KDE__accidents_involving_bicycles__2011_2020. Month 0-1

Kernel density estimation

KDE__accidents_on_principal_roads__2011_2020. Average temperature range: -12.3--9.3 °C

Questions ?

References

- The git repository can be found here: https://github.com/massstab/ESC403_project.git

- Also the raw data and instructions how to tidy up the data and the codebook with the explanation of all labels can be found there.

- The main accidents dataset can be found here: https://opendata.swiss/de/dataset/polizeilich-registrierte-verkehrsunfalle-auf-dem-stadtgebiet-zurich-seit-2011/resource/3bf3f12a-bf09-4e69-8cde-0df9e268d54b

- The meteo data can be found here: https://data.stadt-zuerich.ch/dataset/ugz_meteodaten_stundenmittelwerte

- The count of the pedestrian/cars/bike data can be found here: https://data.stadt-zuerich.ch/dataset/ted_taz_verkehrszaehlungen_werte_fussgaenger_velo