

Team Tetrahedron

Milestone #4 - Data Analysis

Team members:

Lauro Fialho Müller

Chandan Radhakrishna

Raghava Vinaykanth Mushunuri

Kavya Vajja

Arnab Das

Anjan Chatterjee

26.05.2020

Table of contents

Overview	2
Input variables measured	2
Output variables measured	2
Erich-Weinert-Str	3
Experimental observations and final conclusions	4
Am Fuchsberg	5
Experimental observations and final conclusions	6
Leipziger Str. South	7
Experimental observations and final conclusions	8
Leipziger Str. North	9
Experimental observations and final conclusions	10
Traffic lights	11
Bicycles, pedestrians and trams	12
For Trams	13
Car Moving Direction with Count & Probability	13
Additional measurements	16
Inputs:	16
Outputs:	16
Difficulties encountered while obtaining the data	17
Dealing with trade-off between good estimation of distribution function and variance	17
Limitations on the accuracy or validity of the data	20
Cost overview	21
Future work	22

Overview

The data analysis is conducted this milestone. The main motivation here is to identify the theoretical distributions that best fit the data so that we can use them when building our simulation model. Other elements such as pedestrian and traffic light modelling are also considered.

Input variables measured

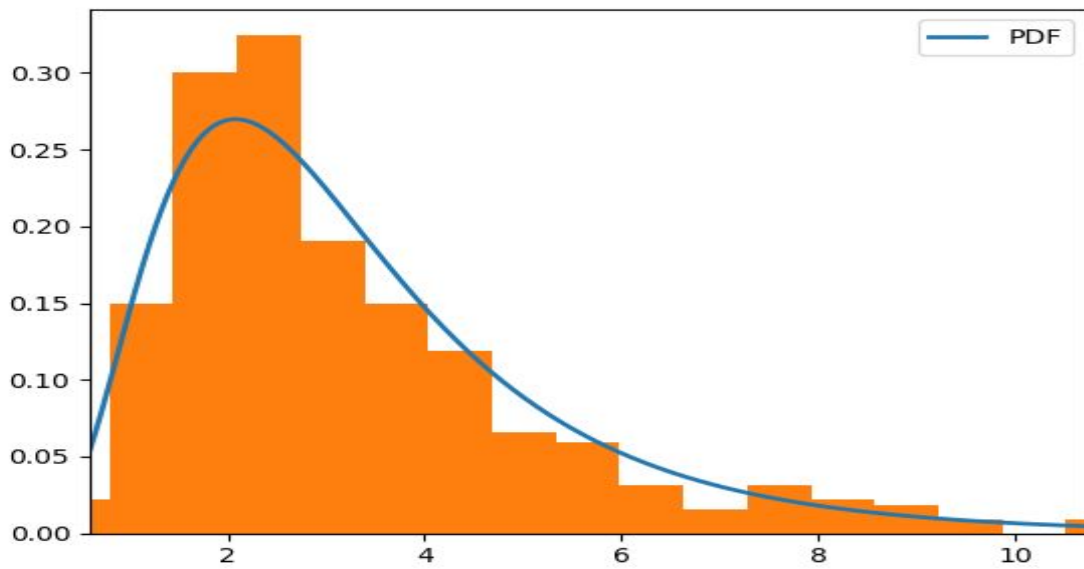
1. Inter arrival times are measured as inputs for estimating the distribution functions.
2. Total bicycles and pedestrians entering the system per hour.

Output variables measured

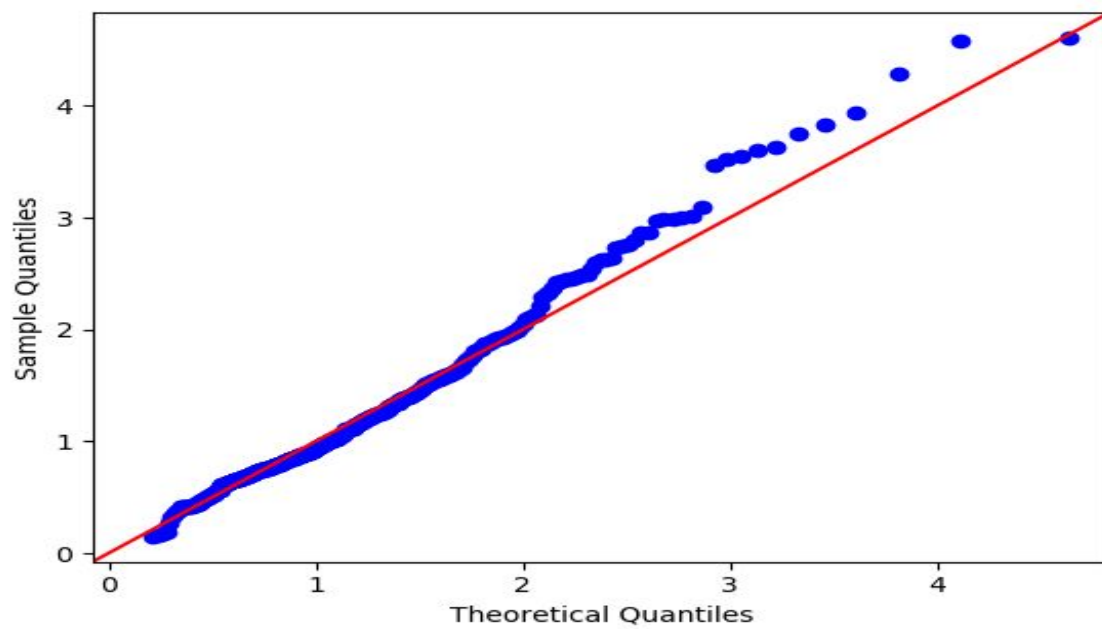
1. Frequency of observed values for k number of bins
2. Expected frequency for each distribution type for k number of bins
3. Inverse values of cumulative distribution function is measured for each distribution type considered.
4. Underlying distribution functions for the vehicles entering each lane of the intersection point.
5. Statistics of each distribution is obtained
6. Chi square values for each distribution type taken at every possible k value
(that is in the range of $k = \sqrt{n}$ to $n/5$, n is total number of samples)
7. Probabilities for the vehicle to take a turn is measured.

Erich-Weinert-Str

Histogram:



QQ plot:



Experimental observations and final conclusions

Observations:

- Distribution type: Lognormal
- Statistics:
 - Mean = 1.161
 - Standard deviation = 0.533
 - For $k = 31$, chi-square calculated less than critical

From the above comparisons we can conclude the test failed to reject the null hypothesis and the observed values and Expected values have correlation and the expected values are not occurring just by a mere coincidence.

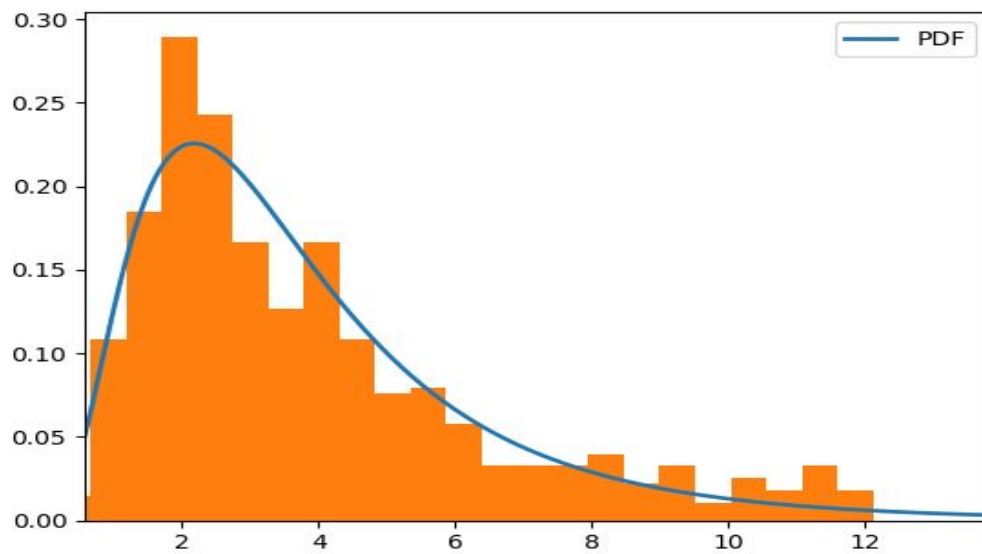
The expected data is following the observed data with 95% confidence level.

Conclusion :

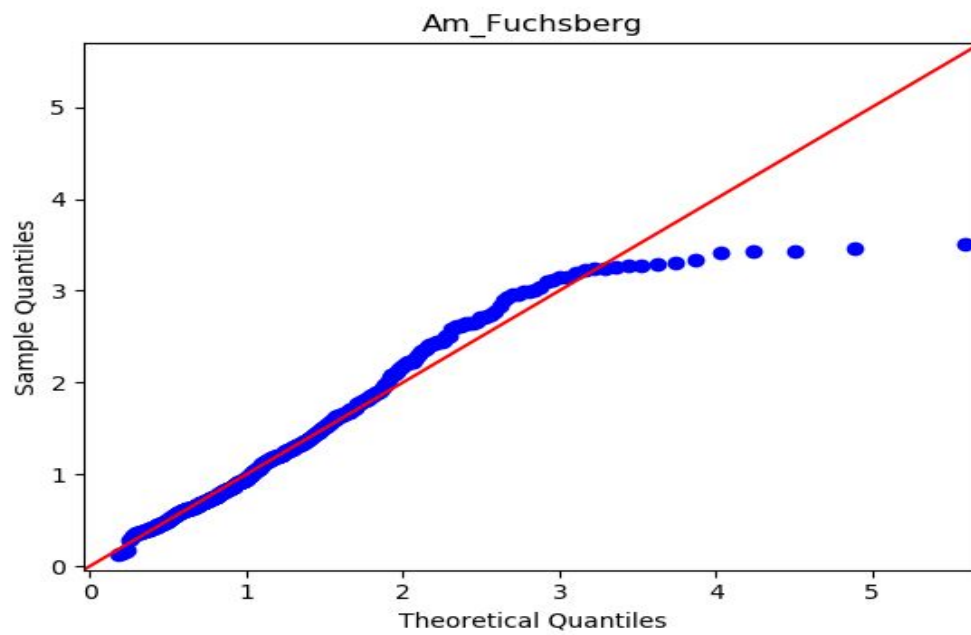
Lognormal with mean = 1.161, standard deviation = 0.533

Am Fuchsberg

Histogram:



QQ plot:



Experimental observations and final conclusions

Observations:

- Distribution type: Lognormal
- Statistics:
 - Mean = 1.267
 - Standard deviation = 0.593
 - For $k = 27$, chi-square calculated less than critical

From the above comparisons we can conclude the test failed to reject the null hypothesis and the observed values and Expected values have correlation and the expected values are not occurring just by a mere coincidence.

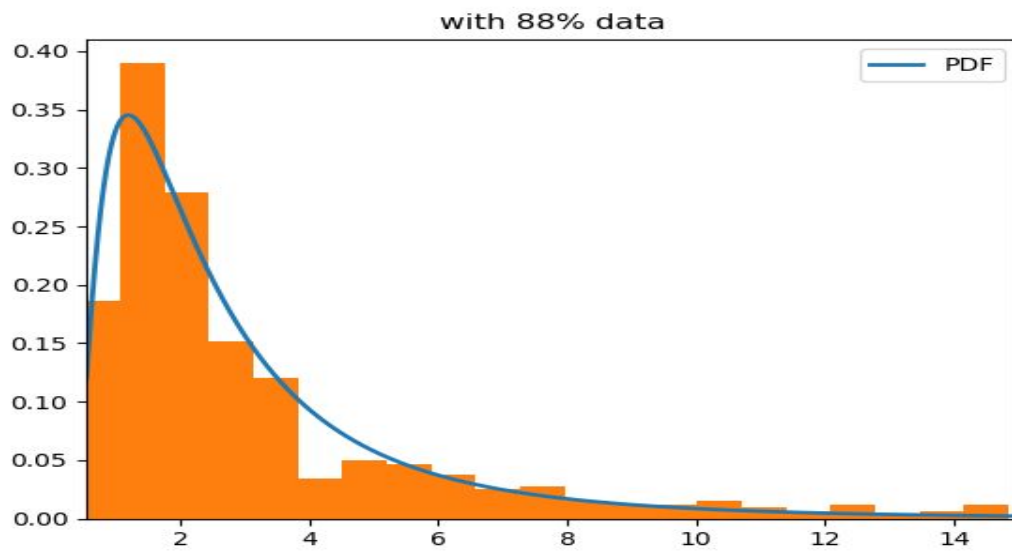
The expected data is following the observed data with 95% confidence level.

Conclusion :

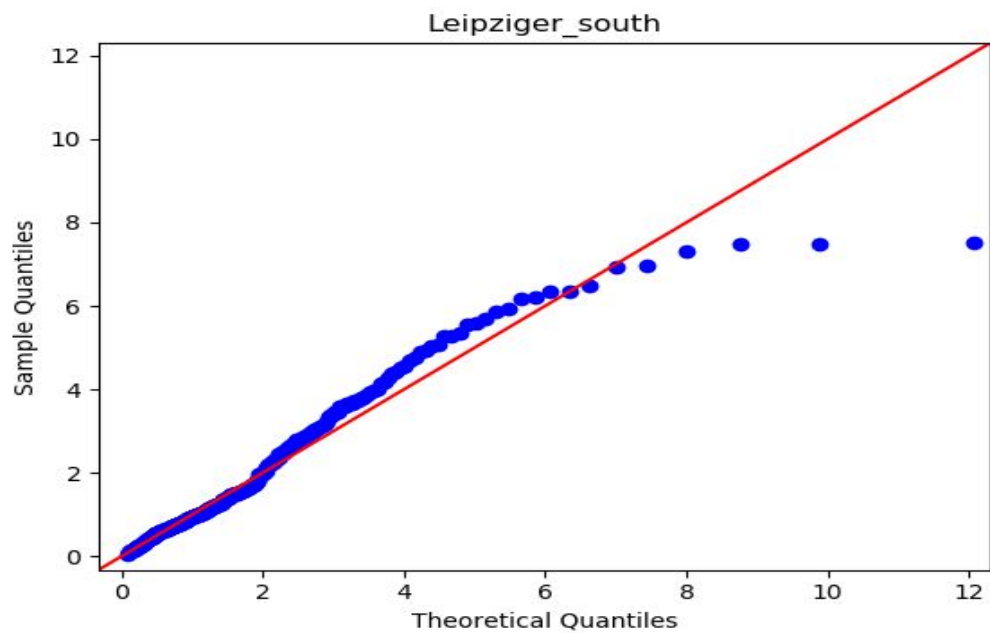
Lognormal with mean =1.267, standard deviation = 0.593

Leipziger Str. South

Histogram:



QQ plot:



Experimental observations and final conclusions

Observations:

- Distribution type: Lognormal
- Statistics:
 - Mean = 0.662
 - Standard deviation = 0.871
 - For $k = 34$, chi-square calculated less than critical

From the above comparisons we can conclude the test failed to reject the null hypothesis and the observed values and Expected values have correlation and the expected values are not occurring just by a mere coincidence.

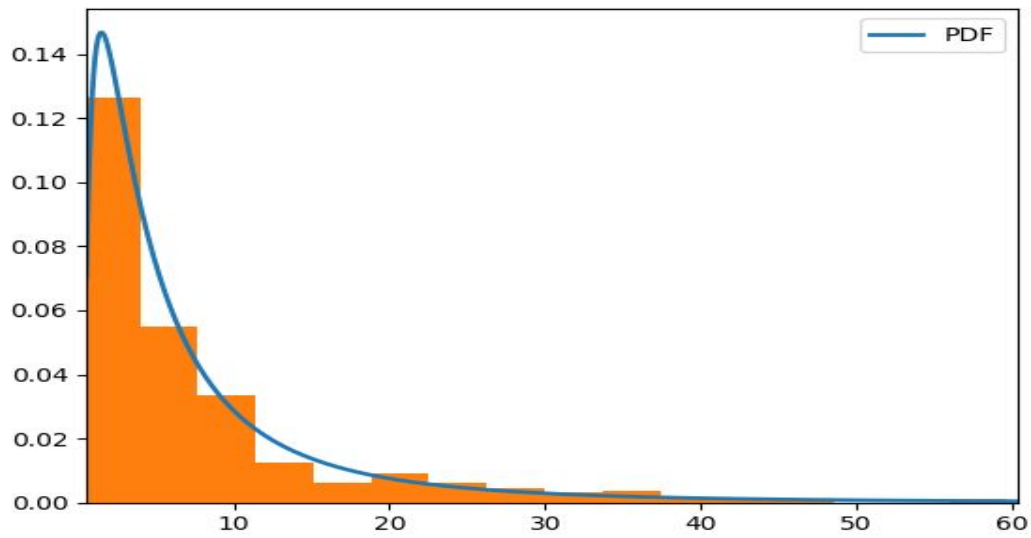
The expected data is following the observed data with 95% confidence level.

Conclusion:

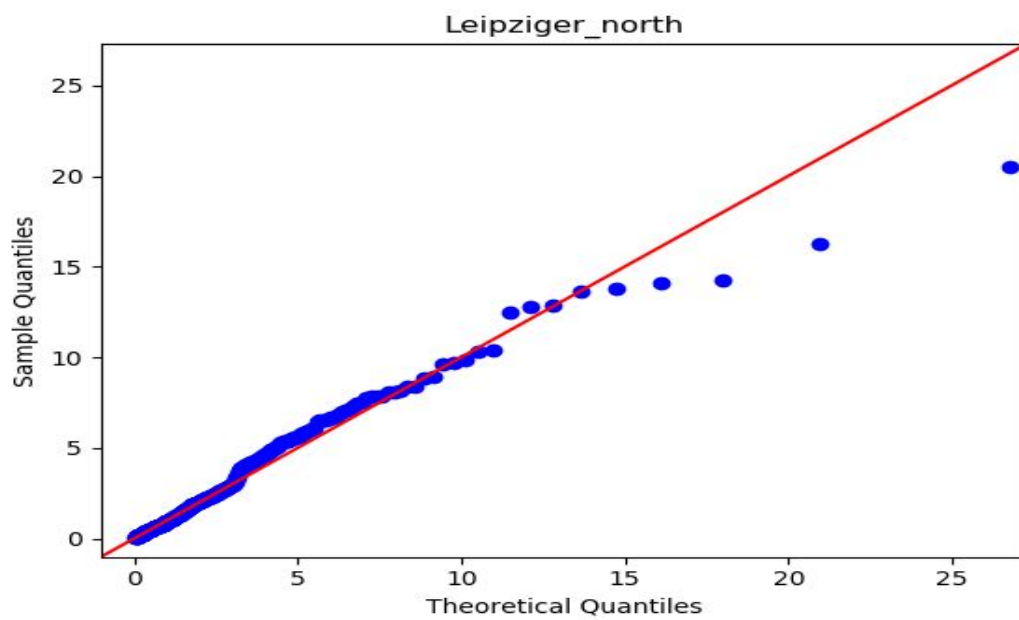
Lognormal with mean = 0.662, standard deviation = 0.871

Leipziger Str. North

Histogram:



QQ plot:



Experimental observations and final conclusions

Observations:

- Distribution type: Lognormal
- Statistics:
- Mean = 1.511
- Standard deviation = 1.112
- For $k = 25$, chi-square calculated less than critical

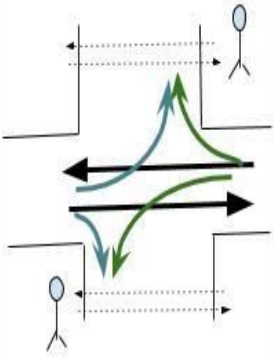








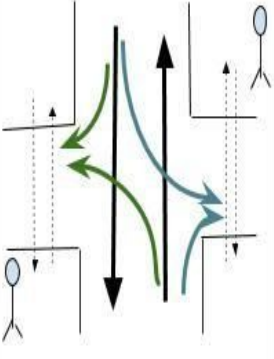








From the above comparisons we can conclude the test failed to reject the null hypothesis and the observed values and Expected values have correlation and the expected values are not occurring just by a mere coincidence.

The expected data is following the observed data with 95% confidence level.

Conclusion:

Lognormal with mean = 1.511, standard deviation = 1.112

Traffic lights

Timing	P1	P2	P3	P4	P5	P6	P7	P8
								
								

P1 = uniform_discr(29, 30)

P2,P4,P6,P8 = 4

P5 = uniform_discr(39, 41)

P3 = uniform_discr(51, 62)

P7 = uniform_discr(65, 70)

Bicycles, pedestrians and trams

For Erich-Weinert-Str to Am Fuchsberg:

- **Distribution type: Uniform-discrete [9 sec,34 sec]**
- Minimum value = 386/hour (9 sec)
- Maximum value = 106/hour (34 sec)

For Am Fuchsberg to Erich-Weinert-Str:

- **Distribution type: Uniform-discrete [10sec, 20sec]**
- Minimum value = 393/hour ($3600/393 = 10/\text{sec}$)
- Maximum value = 186/hour ($3600/186 = 20/\text{sec}$)

For Leipziger-str south to Leipziger-str North:

- **Distribution type: Uniform-discrete [10sec, 27sec]**
- Minimum value = 363/hour (10sec)
- Maximum value = 137/hour (27sec)

For Leipziger-str North to Leipziger-str south:

- **Distribution type: Uniform-discrete [12 sec, 28 sec]**
- Minimum value = 290/hour (12 sec)
- Maximum value = 129/hour (28 sec)

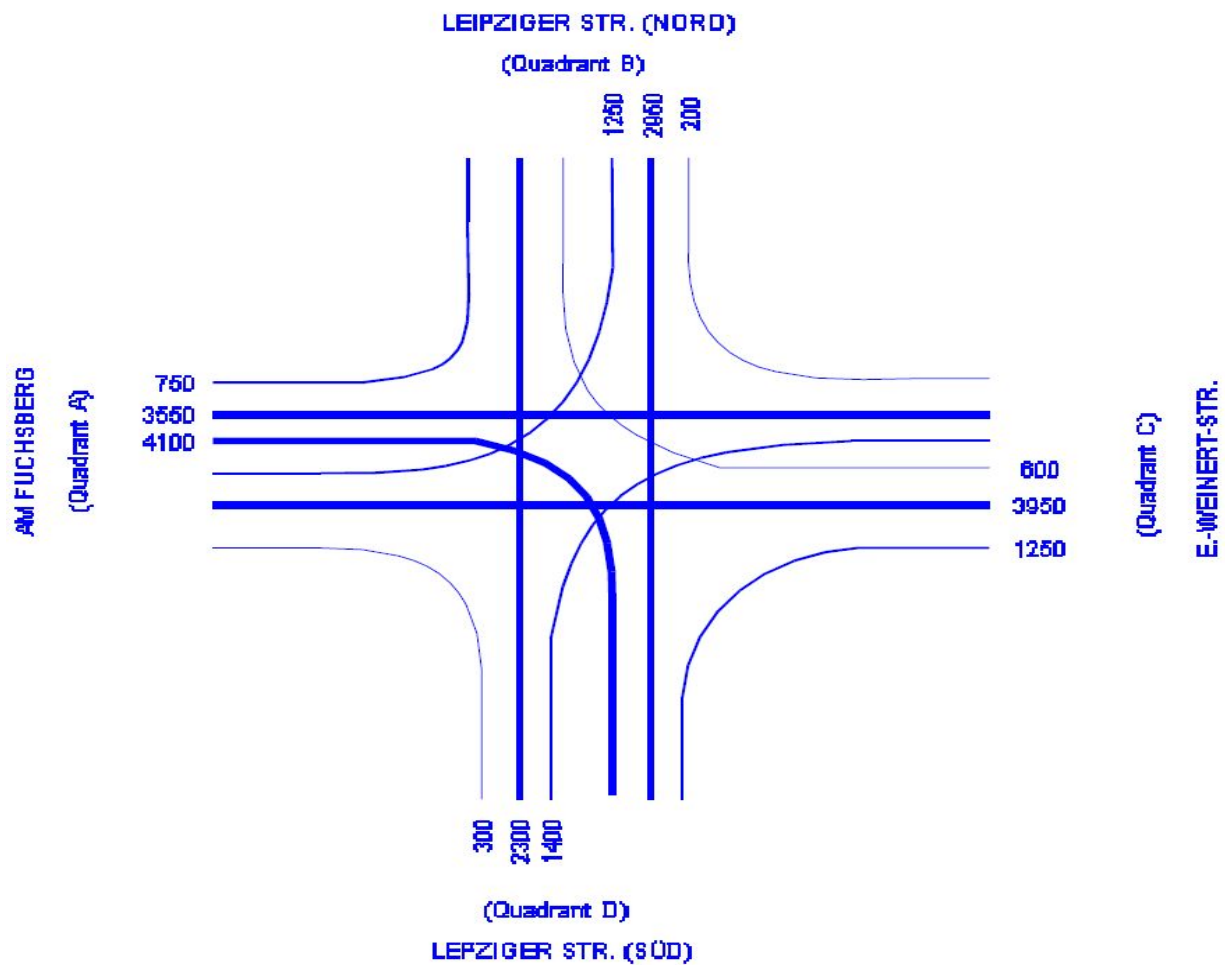
For Trams

- **Distribution: Truncated normal**
- Min value = 8 minutes, Max value = 12 minutes,
- Standard deviation = 2 minute, mean = 10minutes

Car Moving Direction with Count & Probability

Data Provided by Landeshauptstadt MAGDEBURG

Year : 2015



Am Fuchsberg	to	Leipziger strasse South	300
		Leipziger strasse North	1250
		Straight to Erich-Weinert-Str	3950
		Total	5510

$$P1 = 300/5500 = 0.05$$

$$P2 = 1250/5510 = 0.23$$

$$P3 = 3950/5510 = 0.72$$

Erich-Weinert-Str	to	Leipziger strasse North	200
		Leipziger strasse South	1400
		Straight to Am Fuchsberg	3550
		Total	5150

$$P4 = 200/5150 = 0.04$$

$$P5 = 1400/5150 = 0.27$$

$$P6 = 3550/5150 = 0.69$$

Leipziger Str. South	to	Am Fuchsberg	4100
		To Erich-Weinert-Str	1250
		Straight to Leipziger Str. North	2950
		Total	8300

$$P7 = 4100/8300 = 0.49$$

$$P8 = 1250/8300 = 0.15$$

$$P9 = 2950/8300 = 0.36$$

Leipziger Str. North	to	Am Fuchsberg	750
		To Erich-Weinert-Str	600
		Straight to Leipziger Str. South	2300
		Total	3650

$$P10 = 750/3650 = 0.21$$

$$P11 = 600/3650 = 0.16$$

$$P12 = 2300/3650 = 0.63$$

Additional measurements

Inputs:

1. Tram times
2. Pedestrian traffic lights

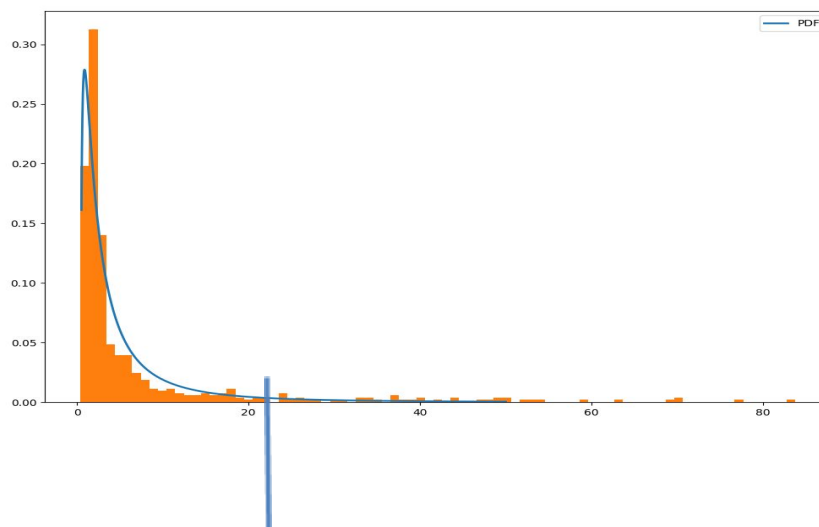
Outputs:

1. Sum of squared errors to check how well the expected value fits with the actual value curve.

Difficulties encountered while obtaining the data

We were unable to collect the data.

While analyzing the data we have encountered the problem of outliers due to which proper estimation of the distribution function is not done as the distribution function tries to fit the entire data which really does not have any impact on the system. The expected values are not in any way related to the actual data points thus leading to a false estimate of distribution function.



Dealing with trade-off between good estimation of distribution function and variance

Data censoring is done in order to reduce the variance and the mean is shifted to a point where the amount of data is high so that we can estimate a good fit for the data.

Following are the observations:

Leipziger str south:

Before removing:

- Variance: 145.705501
- Expon SSE: 22771.40234860699, rejected null hypothesis with 95% C.I
- Log norm SSE: 10547.130880170222, rejected null hypothesis with 95% C.I
- Norm SSE: 43318.018614922636, rejected null hypothesis with 95% C.I

Data < 15

exp: close fit in q-q plot but Rejected null hypothesis with 95% confidence interval as chi value is not less than tabulated (close to theoretical), sse = 96.99359997385118

Lognormal: Failed to reject null hypothesis with 95% confidence interval at $k = 34$ $\chi = 40.55014279825781 < 44.985$ ($df = 31$) (not only 34 but for most of them), sse = 294.0773265479426

Norm: Rejected null hypothesis with 95% confidence interval, chi value is not less than tabulated, sse = 959.3228009715183

Variance: 7.719009

Data < 25

exp: Rejected null hypothesis with 95% confidence interval but close fit in qq plot, sse = 908.1043538692305

Lognormal: failed to reject null hypothesis at $k = 80$, $\chi^2 = 96.8853491051358 < 98.484$ ($df = 77$) and for few other values after 81, close fit in qq plot, $sse = 1139.8242136414551$

Norm: No proper fit and Rejected null hypothesis with 95% confidence interval, $sse = 3620.8693617585723$

Variance: 19.972214

Data < 30

exp: close fit but Rejected null hypothesis with 95% confidence interval, χ^2 value is not less than tabulated for any k , $sse = 1506.8272257584504$

Lognormal: failed to reject null hypothesis with 95% confidence level at $k=83$, $\chi^2 = 97.01192955789713 < 101.88$ ($df = 80$), $sse = 1458.0231606620443$

Norm: Rejected null hypothesis with 95% confidence interval, χ^2 value is not less than tabulated for any k , $sse = 5141.926255269538$

Variance: 26.079071

Data < 40

exp: Rejected null hypothesis with 95% confidence interval, χ^2 value is not less than tabulated for any k , $sse = 4586.7407320703205$

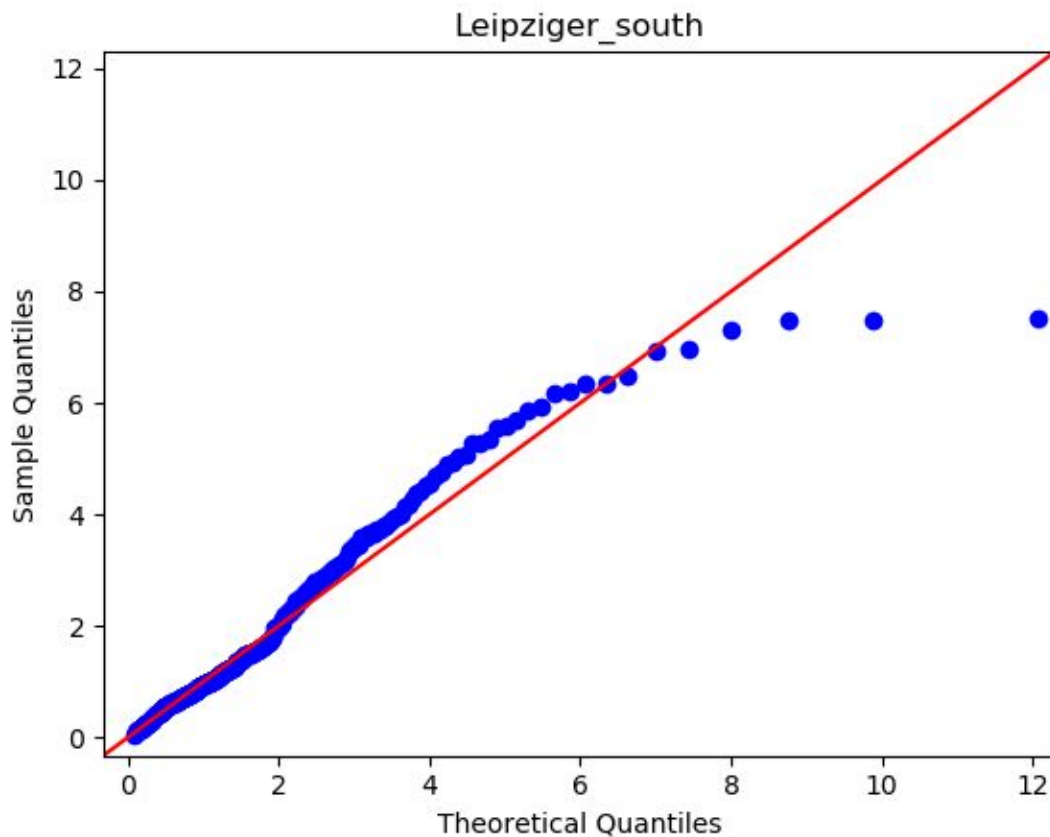
Lognormal: failed to reject null hypothesis at ($k=91$) = $\chi^2 = 108.835 < 110.90$ ($df = 88$), $sse = 3141.7580461553866$

Norm: Rejected null hypothesis with 95% confidence interval, χ^2 value is not less than tabulated for any k , $sse = 11457.69245570545$

Variance: 48.390222

CONCLUSION:

lognormal Data<15 would be a best choice as it failed to reject null hypothesis and q-q plot fits well with slope =1 for considerable range.



As the variance decreased to a greater extent the mean is shifted and does not consider the data which has very less impact on the system.

Limitations on the accuracy or validity of the data

1. The data is cleaned for the sole purpose of getting an optimum distribution function neglecting the data which does not really impact the system.
2. The data is biased towards smaller interarrival times.
3. Patterns in the data are likely to change over time and timely analysis has to be done.
4. Pedestrian data is not precise enough to do the analysis so we have to go with uniform distribution.

Cost overview

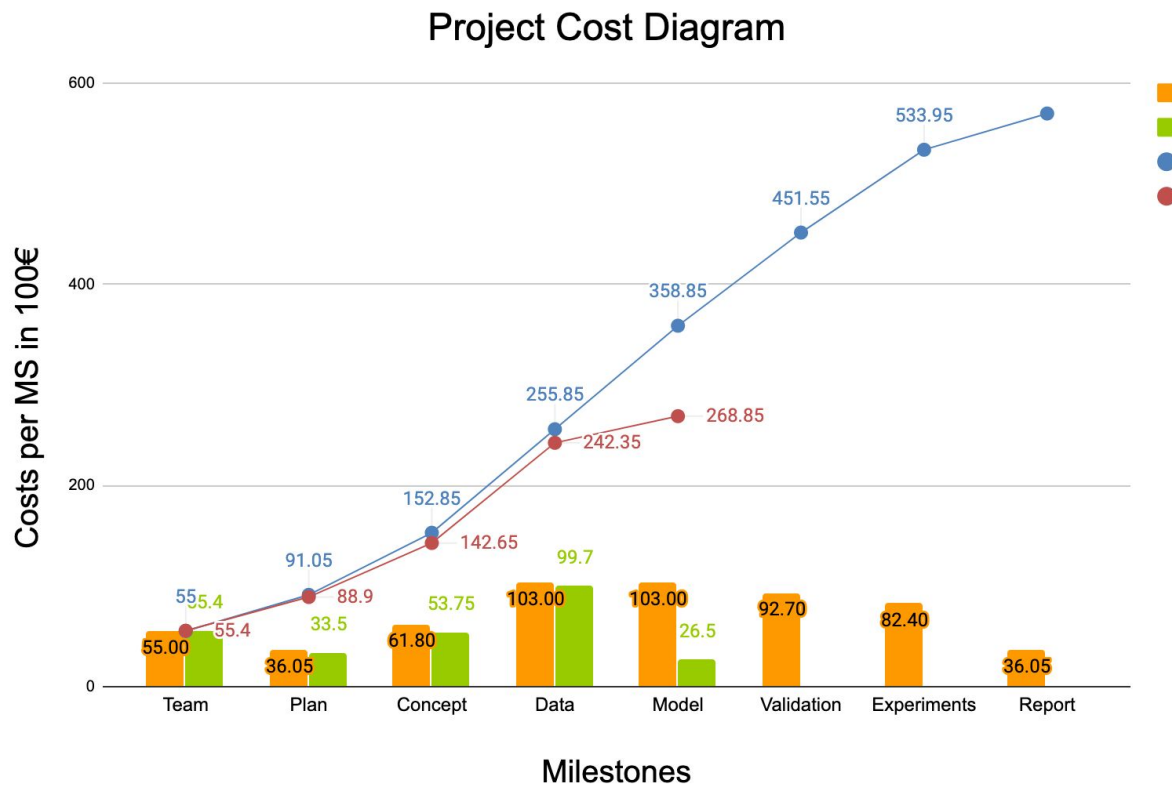
The chart below presents a two-dimensional breakdown of costs, aggregated by both milestone and individual members.

Name

Team Tetrahedron

	Lauro	Chandan	Vinay	Kavya	Arnab	Anjan	Total
Milestone 1 (hrs)	12.60	8.10	8.40	8.70	9.05	8.55	55.40
Milestone 2 (hrs)	5.75	3.50	4.00	5.05	4.25	10.95	33.50
Milestone 3 (hrs)	12.25	17.00	4.00	7.30	8.50	4.70	53.75
Milestone 4 (hrs)	19.50	9.30	41.00	8.30	12.50	9.10	99.70
Milestone 5 (hrs)	10.50			11.00	5.00		26.50
Milestone 6 (hrs)							0.00
Milestone 7 (hrs)							0.00
Milestone 8 (hrs)							0.00
Total hrs	60.60	37.90	57.40	40.35	39.30	33.30	268.85
Billing rate (hourly)							€100.00
							€26,885.00

Additionally, the chart below shows the cumulative cost of the project so far. The orange bars represent the planned milestone costs, and the blue line the planned cumulative cost. The green bars represent the actual milestone costs, and the red line the actual cumulative cost.



Future work

Complete the simulation model and validate the model.