

CLASSIFYING ROAD RISKS

Merging and analysing data from geographical data sources



Bachelor's thesis

Hämeenlinna University Centre,
Degree Programme in Business Information Technologies

March 2020

Sebastiaan Reggers

Business Information Technologies
Hämeenlinna University Centre

Author	Sebastiaan Reggers	Year 2020
Title	Classifying road risks: Merging and analysing data from geographical data sources	
Supervisor(s)	Deepak Kc	

ABSTRACT

The aim of this thesis was to research the possibilities of classifying road risks by combining datasets from different geographical data sources and merging these sets into one final dataset. The purpose of this dataset is to serve as data input for a data analytics solution containing one or more map visualisations that can be used to analyse road risks. The commissioning party is the HAMK Smart Research unit and a potential client in the logistics sector was mentioned.

In order to research the possibilities of this geographical data, a thorough literary research was performed. The results of this research were then used to create such a dataset and was tested in the form of a basic data analytics solution. The merged dataset was created using the RStudio toolchain and the analytics solution was created in Power BI Desktop.

This resulted in a compact, but clean and efficient dataset, accompanied by a functioning analytics solution. This solution remains just a demo but could already be used to draw some conclusions about road safety. This proves that classifying road risks by using geographical data from different sources is feasible. However, in order to make this process more optimal, other technologies than the ones used in this thesis should also be researched.

Keywords Open Data, Power BI, Data Analytics, Transport, Safety

Pages 40 pages including 8 pages of appendices

Abbreviations

WFS	Web Feature Service
WMS	Web Map Service
FTIA	Finnish Transport Infrastructure Agency
CRS	Coordinate Reference System
GIS	Geographic Information System
OGC	Open Geospatial Consortium
XML	eXtensible Markup Language
CSV	Comma-Separated Values
URL	Uniform Resource Locator
EPSG	European Petroleum Survey Group
FTP	File Transfer Protocol
QGIS	Quantum GIS

CONTENTS

1	INTRODUCTION	1
2	THEORETICAL FRAMEWORK	2
2.1	Geographical data services	2
2.1.1	WFS	2
2.1.2	WMS	3
2.2	Data sources	3
2.2.1	Road traffic accidents	4
2.2.2	FTIA open road network data	5
2.3	Analysis software	6
2.3.1	Power BI	7
2.3.2	QGIS	8
2.3.3	R	9
3	MERGING DATA SOURCES	11
3.1	Analysing the data source structure	11
3.1.1	Road traffic accidents	11
3.1.2	Finnish Transport Infrastructure Agency	12
3.2	Extracting the data	13
3.3	Cleaning up the datasets	14
3.4	Preparing the datasets for merging	16
3.5	Merging the datasets	17
4	DATA AND ADVANCED ANALYTICS SOLUTION	18
4.1	Solution overview	18
4.1.1	Case 1: Traffic accidents in Hämeenlinna and dangerous months	18
4.1.2	Case 2: Sight distance in Hämeenlinna with average comparison	19
4.1.3	Case 3: Height limit influence on accident risk at an intersection	21
4.1.4	Case 4: Influences on accidents along a section of the road	22
4.2	Business implications	23
5	CONCLUSION	24
5.1	The process	24
5.2	The result	25
5.2.1	What data sources are useful for road risk analysis?	25
5.2.2	Which techniques can be used for merging data from different sources to create one dataset?	25
6	SUMMARY	26
	REFERENCES	27

Appendices

Appendix 1 R code used for data manipulation

1 INTRODUCTION

The aim of this thesis is to research the possibilities of using open geospatial data in Finland to analyse and classify road risks and to lay the foundations of a data solution for this purpose.

Open data is the general term for data collections that can be freely accessed, used, shared and modified by anyone. It is a resource that has many uses for social and commercial activities. (Open Knowledge Foundation, n.d.)

With more than 2000 open data tables for around 26 different statistical topics, made available by statistics Finland, The Finnish Open Data Programme, set up by the Ministry of Finance in Finland and the data collected and made public by the Finnish Transport Infrastructure Agency, there is an immense amount of Open Data available in Finland. (Finnish Transport Infrastructure Agency, 2018; Statistics Finland, 2018)

Logistics is one of the sectors that could benefit a lot from the open data. An example of this is using geographical road data published by the government to aid in the process of transport Planning. Analysis of this data offers support at the “back-end” as a database management service and at the “front-end” to produce visualizations. (Miller & Shaw, 2001)

The topic was given by Juuso Saarinen and Olli Koskela who work for the Häme University of Applied Sciences. Their aim is to eventually involve an interested Finnish logistics company as a third party. In order to obtain the objectives of the thesis work, the following research questions were set:

- What data sources are useful for road risk analysis?

For a data solution to be useful, it needs useful data. In this thesis, the usefulness of data is determined on 3 levels:

1. Data source level (Chapter 2.2)
2. Dataset level (Chapter 3.3)
3. Data solution level (Chapter 4.1)

The search for data sources in this thesis is restricted to open data due to budget constraints.

- Which techniques can be used for merging data from different sources to create one dataset?

In order to create a solid data solution with data from different sources, the data needs to be uniform and preferably in one single dataset.

2 THEORETICAL FRAMEWORK

In order to formulate an answer to both research questions, an amount of background knowledge is required. First of all, the basic functionalities of different geographical data services have to be known. On top of that, the metadata and basic structure of the different data sources must be researched. Finally, in order to manipulate and visualise this data, several different software packages are required and thus need to be explored.

2.1 Geographical data services

Road risks are not static but dynamic. In other words, the data influencing road risks changes constantly. This means that the traditional method of employing and analysing historical data is inefficient. Instead, live data deployment services and analysing techniques should be used. WFS and WMS are such services and are most commonly used for deploying geospatial data. (Strong, 2017)

2.1.1 WFS

Web Feature Service or WFS is a standard set by the OGC for sharing access to geographic information over the web. The difference between this service and traditional services such as FTP is that WFS offers direct access to the information at feature and feature property levels. This means that a client can use the service to only retrieve or to modify specific data in the form of queries. (Open Geospatial Consortium Inc., 2010)

The queries are generated by executing an operation in the form of a URL (e.g. www.wfsresource.com/wfs?service=WFS). There are several types of operations in the standard, of which two types are commonly employed to retrieve data.

1. Discovery operations: Operations used to analyse the structure of the data source.
2. Query operations: Operations used to retrieve and/or modify data from the data source.

Two of these operations are employed in this thesis. (Open Geospatial Consortium Inc., 2010)

GetCapabilities is a discovery operation and is used to acquire more information about a WFS source. This information shows, among other things, the different layers and CRS formats available. (Open Geospatial Consortium Inc., 2010)

GetFeature is a query operation and is used to retrieve a dataset from the source. This operation makes use of several parameters that can be

specified by the client for a more specific dataset in a preferred format. A few of these parameters are:

- Typename: Specifies the name of the layer that is to be retrieved.
- Srsname: Specifies the CRS format of the coordinates. The original coordinates will be converted to use this format.

The amount of feature specification that can be achieved with the use of just these three parameters demonstrates the power of WFS. (Open Geospatial Consortium Inc., 2010)

2.1.2 WMS

Web Map Service or WMS is a standard set by the OGC for producing maps of spatially referenced data, based on geographic information. The standard defines three operations: GetCapabilities, GetMap and GetFeatureInfo. (Open Geospatial Consortium Inc., 2006)

GetCapabilities for WMS has a similar functionality as GetCapabilities for WFS. The operation returns an overview of service level metadata, which can be used to determine the different layers and CRS formats available. (Open Geospatial Consortium Inc., 2006)

GetMap returns a map, based on the request parameters specified by the client. A few of these parameters are:

- Layers: Specifies the name of the layers that are to be rendered
- CRS: Specifies the CRS format of the map
- BBOX: Specifies the borders of the map in CRS specific units
- Width: Specifies the width of the picture in pixels
- Height: Specifies the height of the picture in pixels

The result of a GetMap service request is a map in picture format. (Open Geospatial Consortium Inc., 2006)

GetFeatureInfo is an optional operation and is not always supported by every layer of a WMS service. The purpose of the operation is to retrieve more information about specific features in the picture of the map. (Open Geospatial Consortium Inc., 2006)

2.2 Data sources

In this thesis only open data sources are used. Open data are data that don't have access or usage restrictions (Open Knowledge Foundation, n.d.). This means that all the data sources used in this thesis are openly accessible, without any extra costs. Another criterium, to ensure the reliability and veracity of the data, is that only governmental owned or steered data sources are considered useful.

When it comes to the datasets in the data sources, the only specific criterium is that the coordinates are of GIS spatial point vectors data type.

The CRS of this data type can easily be converted to another CRS. The simplicity of this data also makes visualization on a map easier. (GISGeography, 2018)

2.2.1 Road traffic accidents

This data contains info about traffic accidents on the road that are reported to the Finnish police. The biggest benefit of this data is that it comes with coordinates. This facilitates the visualisation of the data on a map. (Statistics Finland, n.d.)

The producer and owner of this data is Statistics Finland, an organization founded in 1865. This organization is, in their own words: “the only Finnish public authority specifically established for statistics”. Their main goal is to combine collected data and produce statistics based on this data. (Statistics Finland, 2019)

Each data entry in this dataset contains a set of properties. Illustrated in table 1 is an explanation of each property using a single result (figure 1) from the dataset as an example:

```

vvonn kkonn kello    vakav onntyyppi lkmhapa lkmlaka lkmjk lkmp lkmmpo lkmmp lkmmuukulk    x    y
2018    1 17.00-      2      8      1      0      0      0      0      0      0 398360.9 6680606
        17.59

```

Figure 1. One row of the Road traffic accidents dataset.

Table 1. Road traffic accident dataset properties explained with example.

Property	Explanation	Example value
vvon	Year of the accident	2018
kkonn	Month of the accident	1 (January)
kello	Time of the accident	17:00 – 17:59
vakav	Seriousness of the accident: 1. accident resulting in death 2. accident resulting in injury 3. accident resulting in serious injury	2 (Injury)
onntyyppi	Type of accident 0. same direction of travel (going straight) 1. same direction of travel (turning) 2. opposite direction of travel (going straight) 3. opposite direction of travel (turning) 4. intersecting direction of travel (going straight) 5. intersecting direction of travel (turning)	8 (Running of the road)

	6. pedestrian accident (on pedestrian crossing)	
	7. pedestrian accident (elsewhere)	
	8. running off the road	
	9. other accident	
lkmhapa	Number of passenger cars and vans in the accident	1
lkmlaka	Number of buses and lorries in the accident	0
lkmjk	Number of pedestrians in the accident	0
lkmpp	Number of cyclists in the accident	0
lkmmo	Number of mopeds in the accident	0
lkmmp	Number of motorcycles in the accident	0
lkmmuukulk	Number of other vehicles in the accident	0
x	x coordinate of the accident ()	398360.9
y	y coordinate of the accident	6680606

The coordinates use the EPSG:3067 CRS but can be converted through the WFS and WMS services (See Chapter 2.3.1.). The WFS URL is: <http://geo.stat.fi/geoserver/tieliikenne/wfs?> (Paikkatietohakemisto, n.d.)

This data is openly available in the form of both WMS and WFS services.

2.2.2 FTIA open road network data

The FTIA has collected quite a big amount of data. A big part of this data has been made openly accessible, accomplishing several of the goals of the Finnish Open Data Programme 2013–2015. (Finnish Transport Infrastructure Agency, 2018)

The road network data from this data source is divided in three parts: Digttraffic for real-time traffic data, Digiroad and the Road Register, both for spatial data (Finnish Transport Infrastructure Agency, 2015). This data is openly viewable in the form of map layers on their online application: <https://julkinen.vayla.fi/oskari/?lang=en>.

The datasets in the Transport Network section are also available in the form of both WMS and WFS services. Most of these data are in the vector line and polygon formats which are not data formats supported by Power BI. There are also a few datasets in Shape Point Vector format which can be broken down and converted to decimal coordinates and are supported. (GISGeography, 2018; Iseminger, 2020)

This data source contains many different datasets and only a small selection is used in this thesis (table 2). This selection is based on 2 criteria:

1. The coordinates are in Shape Point Vector format.
2. The dataset is related to road safety.

Table 2. The dataset selection from the FTIA data source, with the reason for selection.

Dataset	English translation	Why useful?
Korkeusrajoitus	Height limit	If the height limit is too low, it could be dangerous to use the road.
Leveysrajoitus	Width limit	If the width limit is too small, it could be dangerous to use the road.
Näkemäpituus	Sight distance	If the sight distance is too short, it could be dangerous to use the road.
Onnettomuusriski	Accident risk, intersection	If the risk of an accident is too high, the intersection could be dangerous.
Palvelualueet	Service areas	If there are many service areas near a road, the road might be safer.
Suojatiet	Pedestrian crossings	If a road has many pedestrian crossings, the risk of hitting a pedestrian could be high.

The coordinates use the EPSG:3067 CRS but can be converted through the WFS and WMS services (See Chapter 2.3.1.). The WFS URL is: <https://julkinen.vayla.fi/inspirepalvelu/avoin/wfs?> (Finnish Transport Infrastructure Agency, 2019).

2.3 Analysis software

Both geospatial and live data need specialised software in order to be employed, manipulated and analysed. To start off, the geospatial data deployed through WFS and WMS data services have the GIS format and thus need software that is specialised in GIS data handling (Open Geospatial Consortium Inc., 2010; Open Geospatial Consortium Inc., 2006). Furthermore, due to the ever-changing nature of live data, software with live-data analysis capabilities is required. (Strong, 2017)

There are several analysis software combinations that meet these requirements, but the one presented in this thesis is QGIS to R to Power BI. At the start, QGIS is used for data source structure analysis. Next, the knowledge gathered from this analysis is then used in order to acquire and manipulate the data in R. At the end, this manipulated data is used in Power BI for visualisation and analytics.

2.3.1 Power BI

Power BI is primarily a solution for business analytics, with the power to easily connect, process and visualize data from multiple sources. On top of that, the Power BI Desktop application is available for free and is thus optimal for research purposes. (Microsoft, 2020)

One of the tools available in this application is the Power Query Editor. This tool has many built-in functionalities to better prepare the data for visualization. On launching the tool, it reads a limited number of rows from the dataset after which the user can start the process of preparing the data. This happens on a step-by-step basis, where the user defines which operations are executed on the data and in which specific order. After the preparation is done and the user chooses to save and apply, the tool will apply these changes to the data model into the application for further use. The biggest advantage of this system is that it allows the user to go back on their changes and further fine-tune the preparation whenever they wish to do so. (Popell & Iseminger, 2019)

There are 2 map visualisations in Power BI for visualizing a collection of geographical coordinates. The first one, ArcGIS Maps, is part of the standard set of Power BI visualizations. The other one, Mapbox, needs to be imported from the Power BI marketplace.

The ArcGIS Maps visualization accepts latitude and longitude coordinates in WGS84 CRS as input and will then display this data in an overlay map theme defined by the user. This theme can be Location Only, Heat Map or Cluster (figure 2). An interesting functionality of the ArcGIS Maps in specific is its possibility to add reference layers to the map. The user can browse through the many layers publicly made available or create their own ones with an ArcGIS online organizational subscription. (Esri, 2020)

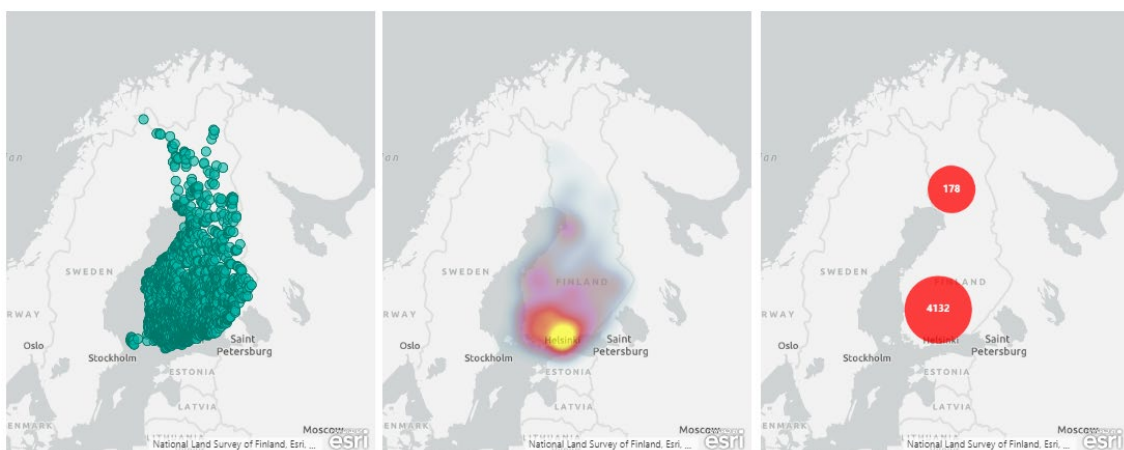


Figure 2. A comparison of the 3 different map themes in ArcGIS Maps.

The Mapbox visualization is in a lot of aspects rather similar to the ArcGIS Maps visualization. It has the same basic functionalities (figure 3), but it lacks the added reference layer functionality. Instead of this however, it adds more customizability, a whole array of extra functions and a basic raster layer functionality that could process a basic WMS request. (Mapbox, n.d.)

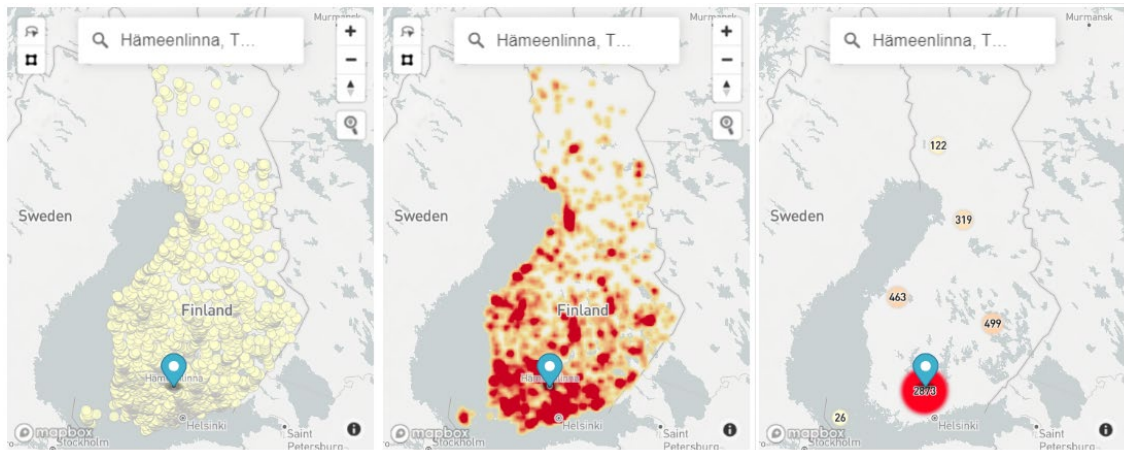


Figure 3. A comparison of 3 layer-types in Mapbox.

2.3.2 QGIS

QGIS is an open source geographic information system, which aims to function as a client as well as a server for geographic information. QGIS Desktop is a client QGIS software for PC that has an extensive toolkit to access, create and share geographical information. (QGIS Development Team, 2020)

One of the functionalities of this application is the Layers toolkit which can, among other things, read, analyse and process WFS and WMS service layers. The application supports the use of multiple layers and handles layer service processing in a user-friendly way. (QGIS Development Team, 2020)

The Add Layer functionality allows the user to connect to a WFS or a WMS service and will immediately show an overview of the different layers contained in said service. The user can then choose between the different layers, configure some of the request parameters like the CRS and add the layer to the map. (QGIS Development Team, 2020)

Furthermore, with the use of the QGIS Network Logger plugin (figure 4) it becomes quite simple to retrieve the service request URL used by QGIS to request a specific WFS or WMS layer in a specific CRS format. (Duivenvoorde, 2019)

All in all, this application allows the user to research the capabilities of a WMS or WFS service in a more user-friendly way than simply using the GetCapabilities operation.

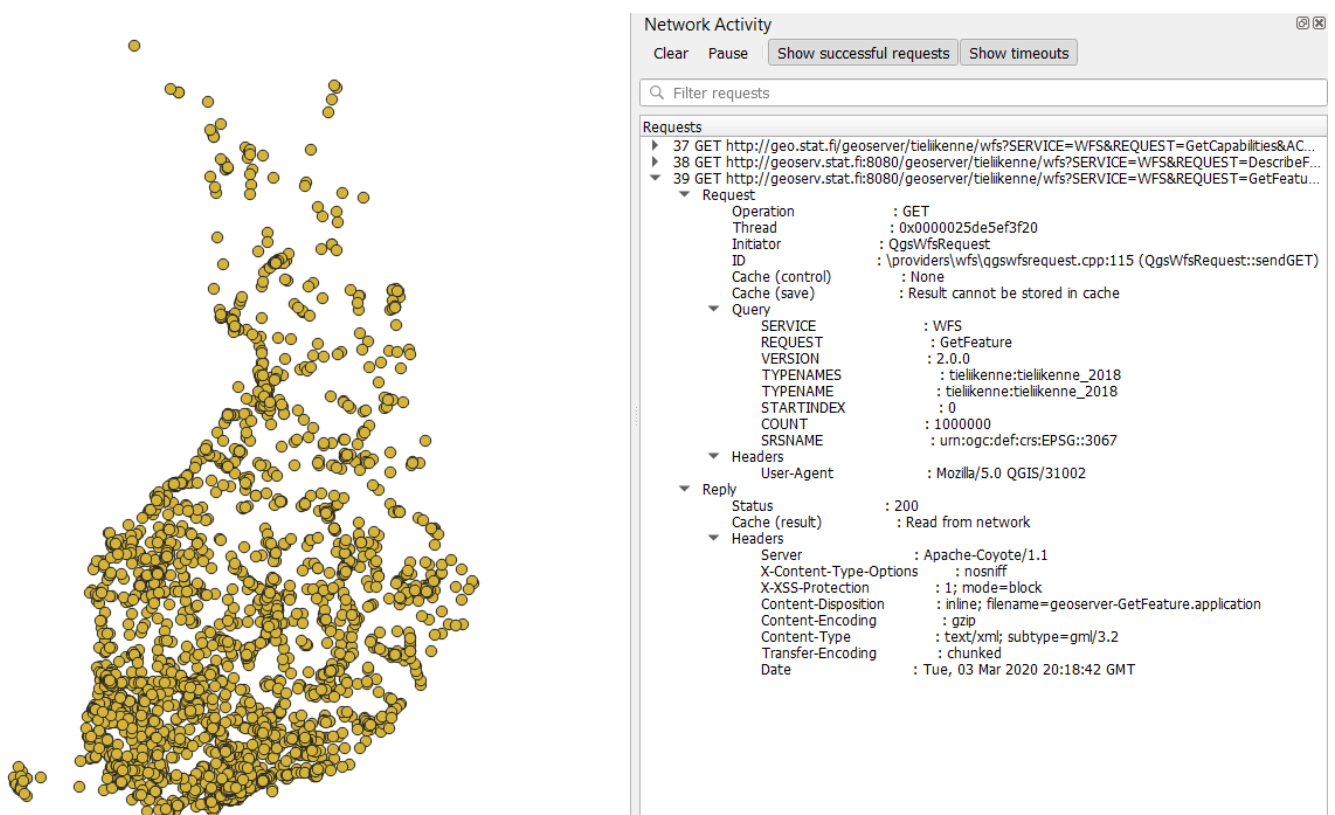


Figure 4. A WFS map layer in QGIS, logged by the network logger.

2.3.3 R

R is a programming language and software environment for statistical computing and graphics. The main goal of the language is to be able to process large amounts of data. The functionalities are quite extensive and include, next to what one would expect from a simple programming language, also data handling and storage facility, operators for calculations on arrays, a collection of tools for data analysis and several graphical facilities. (The R Foundation, n.d.)

One of these tools is Tidyverse, a collection of R packages that work together seamlessly and are designed for data science. This collection greatly decreases the complexity of the code, while increasing its functionality. It also allows for the use of “piping”, a programming technique for increasing the readability of complex code. (Grolemund & Wickham, 2016)

RStudio is an integrated development environment for R and is available as an open source desktop application. RStudio allows the developer to

write, edit and execute R code directly from its source editor. It also contains a big collection of tools for software integration. (RStudio, n.d.)

As illustrated in figure 5, one of the functionalities of RStudio is creating and editing R Notebooks. A notebook allows the user to execute parts of R code, so-called “lines”, independently and interactively, with intermediate outputs shown after every line. This eases the process of writing a bigger chunk of code and facilitates error resolving. It is also useful for researching a dataset, while changing the analysis code bit by bit, comparing the results. (Xie, Allaire, & Golemund, 2019)

```

---
title: "Suomi Traffic Accidents"
output: html_notebook
---

```{r}
library(httr)
wfs_accidents <- "http://geo.stat.fi/geoserver/tieliikenne/wfs?"
```

```{r}
library(knitr)
query <- list(service = "WFS",
 request = "GetFeature",
 version = "1.3.0",
 typeName = "tieliikenne:tieliikenne_2018",
 outputFormat = "csv",
 srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_accidents, query = query)
df <- read.csv(textConnection(content(result, 'text')))
kable(head(df, 5))
```

```

| FID | geom | vvonn | kkonn | kello | vakav | onntyyppi | lkmhapa | lkmlaka | lkmjk | lkmp | lkmn |
|--------------------|--|-------|-------|-------------|-------|-----------|---------|---------|-------|------|------|
| tieliikenne_2018.1 | POINT (60.24944757965415 25.163782757442434) | 2018 | 1 | 17.00-17.59 | 2 | 8 | 1 | 0 | 0 | 0 | |
| tieliikenne_2018.2 | POINT (60.23817205255151 24.858923628964465) | 2018 | 1 | 20.00-20.59 | 2 | 8 | 2 | 0 | 0 | 0 | |
| tieliikenne_2018.3 | POINT (60.21046093801623 25.05870117966777) | 2018 | 1 | 17.00-17.59 | 2 | 4 | 1 | 0 | 0 | 1 | |
| tieliikenne_2018.4 | POINT (60.169831696308414 24.92609015767353) | 2018 | 1 | 03.00-03.59 | 2 | 9 | 1 | 0 | 1 | 0 | |
| tieliikenne_2018.5 | POINT (60.24146305859899 24.849496287820408) | 2018 | 1 | 12.00-12.59 | 2 | 0 | 4 | 0 | 0 | 0 | |

Figure 5. A WFS service request and table display in R

3 MERGING DATA SOURCES

Merging different datasets from different sources can't and shouldn't be done without careful preparation (Osborne, 2013). This preparation is represented on a step-by-step basis. First, the structure from each of the data sources is analysed, after which the datasets can be extracted. Subsequently, these datasets are trimmed until only the essential data remains. To finish the preparation, all datasets are transmuted to have a uniform structure.

3.1 Analysing the data source structure

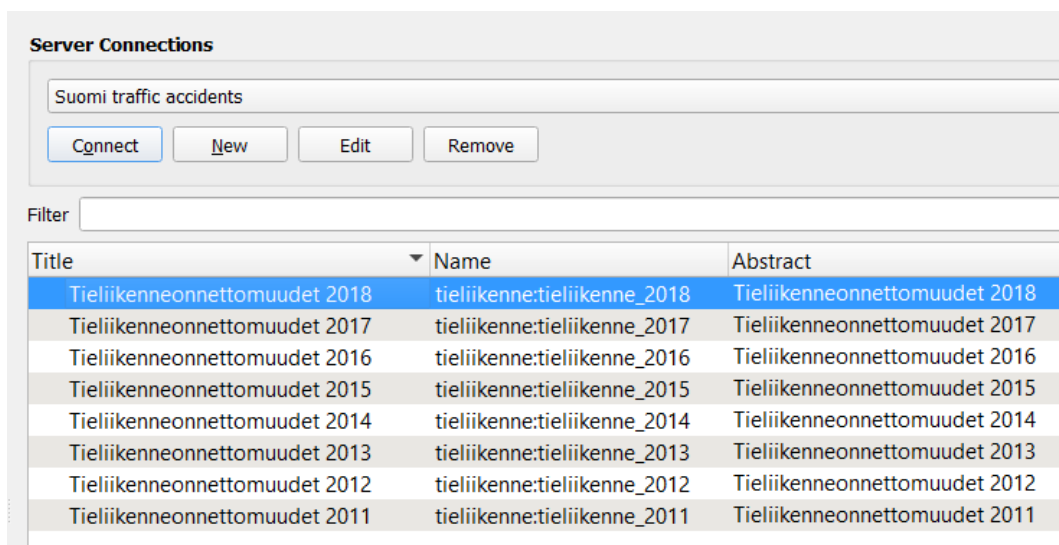
WMS and WFS data need some preparation in order to be successfully interpreted by data analysis and visualisation software. The data is encapsulated in an XML structure and needs to be extracted. To accomplish this the structure of each source needs to be known. (Open Geospatial Consortium Inc., 2010; Open Geospatial Consortium Inc., 2006)

There are several tools that can be used to analyse this structure. Two of these – QGIS (Chapter 2.5.2) and R (Chapter 2.5.3) – are used in this sub-chapter. The analysis results are compared to each other for greater accuracy. Both data sources from Chapter 2.4 are analysed here this way.

3.1.1 Road traffic accidents

The structure of this data source is not that complicated (figures 6 and 7). The data is divided in layers on annual basis, from 2011 up until 2018. The name format is:

```
tieliikenne:tieliikenne_201X
```



The screenshot shows the 'Server Connections' dialog in QGIS. At the top, there is a text box containing 'Suomi traffic accidents' and four buttons: 'Connect', 'New', 'Edit', and 'Remove'. Below this is a 'Filter' input field. The main part of the dialog is a table with three columns: 'Title', 'Name', and 'Abstract'. The table lists ten layers, one for each year from 2011 to 2018. The 2018 layer is highlighted in blue.

| Title | Name | Abstract |
|-------------------------------|------------------------------|-------------------------------|
| Tieliikenneonnettomuudet 2018 | tieliikenne:tieliikenne_2018 | Tieliikenneonnettomuudet 2018 |
| Tieliikenneonnettomuudet 2017 | tieliikenne:tieliikenne_2017 | Tieliikenneonnettomuudet 2017 |
| Tieliikenneonnettomuudet 2016 | tieliikenne:tieliikenne_2016 | Tieliikenneonnettomuudet 2016 |
| Tieliikenneonnettomuudet 2015 | tieliikenne:tieliikenne_2015 | Tieliikenneonnettomuudet 2015 |
| Tieliikenneonnettomuudet 2014 | tieliikenne:tieliikenne_2014 | Tieliikenneonnettomuudet 2014 |
| Tieliikenneonnettomuudet 2013 | tieliikenne:tieliikenne_2013 | Tieliikenneonnettomuudet 2013 |
| Tieliikenneonnettomuudet 2012 | tieliikenne:tieliikenne_2012 | Tieliikenneonnettomuudet 2012 |
| Tieliikenneonnettomuudet 2011 | tieliikenne:tieliikenne_2011 | Tieliikenneonnettomuudet 2011 |

Figure 6. Road traffic accidents data source structure in QGIS

```

INFO: Open of `WFS:http://geo.stat.fi/geoserver/tieliikenne/wfs?`
      using driver `WFS` successful.
Metadata:
  ABSTRACT=Tilastokeskuksen palvelurajapinta (WFS)
  PROVIDER_NAME=Tilastokeskus
  TITLE=Tilastokeskuksen palvelurajapinta (WFS)
1: tieliikenne:tieliikenne_2011 (title: Tieliikenneonnettomuudet 2011) (Point)
2: tieliikenne:tieliikenne_2012 (title: Tieliikenneonnettomuudet 2012) (Point)
3: tieliikenne:tieliikenne_2013 (title: Tieliikenneonnettomuudet 2013) (Point)
4: tieliikenne:tieliikenne_2014 (title: Tieliikenneonnettomuudet 2014) (Point)
5: tieliikenne:tieliikenne_2015 (title: Tieliikenneonnettomuudet 2015) (Point)
6: tieliikenne:tieliikenne_2016 (title: Tieliikenneonnettomuudet 2016) (Point)
7: tieliikenne:tieliikenne_2017 (title: Tieliikenneonnettomuudet 2017) (Point)
8: tieliikenne:tieliikenne_2018 (title: Tieliikenneonnettomuudet 2018) (Point)

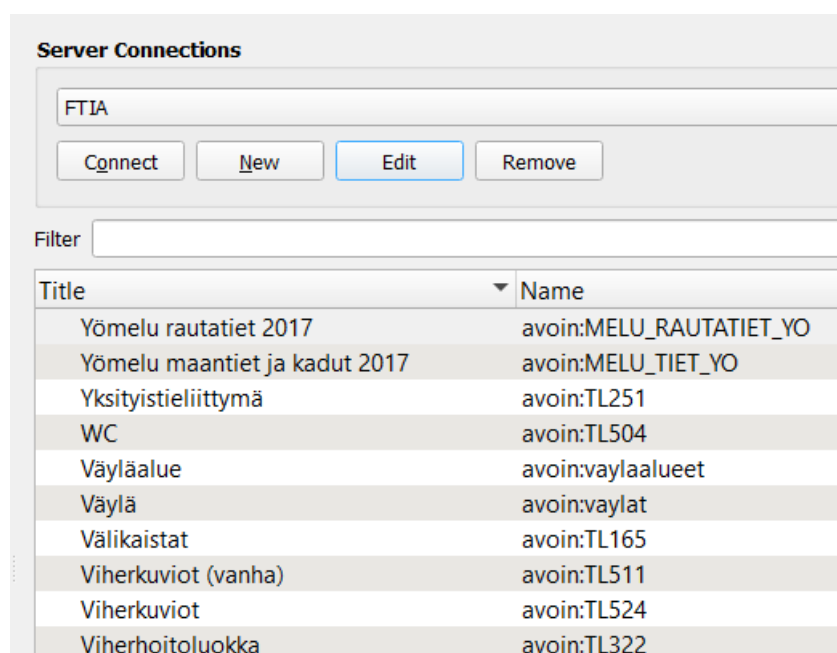
```

Figure 7. Road traffic accidents data source structure in R

3.1.2 Finnish Transport Infrastructure Agency

This data source structure (figure 8) is more complicated. There are many different layers with many different vector coordinate types. For the purpose of this thesis only Point vectors are used due to their simplicity.

This is where R shows its usefulness. In R it is possible to filter the connection result with substring functionalities. This means that the user can filter on a word, e.g. "Point", and the program result will only show rows containing that word. As illustrated in figure 9, this results in a quick overview of the specific layers that are useful for the user, saving a lot of time otherwise spent on manual research.



| Title | Name |
|-------------------------------|-------------------------|
| Yömelu rautatiet 2017 | avoin:MELU_RAUTATIET_YO |
| Yömelu maantiet ja kadut 2017 | avoin:MELU_TIET_YO |
| Yksityistieliittyvä | avoin:TL251 |
| WC | avoin:TL504 |
| Väyläalue | avoin:vaylaalueet |
| Väylä | avoin:vaylat |
| Välikaistat | avoin:TL165 |
| Viherkuviot (vanha) | avoin:TL511 |
| Viherkuviot | avoin:TL524 |
| Viherhoitoluokka | avoin:TL322 |

Figure 8. FTIA data source structure in QGIS, last 10 layers


```

4: avoin:TL262 (title: Alikulkupaikka) (Point)
9: avoin:rata_baliisi (title: Baliisi) (Point)
10: avoin:TL508 (title: Bussipysäkin katokset) (Point)
11: avoin:TL507 (title: Bussipysäkin varusteet) (Point)
13: avoin:rata_erotusjakso (title: Erotusjakso) (Point)
16: avoin:TL516 (title: Hiekkalaatikot) (Point)
19: avoin:rata_ilmainen (title: Ilmainen) (Point)
22: avoin:DR_PYSAKKI (title: Joukkoliikenteen pysäkit (Digiroad)) (Point)
23: avoin:TL505 (title: Jätehuolto) (Point)
29: avoin:TL211 (title: Kantavuusmittaus) (Point)

```

Figure 9. FTIA data source structure in R, first 10 layers containing the word “Point”

3.2 Extracting the data

Once the data source structure is analysed, extracting the data is not that complicated. Only one custom request parameter is required, and two others can be very useful. These request parameters are the layer to be requested, the CRS and the output format.

In order to keep this chapter short one example dataset is used. The method of constructing this dataset connection can be applied to all the dataset connections. The dataset used in this example is Road Accidents in 2018.

The data source structure analysis shows that the name of this dataset is tieliikenne:tieliikenne_2018. The preferred CRS is EPSG:4326 for easy to use global coordinates in latitude and longitude (Klokantech Technologies GmbH, 2007). An easy to use output format is CSV. As illustrated in figure 10, the Network Logger plugin in QGIS can be used to show the required format of some of these parameters.

▼ Query

| | |
|------------|--------------------------------|
| SERVICE | : WFS |
| REQUEST | : GetFeature |
| VERSION | : 2.0.0 |
| TYPENAMES | : tieliikenne:tieliikenne_2018 |
| TYPENAME | : tieliikenne:tieliikenne_2018 |
| STARTINDEX | : 0 |
| COUNT | : 1000000 |
| SRSNAME | : urn:ogc:def:crs:EPSG::3067 |

Figure 10. QGIS Network Logger WFS request query parameters. Circled in red are the custom parameters and in blue are the common parameters.

Using the information from this short analysis, the custom request parameters are:

```

typeNameNames = tieliikenne:tieliikenne_2018
srsName        = urn:ogc:def:crs:EPSG::4326
outputFormat   = csv

```

In order to employ the data in a browser or a Power BI solution, a connection URL is required. Constructing this URL can be done using the following steps:

1. Take the base URL:

`http://geo.stat.fi/geoserver/tieliikenne/wfs?`

2. Add the common request parameters:

`http://geo.stat.fi/geoserver/tieliikenne/wfs?service=WFS&request=GetFeature&version=2.0.0`

3. Add the custom request parameters:

`http://geo.stat.fi/geoserver/tieliikenne/wfs?service=WFS&request=GetFeature&version=2.0.0&typeName=tieliikenne:tieliikenne_2018&srsName=urn:ogc:def:crs:EPSG::4326&outputFormat=csv`

To employ the data in R a query can be constructed using a list of request parameters as input. This query can then be executed using a GET method (figure 11). This creates a more readable overview of these parameters and allows for easier customization in case the request needs to be changed.

```
query <- list(service = "WFS",
             request = "GetFeature",
             version = "2.0.0",
             typeName = "tieliikenne:tieliikenne_2018",
             outputFormat = "csv",
             srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_accidents, query = query)
```

Figure 11. Using a request parameter list as query input for a GET method in R

3.3 Cleaning up the datasets

Merging different datasets becomes easier when the datasets only have a few columns. To accomplish this only the essential columns of each dataset are kept. This includes the dates, the locations and the most important dataset specific data. (Osborne, 2013)

First up is the location data. Because of the CRS specification each dataset has a column containing the coordinates in the form of a Point vector. In the Road traffic accidents datasets this data is found in the “geom” column. In the FTIA datasets the column is called “SHAPE”. Next to the coordinates, all FTIA datasets also contain a “KUNTA” column with the name of the municipality. (Finnish Transport Infrastructure Agency, 2019; Paikkatietohakemisto, n.d.)

Next up, when it comes to dates, the Road traffic accidents datasets contain a column called “vvon” with the year and a column called “kkonn” with the month of the accident. The datasets from FTIA have a column called “MUUTOSPVM” containing the full date of the latest update to the measurement. (Finnish Transport Infrastructure Agency, 2019; Paikkatietohakemisto, n.d.)

Finally, for the dataset specific information, table 3 displays the columns that were kept, accompanied with the first results in figure 12.

Table 3. Overview of the most important columns per dataset with name, information and value

| Dataset | Column name | Information | Value |
|---|-------------------|---|---|
| Road traffic accidents 2018 | vakav | Seriousness of the accident | 1 = accident resulting in death
2 = accident resulting in injury
3 = accident resulting in serious injury |
| Korkeusrajoitus
Leveysrajoitus
Näkemäpituus
Onnettomuusriski | ALIKKO | Height limit | Height in cm |
| | MAXLEV | Width limit | Width in cm |
| | NAKEMA | Sight distance | Distance in m |
| | RRO
RRK
RRV | Risk of light injury
Risk of serious injury
Risk of death | Amount of accidents per 100 billion vehicles passing the intersection |
| Palvelualueet | PATY | Service area | 1 = Rest area deluxe
2 = Rest area
3 = Private rest area
4 = Parking area deluxe
5 = Parking area
6 = Loading area
7 = Pier
8 = Other area |
| Suojatiet | VARO310 | Pedestrian crossing | 1 = with warning in advance
0 = without warning in advance |

| | geom | vvonn | kkonn | vakav |
|---|--|-------|-------|-------|
| 1 | POINT (60.24944757965415 25.163782757442434) | 2018 | 1 | 2 |
| 2 | POINT (60.23817205255151 24.858923628964465) | 2018 | 1 | 2 |
| 3 | POINT (60.21046093801623 25.05870117966777) | 2018 | 1 | 2 |
| 4 | POINT (60.169831696308414 24.92609015767353) | 2018 | 1 | 2 |
| 5 | POINT (60.24146305859899 24.849496287820408) | 2018 | 1 | 2 |

Figure 12. First 5 rows of the cleaned Road traffic accidents dataset.

3.4 Preparing the datasets for merging

In order to merge different datasets without complications a few requirements need to be met:

- Every dataset should have the same number of columns.
- Each column should only contain data of 1 data type.
- Every dataset should have the same column names.

(Osborne, 2013)

To fulfil these requirements a list of column names is defined. These column names are chosen in a way that the information from all the different cleaned datasets can be accessed in the merged version. Table 4 shows the list of chosen column names:

Table 4. An overview of the chosen column names for merging with explanations

| Column name | Information | Data type | Example |
|--------------|----------------------------|-----------|------------------------------------|
| lat | Coordinate latitude | Decimal | 65.2390465622 |
| long | Coordinate longitude | Decimal | 25.3866022237 |
| date | Latest date of measurement | Date | 2020-9-13 |
| municipality | Municipality | Text | Oulu |
| value_ref | Referenced value | Text | Accident resulting in light injury |
| value_num | Numerical value | Decimal | 1.53 |

Not every dataset has data for each column. These missing data cells receive an “NA” or Not Available value instead. Most notable missing data are in the “value_ref” and “value_num” columns. Most datasets that have numeric value data, don’t have referenced value data and vice versa. The Road traffic accidents dataset doesn’t have municipality data included.

Before merging the data, one extra step is required. An extra column called “type” is added to each dataset. The goal of this column is to give context to the data in the value-column by specifying which dataset the data originally came from.

In order to accommodate these changes, the Onnettomusriski dataset is split in 3 different datasets. The first one with the risk of light injury, the second one with the risk of serious injury and the last one with the risk of death as numeric value. This way each dataset now has one value column. The result is illustrated in figure 13.

| | lat | long | date | municipality | value_ref | value_num | type |
|---|--------------------|--------------------|--------|--------------|------------------------------------|-----------|---------------|
| 1 | 60.24944757965415 | 25.163782757442434 | 2018-1 | NA | accident resulting in light injury | NA | road_accident |
| 2 | 60.23817205255151 | 24.858923628964465 | 2018-1 | NA | accident resulting in light injury | NA | road_accident |
| 3 | 60.21046093801623 | 25.05870117966777 | 2018-1 | NA | accident resulting in light injury | NA | road_accident |
| 4 | 60.169831696308414 | 24.92609015767353 | 2018-1 | NA | accident resulting in light injury | NA | road_accident |
| 5 | 60.24146305859899 | 24.849496287820408 | 2018-1 | NA | accident resulting in light injury | NA | road_accident |

Figure 13. First 5 rows of the prepped Road traffic accidents dataset.

3.5 Merging the datasets

After the preparation, the data can be merged without further complications. The set of rows from each dataset is added to the final dataset one by one until the final dataset is complete. This dataset now contains all the required data from each dataset (figure 14).

The only thing that is added after merging is one extra column, “id”, with a unique identifier. This means that each row now has a unique number. This column is added for etiquette and is not necessary for the purpose of this thesis, but it makes the dataset viable for use across different data management software (Osborne, 2013).

| lat | long | date | municipality | value_ref | value_num | type | id |
|--------------------|--------------------|------------|--------------|-----------------|-----------|---|--------|
| 60.68207734267091 | 21.705681294116005 | 2020-04-01 | Vehmaa | NA | 4229.00 | intersection_accident_light_injury_risk | 275791 |
| 60.68207734267091 | 21.705681294116005 | 2020-04-01 | Vehmaa | NA | 343.00 | intersection_accident_serious_injury_risk | 288486 |
| 60.68207734267091 | 21.705681294116005 | 2020-04-01 | Vehmaa | NA | 168.00 | intersection_accident_death_risk | 301181 |
| 60.68210615322875 | 21.705957005173413 | 2020-04-01 | Vehmaa | not specified | NA | pedestrian_crossing | 316398 |
| 60.68217956997739 | 21.706780055685638 | 2020-04-01 | Vehmaa | not specified | NA | pedestrian_crossing | 316405 |
| 60.68228716798389 | 21.70829827790736 | 2020-04-01 | Vehmaa | NA | 206.00 | sight_distance | 251328 |
| 60.68233223782231 | 21.70859510275369 | 2020-04-01 | Vehmaa | NA | 60.00 | sight_distance | 253046 |
| 60.68361039933902 | 21.71151466248567 | 2020-04-01 | Vehmaa | NA | 191.00 | sight_distance | 252181 |
| 60.68384534841673 | 21.71188420413103 | 2020-04-01 | Vehmaa | NA | 122.00 | sight_distance | 252176 |
| 60.684410514641115 | 21.71245391052939 | 2020-04-01 | Vehmaa | NA | 100.00 | sight_distance | 252183 |
| 60.68475945816231 | 21.71265278167208 | 2020-04-01 | Vehmaa | without warning | NA | pedestrian_crossing | 316376 |
| 61.41797997285055 | 26.884697176362682 | 2020-04-01 | Mäntyharju | NA | 144.00 | sight_distance | 250039 |
| 61.418244428760715 | 26.886239502105635 | 2020-04-01 | Mäntyharju | NA | 145.00 | sight_distance | 250042 |
| 61.41858742773706 | 26.887558373188877 | 2020-04-01 | Mäntyharju | not specified | NA | pedestrian_crossing | 316395 |
| 61.41868682965211 | 26.88784264948832 | 2020-04-01 | Mäntyharju | NA | 103.00 | sight_distance | 252614 |

Figure 14. First 15 rows of the final dataset, sorted on date.

4 DATA AND ADVANCED ANALYTICS SOLUTION

Advanced analytics is the process of discovering your data and making them useful. Unfortunately, this process takes a long time to come into fruition. However, once it has reached maturity, it has the possibility of creating a sizable profit. (Russom, 2011)

The solution discussed in this chapter takes the first few steps in that process and can serve as a foundation to build upon. It showcases several specific cases that each look into the data from a different angle. These cases are thematical visualisations of the data but can also serve for basic analytical purposes.

An advantage of using WFS sources for data input is that these sources are live and constantly receive updates. This creates the opportunity to analyse data as it is being written, making sure that the solution always stays up to date. The caveat is that this requires technologies that are optimized for this sort of analytics. (Strong, 2017)

4.1 Solution overview

The solution is created in the Power BI Desktop application as an interactive report. This means that the data is not only visualised, but that the user can also further specify parameters in the visualisation (Microsoft, 2020). Each case receives a new page in this report and consists of several visualisations.

The database connection in Power BI is created through an R script. This R script is a simplified version of the scripts in the R notebook from Chapter 3 (See Appendix 1). Using this method, a clean separation is made between the data manipulation in R and the data visualisation and analytics in Power BI. Employing the latest data is done by pressing a refresh button in Power BI, which re-runs the R script.

Separating the solution in frontend and backend allows for the separation of tasks on professional basis. This means that a Data Architect can work alongside a Data Analyst without interference. This separation also increases the adaptability of this solution to newer technologies. A company could for example decide to replace the front end without creating the necessity to rewrite the backend data manipulation.

4.1.1 Case 1: Traffic accidents in Hämeenlinna and dangerous months

This case is used for analysing the traffic accidents in Hämeenlinna (figure 15). Only the data with "road_accident" as type-value are loaded. This

means that all the data displayed on this page comes from the Road traffic accidents 2018 dataset.

The map visualisation in this case shows the road traffic accidents in Hämeenlinna. These are then linked to the stack bar visualization on the right, which compares these data on a monthly scale. The three buttons in the bottom left are used to filter on the seriousness of the accident.

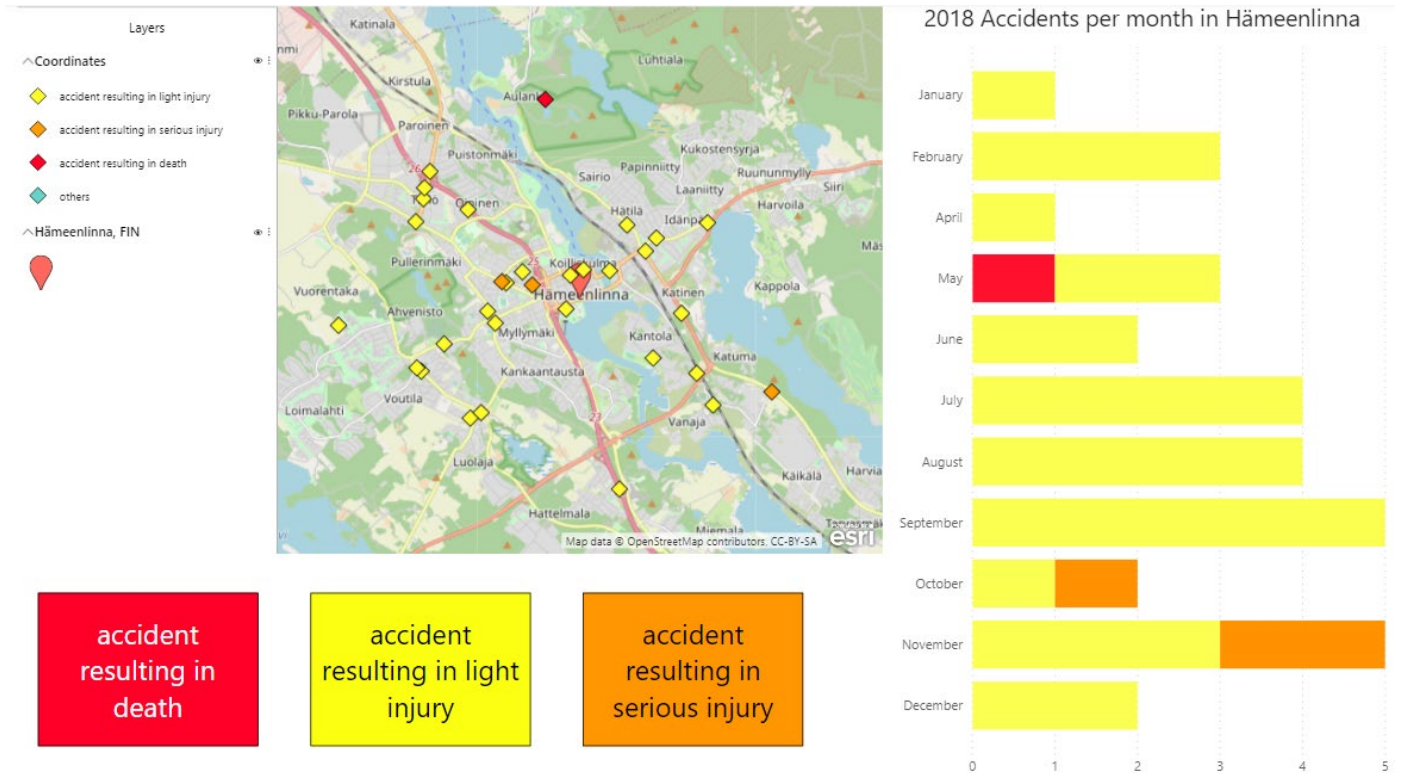


Figure 15. Case 1 Power BI report page

4.1.2 Case 2: Sight distance in Hämeenlinna with average comparison

This case is used for comparing roads on sight distance, using the average sight distance in Finland as a reference (figures 16 and 17). Only the "sight_distance" data are loaded. This dataset is quite big and thus, in order to increase the performance, the focus is put on one province: The Hämeenlinna municipality.

The map visualisation on this report page shows the sight distance measurements in the Hämeenlinna municipality. The gauge visualisation compares the average sight distance in Hämeenlinna to the Finnish average. The bar visualisation shows the sight distance of the selected measurement on the map to the Finnish average.

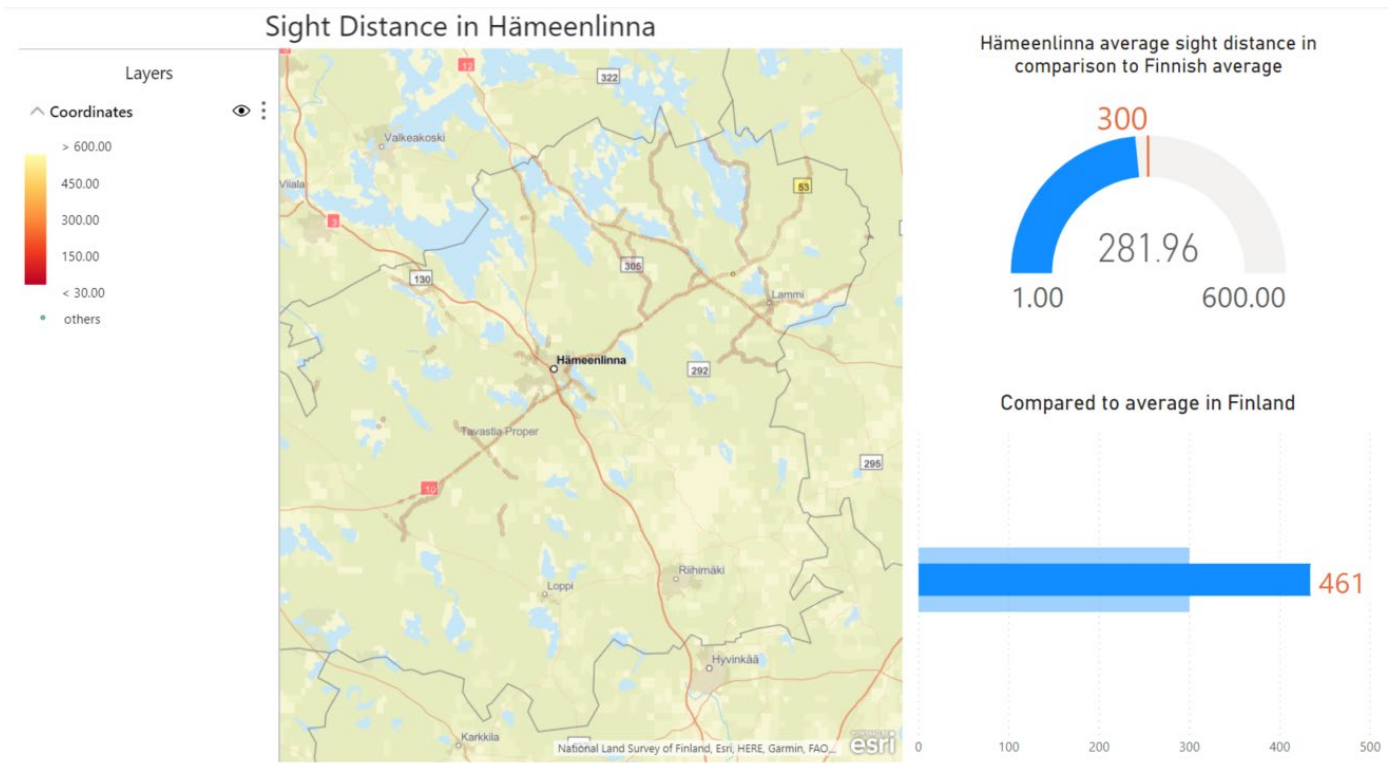


Figure 16. Case 2 Power BI report with one location selected

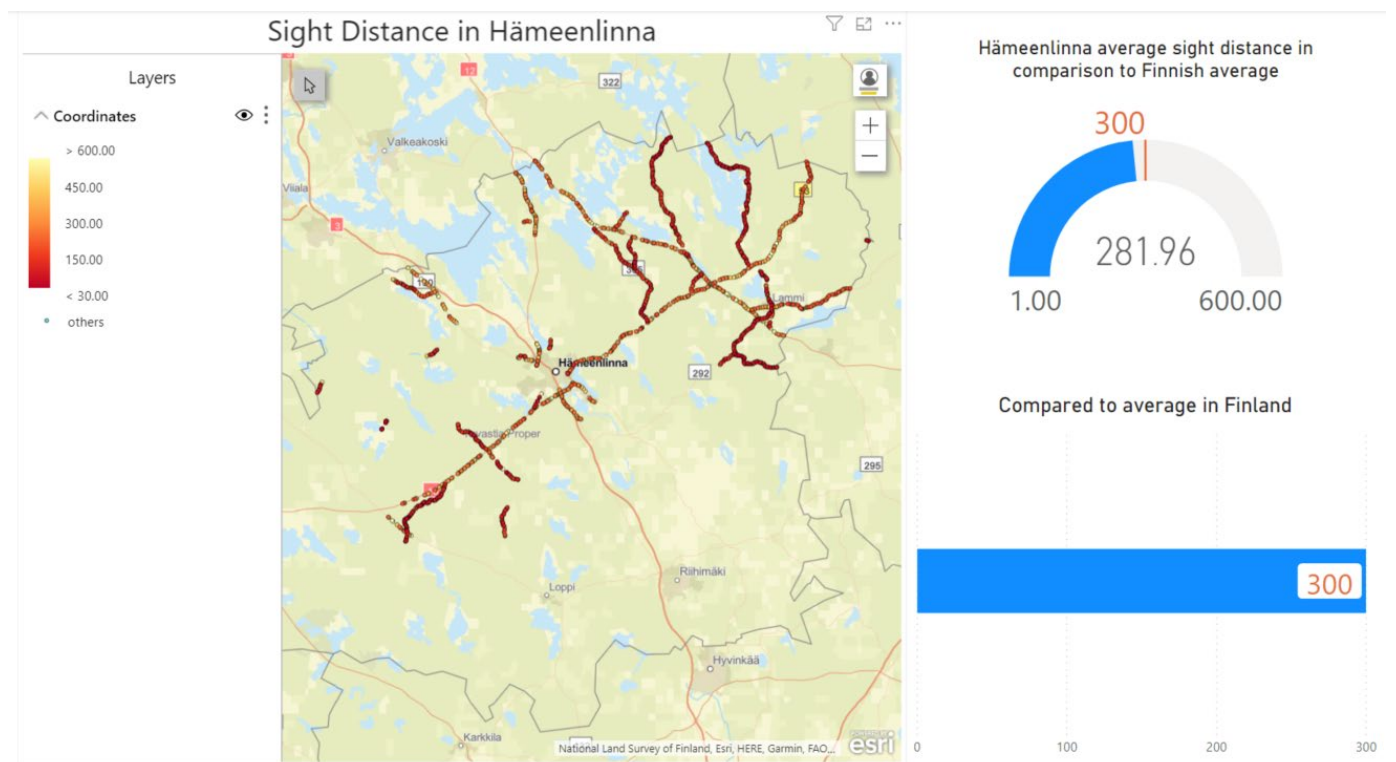


Figure 17. Case 2 Power BI report with no locations selected

4.1.3 Case 3: Height limit influence on accident risk at an intersection.

This case focuses on the situation near a certain intersection (figure 18). It considers the height limitation at the crossing and its influence on the risk of a serious accident at that crossing. Both “max_height” and “intersection_accident_serious_injury_risk” data are loaded. The crossing is located in the Hämeenlinna municipality.

There are 2 map visualisations focussed on the same intersection in this report page. One displays the height limit measurements at that intersection and the other the risk of having an accident with a serious injury. A gauge visualisation on the right compares the risk at this crossing to the average risk on intersections in the Hämeenlinna municipality.

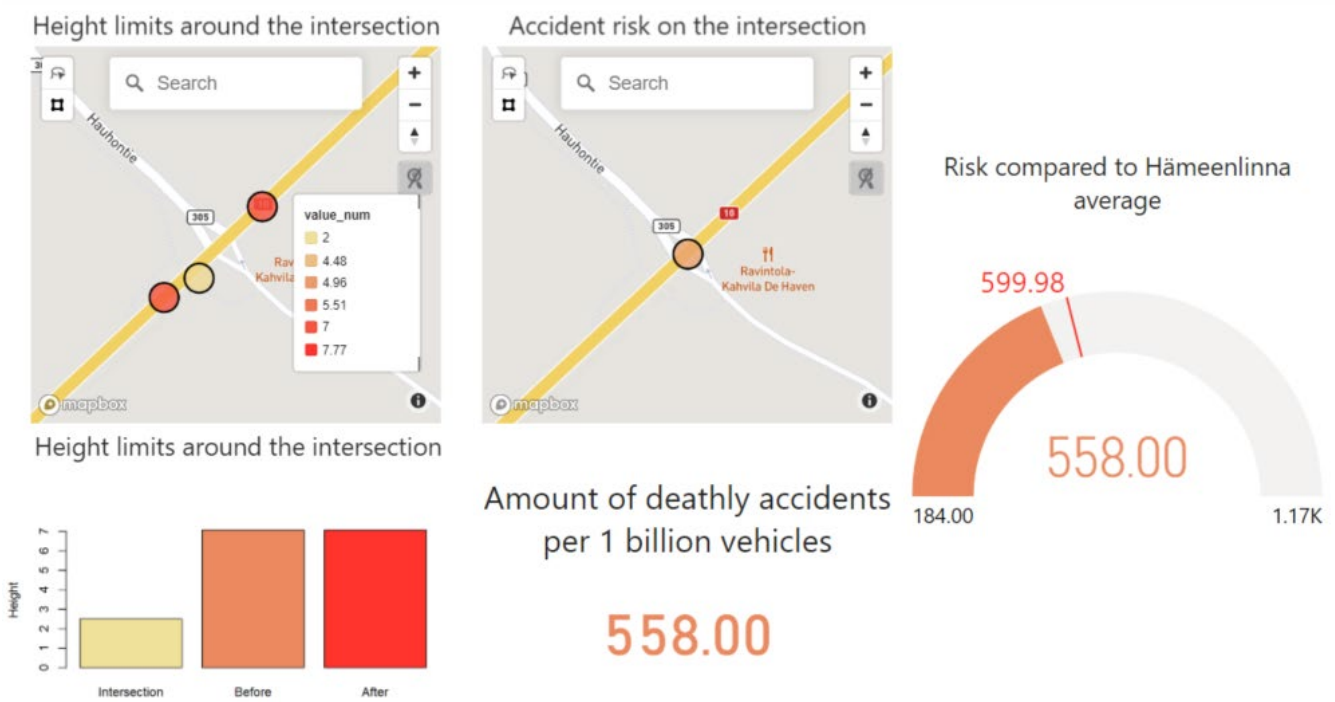


Figure 18. Case 3 Power BI report page

4.1.4 Case 4: Influences on accidents along a section of the road

This case focusses on the safety of a specific stretch of a road (figure 19). In the example a part of the road nr 312 near Villähde is showcased on the map. This focus on the map can be shifted and the rest of the visualisations will automatically adapt. The data loaded are “pedestrian_crossing”, “road_accident” and “service_area” type data.

There is one map visualisation on this page that displays the 3 types of data. Each type of data is further highlighted with visualisations that best display their values. First, the pedestrian crossing data are visualised in 2 different cards. One displays the most common warning situation and the other displays the number of pedestrian crossings in this map section. Next, the service area data is visualised with a multi-row card. This card displays the amount of service areas per category in this map section. Finally, the accident data is visualised in a pie chart. This chart creates a visual overview of the distribution between the levels of seriousness of the accidents.

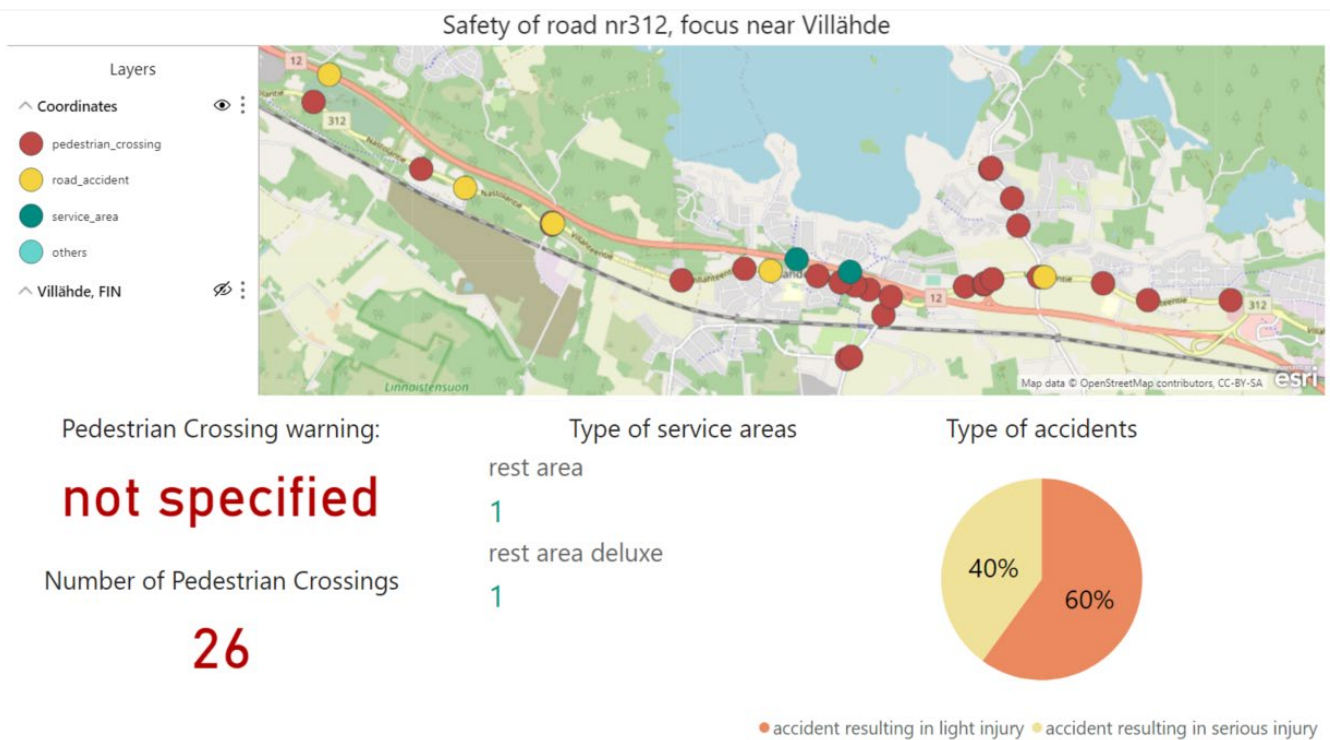


Figure 19. Case 4 Power BI report page

4.2 Business implications

Over the recent years, experts in the fields of big data, economics and finance have come to a general sentiment that the increased use of data in businesses is having a positive economic impact. The one caveat is that the data should be business specific and concrete in order to have this positive impact. In other words, it is better to have a smaller amount of specific and well-defined data than to have a huge amount of data without context. (Schroeder, 2016)

These observations have all been considered while creating this merged dataset and solution. The purpose of this merged dataset is well defined and specific: Analysing and classifying road risks. Each data source has detailed metadata documentation and only the essential data from each dataset was kept. This results in a clean and focused final dataset.

The main audience for this solution is the logistics sector. This is because one of the most crucial tasks in this sector is transport network planning (Jenelius, Petersen, & Mattsson, 2006). When creating such a network it is vital to factor in road risks as a cause of delays and transport disruption (Lin & Wei, 2019). This is also the main purpose of cases 3 and 4 in the solution. Both cases target a specific section of the road and make an analysis based on one or several risk factors. Subsequently case 2 can be used to find safer road sections based on a broader road overview.

Another potential audience for this solution is the Finnish Ministry of Transport. By analysing the road risks, targeted road improvements can be issued. This would improve the general safety of the road and thus reduce the amount of accidents and improve the efficiency of transport networks. To start off, Case 1 of the solution can be used to find accident hotspots. After these have been found, cases 3 and 4 can be used to analyse the causes of the accidents. (Shi & Abdel-Aty, 2015)

5 CONCLUSION

The goal in mind while creating this thesis was to create one dataset that can be used in visualisation and analytics software, such as Power BI, in order to classify road risks in Finland. This goal has been accomplished by merging data from different sources and documenting this merging process. Presented in this conclusion are the limitations that had an influence on the writing process of this thesis, a summary of the merging process itself and an overview of the result, based on the research questions.

5.1 The limitations

First, due to the amount of research required for understanding geographical data and data services, the scope of this thesis was limited to the merging process and a basic data solution. This also shifted the theory/practice text distribution more in favour of the theoretical part of this thesis.

Subsequently does Power BI also have its limitations. The software was publicly released on July 24th of 2015, which is relatively recent (The Power BI Team, 2015). This shows in the lack of more extensive functionalities, map visualisation in specific. Software such as Qlikview/QlikSense or Sisense should currently be considered as more viable alternatives.

A final limitation is the author's lack of knowledge about the Finnish language. Most of the data source documentation is written in Finnish, which created some difficulty in researching the specifications and metadata of the data. A lot of time was invested in making sure there was no misunderstanding about the data context.

5.2 The process

Data manipulation in R is quite straightforward and the code is highly adaptable for future expansions of the dataset. Using an R notebook also enables easy documentation of the code. This, in combination with the increased readability due to the packages included in Tidyverse, creates a nice overview of the merging process. This can be observed in Appendix 1.

The creation of the data solution in Power BI was a bit more complicated. The visualisations were quite limited in customizability and the linking between visualisations often didn't work as expected. The existing map visualisations are not originally made by the Power BI Team and thus lack decent integration.

5.3 The result

The goal of this thesis was to start construction on a road risk classification system. In order to bring structure to this research, 2 research questions were formulated. Now that the research has been completed, both questions have been answered. Represented here are the summaries to these answers.

5.3.1 What data sources are useful for road risk analysis?

Open governmental data sources, that contain transport network data and provide a WFS data service for deploying the data can be considered useful if the data contains coordinates in a Vector Point Shape format. This data can then easily be employed in a data solution for advanced analytics.

5.3.2 Which techniques can be used for merging data from different sources to create one dataset?

Datasets from different sources can easily be merged using the Tidyverse package in R, on condition that the following steps are followed in order:

1. Make a detailed analysis of the data source, it's datasets and their metadata.
2. Extract the datasets from each data source individually.
3. Keep only the essential data of each dataset, remove the rest.
4. Transform all datasets in order to have a uniform structure.
5. Merge the datasets and add an index column.

6 SUMMARY

Basic answers to both research questions have been formulated. This however does not mean that the research is completed. There are many alternatives to the elements of this thesis that have not been explored due to its limited timespan.

First of all, there exist other data services for geospatial data and of the ones discussed in this thesis, WMS has only been superficially researched. Subsequently, due to their complexity, other vector shape data types than Point data have been ignored. Finally, when it comes to software, Power BI was chosen for visualisation due to the author's familiarity with the program. This software however ended up having difficulties handling geospatial data.

However, one conclusion that can be taken from this thesis is that it is definitely possible to merge datasets from different sources for the purpose of classifying road risks. The final dataset is successfully created and employed in a data solution for this purpose. The next step is to systematically do further analysis of the data itself in order to find specific relations between the data and create a full-fledged road risk classification.

REFERENCES

Duivenvoorde, R. (2019, November 20). *qgisnetworklogger/README*. Retrieved March 3, 2020, from <https://github.com/rduivenvoorde/qgisnetworklogger/blob/master/README.md>

Esri. (2020, January 13). *ArcGIS Maps for Power BI*. Retrieved March 3, 2020, from ArcGIS Documentation: <https://doc.arcgis.com/en/maps-for-powerbi/>

Finnish Transport Infrastructure Agency. (2015, October 27). *Materials*. Retrieved February 29, 2020, from Finnish Transport Infrastructure Agency: <https://vayla.fi/web/en/open-data/materials>

Finnish Transport Infrastructure Agency. (2018, December 31). *Open Data*. Retrieved February 20, 2020, from Finnish Transport Infrastructure Agency: <https://vayla.fi/web/en/open-data>

Finnish Transport Infrastructure Agency. (2019, December 20). *Tierekisteri Tietosisallon Kuvaus*. Retrieved February 29, 2020, from Finnish Transport Infrastructure Agency: https://julkaisut.vayla.fi/tierekisteri/tierekisteri_tietosisallon_kuvaus.pdf

GISGeography. (2018, February 18). *Vector vs Raster: What's the Difference Between GIS Spatial Data Types?* Retrieved February 29, 2020, from GIS Geography: <https://gisgeography.com/spatial-data-types-vector-raster/>

Grolemund, G., & Wickham, H. (2016). *R for Data Science*. O' Reilly Media. Retrieved April 18, 2020, from <https://r4ds.had.co.nz/index.html>

Iseminger, D. (2020, January 7). *Data types in Power BI Desktop*. Retrieved February 29, 2020, from Microsoft Docs: <https://docs.microsoft.com/en-us/power-bi/desktop-data-types>

Jenelius, E., Petersen, T., & Mattsson, L.-G. (2006, August). Importance and exposure in road network vulnerability analysis. *Transportation Research Part A: Policy and Practice*, 40(7), 537 - 560. Retrieved April 18, 2020, from <http://www.sciencedirect.com/science/article/pii/S096585640500162X>

Klokan Technologies GmbH. (2007, August 27). *EPSG:4326*. Retrieved April 4, 2020, from epsg.io: <https://epsg.io/4326>

Lin, H.-Z., & Wei, J. (2019). Optimal transport network design for both traffic safety and risk equity considerations. *Journal of Cleaner Production*, 218, 738 - 745. Retrieved April 4, 2020, from <http://www.sciencedirect.com/science/article/pii/S095965261930455X>

- Mapbox. (n.d.). *Create data visualizations with the Mapbox Visual for Power BI*. Retrieved March 3, 2020, from Mapbox | Docs: <https://docs.mapbox.com/help/tutorials/power-bi/>
- Microsoft. (2020, February 28). *Power BI Desktop-Interactive Reports*. Retrieved March 1, 2020, from Power BI: <https://powerbi.microsoft.com/en-us/desktop/>
- Miller, H. J., & Shaw, S.-L. (2001). *Geographic information systems for transportation: principles and applications*. Oxford University Press on Demand.
- Open Geospatial Consortium Inc. (2006, March 15). OpenGIS® Web Map Server Implementation Specification. Retrieved from http://portal.opengeospatial.org/files/?artifact_id=14416
- Open Geospatial Consortium Inc. (2010, November 2). OpenGIS Web Feature Service 2.0 Interface Standard. Retrieved from http://portal.opengeospatial.org/files/?artifact_id=14416
- Open Knowledge Foundation. (n.d.). *Why open data?* Retrieved February 20, 2020, from Open Knowledge Foundation: <https://okfn.org/opendata/why-open-data/>
- Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.
- Paikkatietohakemisto. (n.d.). *Road traffic accidents*. Retrieved February 29, 2020, from Paikkatietohakemisto: <https://www.paikkatietohakemisto.fi/geonetwork/srv/eng/catalog.search#/metadata/de71e0a1-4516-4d50-bd54-e384e5174546>
- Popell, C., & Iseminger, D. (2019, June 27). *Power Query*. Retrieved March 2, 2020, from Microsoft Docs: <https://docs.microsoft.com/en-us/power-query/>
- QGIS Development Team. (2020, March 3). QGIS User Guide. (3.4). Retrieved March 3, 2020, from <https://docs.qgis.org/3.4/pdf/en/QGIS-3.4-UserGuide-en.pdf>
- RStudio. (n.d.). *RStudio IDE Features*. Retrieved March 3, 2020, from RStudio: <https://rstudio.com/products/rstudio/features/>
- Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19(4)*, 1-34.
- Schroeder, R. (2016). Big data business models: Challenges and opportunities. *Cogent Social Sciences*, 2(1), 1166924. Retrieved April 18, 2020, from <https://www.tandfonline.com/doi/pdf/10.1080/23311886.2016.1166924>
- Shi, Q., & Abdel-Aty, M. (2015). Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58(Part B), 380 - 394. Retrieved April 18, 2020, from <https://doi.org/10.1016/j.trc.2015.02.022>

Statistics Finland. (2018). *Statistics Finland*. Retrieved February 20, 2020, from Statistics Finland:
http://www.stat.fi/static/media/uploads/org_en/infograafi/organisaatioinfograafi_enuku_2019_tk-sivut_500x.svg

Statistics Finland. (2019, August 8). *Statistics Finland*. Retrieved 02 29, 2020, from Statistics Finland: http://www.stat.fi/org/index_en.html

Statistics Finland. (n.d.). *Tieliikenneonnettomuudet*. Retrieved February 29, 2020, from Statistics Finland:
<http://www.stat.fi/org/avoindata/paikkatietoaineistot/tieliikenneonnettomuudet.html>

Strong, C. (2017, August 8). The Value Of Real-Time Data Analytics. *Forbes*. Retrieved April 11, 2020, from
<https://www.forbes.com/sites/forbestechcouncil/2017/08/08/the-value-of-real-time-data-analytics/#4ad83b6b1220>

The Power BI Team. (2015, July 10). *Announcing Power BI general availability coming July 24th*. Retrieved April 18, 2020, from Microsoft Power BI Blog:
<https://powerbi.microsoft.com/en-us/blog/announcing-power-bi-general-availability-coming-july-24th/>

The R Foundation. (n.d.). *What is R?* Retrieved March 3, 2020, from R Project:
<https://www.r-project.org/about.html>

Xie, Y., Allaire, J. J., & Golemund, G. (2019, December 2). R Markdown: The Definitive Guide. Retrieved March 3, 2020, from <https://bookdown.org/yihui/rmarkdown/>

R notebook

Suomi Traffic Accidents Data Manipulation in R

Variables declaration

```
wfs_accidents <- "http://geo.stat.fi/geoserver/tieliikenne/wfs?"  
wfs_road <- "https://julkinen.vayla.fi/inspirepalvelu/avoin/wfs?"  
"
```

Data source discovery

Layers

```
library(gdalUtils)  
info <- ogrinfo(paste("WFS", wfs_accidents, sep = ":"), so = TRUE,  
E, ro = TRUE)
```

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse  
rse 1.3.0 --
```

```
## v ggplot2 3.3.0      v purrr   0.3.3  
## v tibble  3.0.0      v dplyr   0.8.5  
## v tidyr   1.0.2      v stringr 1.4.0  
## v readr   1.3.1      v forcats 0.5.0
```

```
## -- Conflicts ----- tidyverse_co  
nflicts() --
```

```
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()     masks stats::lag()
```

```
info %>%  
  str_subset(pattern = "Point") %>%  
  head(5) %>%  
  cat(sep = "\n")
```

```
## 1: tieliikenne:tieliikenne_2011 (title: Tieliikenneonnettomuu  
det 2011) (Point)  
## 2: tieliikenne:tieliikenne_2012 (title: Tieliikenneonnettomuu  
det 2012) (Point)  
## 3: tieliikenne:tieliikenne_2013 (title: Tieliikenneonnettomuu  
det 2013) (Point)  
## 4: tieliikenne:tieliikenne_2014 (title: Tieliikenneonnettomuu  
det 2014) (Point)  
## 5: tieliikenne:tieliikenne_2015 (title: Tieliikenneonnettomuu  
det 2015) (Point)
```

Layer example

```
library(httr)
query <- list(service = "WFS",
              request = "GetFeature",
```

Appendix 1/2

```
              version = "2.0.0",
              typeName = "tieliikenne:tieliikenne_2018",
              outputFormat = "csv",
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_accidents, query = query)
road_accidents_df <- read.csv(textConnection(content(result, 'text')))
head(road_accidents_df, 5)
```

```
##          FID
geom vvonn kkonn
## 1 tieliikenne_2018.1 POINT (60.24944757965415 25.163782757442
434) 2018 1
## 2 tieliikenne_2018.2 POINT (60.23817205255151 24.858923628964
465) 2018 1
## 3 tieliikenne_2018.3 POINT (60.21046093801623 25.05870117966
777) 2018 1
## 4 tieliikenne_2018.4 POINT (60.169831696308414 24.92609015767
353) 2018 1
## 5 tieliikenne_2018.5 POINT (60.24146305859899 24.849496287820
408) 2018 1
##          kello vakav onntyyppi lkmhapa lkmlaka lkmjk lkmpp lkm
mo lkmmp
## 1 17.00-17.59      2          8          1          0          0          0
0 0
## 2 20.00-20.59      2          8          2          0          0          0
0 0
## 3 17.00-17.59      2          4          1          0          0          1
0 0
## 4 03.00-03.59      2          9          1          0          1          0
0 0
## 5 12.00-12.59      2          0          4          0          0          0
0 0
##  lkmmuukulk          x          y
## 1          0 398360.9 6680606
## 2          0 381449.2 6679860
## 3          0 392417.8 6676432
## 4          0 384928.2 6672132
## 5          0 380939.2 6680243
```

Dataset manipulation

Getting

```
library(httr)
```

```
query <- list(service = "WFS",
              request = "GetFeature",
              version = "2.0.0",
              typeNames = "tieliikenne:tieliikenne_2018",
              outputFormat = "csv",
```

Appendix 1/3

```
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_accidents, query = query)
road_accidents_df <- read.csv(textConnection(content(result, 'text')))
```

```
query <- list(service = "WFS",
              request = "GetFeature",
              version = "2.0.0",
              typeNames = "avoin:TL263",
              outputFormat = "csv",
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)
height_df <- read.csv(textConnection(content(result, 'text')))
```

```
query <- list(service = "WFS",
              request = "GetFeature",
              version = "2.0.0",
              typeNames = "avoin:TL264",
              outputFormat = "csv",
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)
width_df <- read.csv(textConnection(content(result, 'text')))
```

```
query <- list(service = "WFS",
              request = "GetFeature",
              version = "2.0.0",
              typeNames = "avoin:TL113",
              outputFormat = "csv",
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)
sight_df <- read.csv(textConnection(content(result, 'text')))
```

```
query <- list(service = "WFS",
              request = "GetFeature",
              version = "2.0.0",
              typeNames = "avoin:TL232",
              outputFormat = "csv",
              srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)
intersection_accidents_df <- read.csv(textConnection(content(result, 'text')))
```

```
query <- list(service = "WFS",
              request = "GetFeature",
```

```

    version = "2.0.0",
    typeName = "avoin:TL195",
    outputFormat = "csv",
    srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)

```

Appendix 1/4

```

service_df <- read.csv(textConnection(content(result, 'text')))

query <- list(service = "WFS",
  request = "GetFeature",
  version = "2.0.0",
  typeName = "avoin:TL310",
  outputFormat = "csv",
  srsname = "urn:ogc:def:crs:EPSG::4326")
result <- GET(wfs_road, query = query)
pedestrian_df <- read.csv(textConnection(content(result, 'text')
))

```

Cleaning

```

df <- road_accidents_df
cnames <- c("geom", "vvonn", "kkonn", "vakav")
# 1 = accident resulting in death
# 2 = accident resulting in injury
# 3 = accident resulting in serious injury
road_accidents_df_cleaned <- df[cnames]

df <- height_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "ALIKKO")
# height in cm
height_df_cleaned <- df[cnames]

df <- width_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "MAXLEV")
# width in cm
width_df_cleaned <- df[cnames]

df <- sight_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "NAKEMA") #
# sight distance in m
sight_df_cleaned <- df[cnames]

df <- intersection_accidents_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "RRO", "RRK", "RRV")
# per 100 billion vehicles
# Simplifying the columns
df$RRO = df$RROKE + df$RROKO + df$RROMU
df$RRK = df$RRKKE + df$RRKKO + df$RRKMU
df$RRV = df$RRVKE + df$RRVKO + df$RRVMU
intersection_accidents_df_cleaned <- df[cnames]

```

```
df <- service_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "PATY")
# 1 = Rest area deluxe
# 2 = Rest area
# 3 = Private rest area
# 4 = Parking area deluxe
```

Appendix 1/5

```
# 5 = Parking area
# 6 = Loading area
# 7 = Pier
# 8 = Other area
service_df_cleaned <- df[cnames]

df <- pedestrian_df
cnames <- c("SHAPE", "MUUTOSPVM", "KUNTA", "VARO310") # with or
without warning
pedestrian_df_cleaned <- df[cnames]
```

Prepping

```
library(tidyverse)
```

```
df <- road_accidents_df_cleaned
road_accidents_df_prepped <- df %>%
  extract(geom, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
[.][0-9]+)") %>%
  unite("date", c(vvonn, kkonk), sep = "-") %>%
  mutate(vakav = case_when(
    vakav == 1 ~ "accident resulting in death",
    vakav == 2 ~ "accident resulting in light injury",
    vakav == 3 ~ "accident resulting in serious injury"
  )) %>%
  rename(value_ref = vakav) %>%
  add_column(municipality = NA, .after = "date") %>%
  add_column(value_num = NA, .after = "value_ref") %>%
  add_column(type = "road_accident")
```

```
df <- height_df_cleaned
height_df_prepped <- df %>%
  extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
[.][0-9]+)") %>%
  mutate(ALIKKO = ALIKKO / 100) %>%
  rename(date = MUUTOSPVM, municipality = KUNTA, value_num = ALI
KKO) %>%
  add_column(value_ref = NA, .before = "value_num") %>%
  add_column(type = "max_height")
```

```
df <- width_df_cleaned
width_df_prepped <- df %>%
  extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
[.][0-9]+)") %>%
```

```

mutate(MAXLEV = MAXLEV / 100) %>%
rename(date = MUUTOSPVM, municipality = KUNTA, value_num = MAX
LEV) %>%
add_column(value_ref = NA, .before = "value_num") %>%
add_column(type = "max_width")
df <- width_df_cleaned

```

Appendix 1/6

```

df <- sight_df_cleaned
sight_df_prepped <- df %>%
extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
).[0-9]+)") %>%
rename(date = MUUTOSPVM, municipality = KUNTA, value_num = NAK
EMA) %>%
add_column(value_ref = NA, .before = "value_num") %>%
add_column(type = "sight_distance")

```

```

df <- intersection_accidents_df_cleaned
intersection_accidents_light_df_prepped <- df %>%
extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
).[0-9]+)") %>%
rename(date = MUUTOSPVM, municipality = KUNTA, value_num = RRO
) %>%
select(-RRK, -RRV) %>%
add_column(value_ref = NA, .before = "value_num") %>%
add_column(type = "intersection_accident_light_injury_risk")

```

```

df <- intersection_accidents_df_cleaned
intersection_accidents_serious_df_prepped <- df %>%
extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
).[0-9]+)") %>%
rename(date = MUUTOSPVM, municipality = KUNTA, value_num = RRV
) %>%
select(-RRK, -RRO) %>%
add_column(value_ref = NA, .before = "value_num") %>%
add_column(type = "intersection_accident_serious_injury_risk")

```

```

df <- intersection_accidents_df_cleaned
intersection_accidents_deathly_df_prepped <- df %>%
extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}
).[0-9]+)") %>%
rename(date = MUUTOSPVM, municipality = KUNTA, value_num = RRK
) %>%
select(-RRV, -RRO) %>%
add_column(value_ref = NA, .before = "value_num") %>%
add_column(type = "intersection_accident_death_risk")

```

```

df <- service_df_cleaned
service_df_prepped <- df %>%
extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2}

```

```

}][.][0-9]+)") %>%
  mutate(PATY = case_when(
    PATY == 1 ~ "rest area deluxe",
    PATY == 2 ~ "rest area",
    PATY == 3 ~ "private rest area",
    PATY == 4 ~ "parking area deluxe",
    PATY == 5 ~ "parking area",
    PATY == 6 ~ "loading area",

```

Appendix 1/7

```

    PATY == 7 ~ "pier",
    PATY == 8 ~ "other area"
  )) %>%
  rename(date = MUUTOSPVM, municipality = KUNTA, value_ref = PAT
Y) %>%
  add_column(value_num = NA, .after = "value_ref") %>%
  add_column(type = "service_area")

df <- pedestrian_df_cleaned
pedestrian_df_prepped <- df %>%
  extract(SHAPE, c("lat", "long"), "([0-9]{2}[.][0-9]+) ([0-9]{2
}][.][0-9]+)") %>%
  mutate(VAR0310 = case_when(
    VAR0310 == 0 ~ "without warning",
    VAR0310 == 1 ~ "with warning",
    is.na(VAR0310) ~ "not specified"
  )) %>%
  rename(date = MUUTOSPVM, municipality = KUNTA, value_ref = VAR
0310) %>%
  add_column(value_num = NA, .after = "value_ref") %>%
  add_column(type = "pedestrian_crossing")

```

Merging

```
library(tidyverse)
```

```

merged_df <- bind_rows(list(
  road_accidents_df_prepped,
  height_df_prepped,
  width_df_prepped,
  sight_df_prepped,
  intersection_accidents_light_df_prepped,
  intersection_accidents_serious_df_prepped,
  intersection_accidents_deathly_df_prepped,
  service_df_prepped,
  pedestrian_df_prepped
)) %>%
  mutate(id = row_number())
head(merged_df, 5)

```

```

##           lat           long   date municipality
## 1  60.24944757965415 25.163782757442434 2018-1      <NA>

```


| | | | | |
|------|------------------------------------|--------------------|-----------|---------------|
| ## 2 | 60.23817205255151 | 24.858923628964465 | 2018-1 | <NA> |
| ## 3 | 60.21046093801623 | 25.05870117966777 | 2018-1 | <NA> |
| ## 4 | 60.169831696308414 | 24.92609015767353 | 2018-1 | <NA> |
| ## 5 | 60.24146305859899 | 24.849496287820408 | 2018-1 | <NA> |
| ## | | value_ref | value_num | type |
| id | | | | |
| ## 1 | accident resulting in light injury | | NA | road_accident |
| 1 | | | | |
| ## 2 | accident resulting in light injury | | NA | road_accident |

Appendix 1/8

| | | | | |
|------|------------------------------------|--|----|---------------|
| 2 | | | | |
| ## 3 | accident resulting in light injury | | NA | road_accident |
| 3 | | | | |
| ## 4 | accident resulting in light injury | | NA | road_accident |
| 4 | | | | |
| ## 5 | accident resulting in light injury | | NA | road_accident |
| 5 | | | | |