ZEPPELIN UNIVERSITY

BACHELOR

Different Member State, Different Governance Level, Different Data Quality?

Author:
Kilian LEHN

Examiner:
Dr. Prof. Michael SCHARKOW

Term: Spring Term 2020

Course of Studies: PAIR

Student Number: 16203015

June 25, 2020

Declaration of Authorship

In accordance with §20 Abs. 4 ASPO, I, Kilian LEHN, declare that this thesis titled,
"Different Member State,
Different Governance Level,
Different Data Quality?" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated. The Section Manual (0.1) is excluded from this statement.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

- I'm aware that a false statement will lead to a legal prosecution.

Signed:

Date: 25.06.2020

"At least, every individual must act as if the whole future of the world, of humanity itself, depends on him. Anything less is a shirking of responsibility and is itself a dehumanizing force, for anything less encourages the individual to look upon himself as a mere actor in a drama written by anonymous agents, as less than a whole person, and that is the beginning of passivity and aimlessness." (Weizenbaum, 1976)

"Simple problems beget simple solutions." (Banerjee and Duflo, 2013)

"Comprendre au lieu de juger." (Camus, 1957)

ZEPPELIN UNIVERSITY

Abstract

Politics, International Relations, Public Administration

Bachelor of Arts

Different Member State,
Different Governance Level,
Different Data Quality?

by Kilian LEHN

This work is an audit of the application of Principle 14 (cf. *Eurostat* 2020) of the "European statistics code of practice", which states that "European Statistics are consistent internally, over time and comparable between regions and countries". Data auditing follows the school of the science of quality improvement (cf. Balestracci, 2006) and describes the review of the sanity of data (sets) (cf. Kolb, 2010 and Polyzotis et al., 2018 and Hynes, Sculley, and Terry, 2017) for a certain purpose. This audit consists of a sample of French and German PM10-Emission data sets from each governance level (regional, national, European). A validation schema was developed for this purpose, which found unsystematic inconsistencies both in the vertical comparison of the governance levels in one Member State and in the comparison between two Member States (France and Germany). The data does not support the hypothesis that the count of governance level is at the root of the the discrepancies found in PM10-Emission Data sets, and it suggests that further research is necessary to accurately define the confounding variables responsible for these discrepancies.

Acknowledgements

I thank every being who sticks to the method even if this means that the outcome isn’t what was initially desired.

Contents

List of Figures

List of Tables

List of Abbreviations

cf.	**c**onfer = **V**ergleiche!
DC	**D**ata **C**onsistancy
DS	**D**ata **S**et
DV	**D**ependant **V**ariable
EEA	**E**uropean **E**nvironmental **A**gency
eu	**E**uropean
E.g.	**E**xempli **G**ratia
FRA	**FRA**ance
GER	**GER**many
GL	**G**overnance **L**evel
MA	**M**oving **A**verage
MSt	**M**easuring **St**ation
nat	**nat**ional
reg	**reg**ional
i.e	**i**d **e**st
IV	**I**ndependant **V**ariable
viz.	**v**idelicet = **n**amely

0.1 Manual

This work aims to be reproducible and easily comprehended. Concretely, this means that the reader who is comfortable with Code Editors like R-Studio or Visual Studio Code etc., can open this document in its .Rmd version and reproduce the computational/statistical tests done in this thesis. The reader using self-downloaded data sets should use different programs (Excel/R-Studio/SPSS) to check for data corruption caused by the importing process. File formats and separators as well as row and column nomenclature differ between data sets and can lead to erroneous values. To get the identical results as stated in this thesis is caused by the loading of the data sets (hence fourth described as DS(s)). Execute the following steps to replicate the procedure used in this analysis-report:

1. Download the raw data
2. Open the main .Rmd Version
3. Set the working directory path to the downloaded and extracted zip-folder of the raw data or put the .Rmd-File in a main-folder with the sub-folder "raw data", where the data sets are included (fastest solution)
4. Proceed through the Code-Chunks (If the Editor is R-Studio, use "Show document outline")
5. Compare with the compiled version[1] in order to see if everything works and to get a feeling how the monolithic script is structured

Furthermore, every citation is linked. This means that the reader can get more information by clicking on the hyperlinks. Generally speaking, every coloured letter in this thesis is hyperlinked, from the title page to the bibliography, so if one seeks to know more, one need only click on the coloured letters. If compiler issues arise, please consider going to this GitHub account after 01.09.20 in order to troubleshoot any problem. Graphics are in PDF form and can be magnified as needed.

[1] Packages can be shielded from one another, which leads to bizarre outcomes. However, if the reader loads the R-Chunks per se, the results should be good.

Chapter 1

Fundamental Assumptions

1.1 Introduction

In response to the climate debate, data-driven decisions are increasingly relied upon in Europe. These technocratic decisions (as they are often called in a political context)(cf. Radaelli, 2017) are behind national sanctions like the ones in force in the German city Stuttgart for exceeding the threshold value for air pollution. Because of the all-englobing nature of the problem of climate change, nearly every sector (primary, secondary and tertiary sector[1]) of the economy is touched. But do all stakeholders in these sectors have access to reliable data? Does any citizen, policy maker, or scientist in Europe have access to consistent and reliable data? Does the mayor of a small village in Germany have access to the same environmental data as a policy maker in the eu or any other mayor in the eu. These questions are addressed in the work at hand for the member states France and Germany. The accessibility of data is critically necessary for the social acceptance of the legislation of increasingly stringent environmental measures (cf. Carrete et al., 2012). This principle dovetails with one of the main principles of the Balanced-Scorecard-Topic: ” If you can’t measure it, you can’t manage it. ”(Kaplan and David, 2000, p. 21) or in this context: you can’t make effective managerial decisions without reliable access to consistent data.

1.2 State of Art

The communication[1] "A European strategy for data" was published the 19th February 2020. And the conceptualization of this work preceded by 20 days the publishing of the official communication. The recommendations put forward in the "A European strategy for data" (cf. *COM(2020) 66 final* 2020) confirms what is stated in the introduction (see 1.1) of this work, i.e., that all the stakeholder should have access to the same data in order to make coherent and reliable managerial decisions. The communication states:"Cross-sectoral (or horizontal) measures for data access and use should create the necessary over-arching framework for the data-agile economy, thereby avoiding harmful fragmentation of the internal market through inconsistent actions between sectors and between the Member States." (*COM(2020) 66 final* 2020, p. 12). This means for the work at hand, that it contributes to a new framework. This contribution consists of of "pattern-discovery", and this will be elaborated in the following sections. However, in the following sections it will also be shown, that tools/best practices to reach the goal of the "over-arching framework"[2] (*COM(2020)*

[1] Basic economical sectors, primary: Raw-materials, secondary: Manufacturing, tertiary: Service

[1] In the European context, "communications" between European institutions pave the way to new legislations.

[2] For further information on the "over-arching framework" in the context of environmental data, see *COM(2020) 66 final* 2020, p. 22, 26.

66 final 2020, p. 12) are already included in the *ESS Handbook* 2018. Hence, it might be possible for this work to build an use case[3] for the member states France and Germany of the mentioned "over-arching framework" by referring to the established Principle 14 (cf. *Eurostat* 2020).

1.2.1 Research Gap

The following work is, in the European context, best characterized as one of the "activities of other nature", i.e., "compliance monitoring which is a set of governance processes that are meant to stimulate EU Member States to respect their obligation to apply systematically the EU statistical legislation." (*ESS Handbook* 2018, p. 6). Whereas, in the terms of a general scientific context (cf. Colquitt and Zapata-Phelan, 2007), this work bears toward an examination of Principle 14 (cf. *Eurostat* 2020) with previously unexplored relationships, which in this context would be the application of the developed schema (Sufficient Condition) by testing the potentially explaining variable "Count of governance level" with existing validation approaches (1.1).

1.3 Deduction of Hypotheses

Principle 14 of the "European Statistics Code of Practice" states: "European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources." (cf. *Eurostat* 2020). If the 14th Principle (cf. Estat:eurostat, 2018) is accurate and true, then one can assume that data sets (henceforth, described as DSs) are comparable as the different statistical providers (e.g. regional or national authorities) have used the same harmonised statistical system. But is this the case? In order to answer this question, following hypotheses are formulated:

1.3.1 Research Question

Is there a correlation between a greater count of governance levels (henceforth, described as GL(s)) and inconsistencies between the DSs procured from the homepages of the respective GL?

H_0 = null hypothesis, H_1= alternative hypothesis

H_0 : Assuming that the tested sample lies within the confidence-interval, there is no correlation between a greater count of GL (see Chapter 3) and a degradation of the Data Consistency (henceforth, described as DC) with the confidence of 0.95.

H_1 : With the possibility of less than 0.05 the tested sample doesn't lie within the confidence interval, which suggests that the null-hypothesis is wrong while not suggesting that the contrary is true.

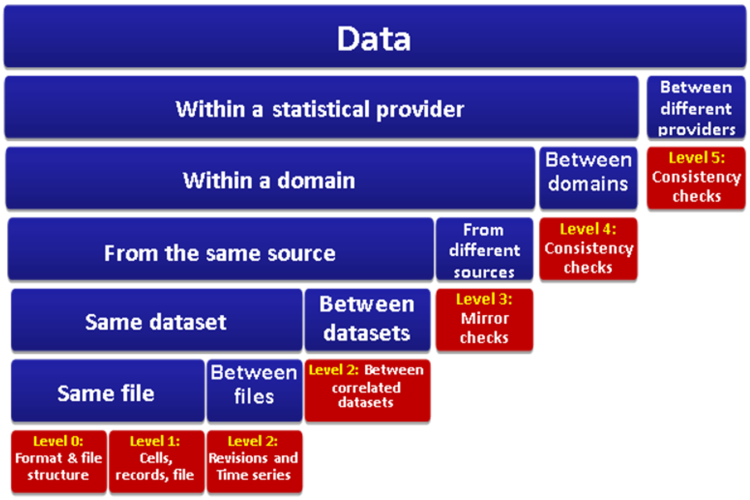

Data
Within a statistical provider
Between different providers
Within a domain
Between domains
Level 5: Consistency checks
From the same source
From different sources
Level 4: Consistency checks
Same dataset
Between datasets
Level 3: Mirror checks
Same file
Between files
Level 2: Between correlated datasets
Level 0: Format & file structure
Level 1: Cells, records, file
Level 2: Revisions and Time series

FIGURE 1.1: Levels DC-Check
ESS Handbook 2018, p. 13

1.3.2 Implementation in eu Framework

The hypotheses stated in the Section 1.3.1 qualify this work as an application of Level 3. The Figure 1.1 should help to place level 3 (the topic of this work) in the overall European-Analysis-Framework. Level 3 is placed between the two poles: Level: "Data" and Level "Same File", whereas the application of the Level: "Data" would consist of making a DC-Check between all data available on the worldwide web. While an analysis of the other pole of the Framework called Level "Same File", consists of comparing the statistical operations made in one and the same file.*"Validation levels 3 is concerned with the check of consistency based on the comparison of the content of the file with the content of "other files" referring to a different data provider on the same harmonised statistical system or domain (sharing common standards with respect to scope, definitions, units and classifications in the different surveys and sources)" (*ESS Handbook *2018, p. 15)*. This level 3 is operationalized in the Chapter 2 through a schema and applied to samples of Germany (henceforth, described as GER) and France (henceforth, described as FRA) DSs (see Chapter 2).

1.4 Research Steps

In the Table 1.1, the research steps conducted in this work are displayed. For the reader who is familiar with empirical works[4], this table might help to navigate through this paper.

1.4.1 Nature of Experiment

The lack of randomization and manipulation of the naturally occurring variables (GL/ State-Structure of the Member State) qualifies this work as a Non-Experiment (cf. Morgan and Winship, 2015, p. 7, 37 ; Bryman, 2008, p. 35-43) (see Section 6.3 for the advantages/disadvantages of this kind of experiment). However, in order

[3] "Each use case is a specific way of using the system [here the over-arching framework] and every execution may be viewed as an instance of the use case" (Jacobson, 1993, p. 129)

[4] "Empirical" means here a work based on data (a data analysis).

N^O	Research Step	Where to find
1	Stratified sampling	State Comparison (3)
2	Non-Experiment	Nature of Experiment (1.4), Execution: .Rmd Version (0.1)
3	Coding/Editing/Cleaning	.Rmd Version (0.1)
4	Time-Series-Analysis, Application Conditions/ Schema	.Rmd Version (0.1)
5	Model Approximation	Numerical Analysis (5)

TABLE 1.1: Research Steps
(Own depiction based on Schnell, Hill, and Esser, 1999, p. 8; see Geddes, 2003, p. 43)

to present a highly technical topic in a more easily understood manner, a pseudo-experimental set-up is employed: GER DSs would be like the test group (=experimental group) with the stimulus: the IV: The Count of GLs. The FRA DSs would take the place of the control-group without the stimulus. This gives rise to the question: Is the stimulus causing an effect in the experimental group? In the attempt to find an answer to that question, the approach chosen is to draw samples from the target population of all the DSs produced by Monitoring-Stations in GER and FRA. PM10 Data is drawn from every "Bundesland" in GER and every "Region" in FRA. The sampling is best described as stratified sampling, as the single DS is one observation from a certain level of the target population (e.g., level: regional/national/European). The representativeness of this one DS of one "Region"/"Bundesland" for the other DSs of the same "Region"/"Bundesland" is estimated as highly probable for two reasons:

1. In some cases more than one DS from multiple Monitoring-Stations of one "Region"/ "Bundesland" were evaluated, all proved to be consistent

2. It is highly improbable that the regional authority doesn't use the same data generating processes for all its municipalities

Chapter 2

Operationalization

2.1 Chapter Objective

The objective of this chapter is to build and apply a Sufficient Condition as part of a data consistency schema, which makes the H_0 falsifiable by taking the count of GLs into account while comparing the DSs of the respective GLs. However, before the actual operationalization is executed, some expressions have to be defined. In order to give a functional explanation, first the process is explained, where the expressions are embedded without definitions. Secondly, the expressions themselves are explained.

2.2 Process

An excellent way to understand the process is to visualize cars driving by a Monitoring-Station. The Monitoring-Station continually monitors motor vehicle exhaust. One type of fuel emission pollutants are particles, which have "an aerodynamic diameter less than or equal to a nominal 10 micrometer" (cf. *PM10* 2020). This particles are called PM10. The reason for keeping track of those particles is because "Their small size allows them to make their way to the air passages deep within the lungs where they may be deposited and result in adverse health effects"(cf. *PM10* 2020). Having this visual explanation in mind, we can start to describe the process in more abstract terms, and standardize the process in order to make it measurable: the Monitoring-Station (x_{ds}) measures PM10 data to the time (x_{ts}), then the Data-Generating-Unit brings the PM10 data in the state of statistical aggregation (x_{tint}). Afterwards the data is shared across the count of GL (n_G).

1. the Monitoring-Station (x_{ds})
2. to the time(x_{ts})
3. in the state of statistical aggregation (x_{tint})
4. across the count of GL (n_G)

2.2.1 Governance Level

A functional definition is applied here too: the only important aspect of GL in this work, is, that they have either a count of two for FRA or a count of three for GER. All of the rest of the process will be explained in the chapter Data-Pulling-Process (3).

2.2.2 Monitoring-Station

A Monitoring-Station (henceforth, MSt) can be:

1. (mechanical) MSt
2. political Level / A Unit in the form of a (human) work unit, which generates/aggregates/ manipulates the data

This approach is chosen because inconsistent data can be produced in both Data-Generating-Processes. In the case of the (mechanical) MSt this can happen if the MSt is not functioning correctly. Concerning the (human) work unit, this can arise because of human error.

2.3 Data-Consistency-Schema

N^O	Definition	Abbreviation
1	M = Quantity of measured PM10, PM10 = Element of rational Numbers	$M \in \{Q\}$
2	The MSt has produced a DS which, is freely downloadable	$x_{ds} \in \{0;1\}$
3	Either the time-frames of the DSs overlap or not	$x_{ts} \in \{0;1\}$
4	Time = (non-gliding) average per hour (1) and hourly data gliding / moving average (henceforth, described as MA) across 24 hours (2)	$x_{tint} \in \{1;2\}$
5	GL: 2 in FRA, 3 in GER (is determined by the state-structure of the member state)	$n_G \in \{1:3\}$

TABLE 2.1: Data-Consistency-Schema

With these definitions, the necessary and sufficient condition for the DC-Check can be formulated

1. Necessary Condition
   $x_{ds}x_{ts}(x_{tint} + n_G) = y$

$y_{GER} \in \{0;4;5\}$:
0 : Necessary condition not fulfilled
4: Necessary condition fulfilled with MA non-gliding
5: Necessary condition fulfilled with MA24h gliding

$y_{FRA} \in \{0;3;4\}$:
0 : Necessary condition not fulfilled
3: Necessary condition fulfilled with MA non-gliding
4: Necessary condition fulfilled with MA24h gliding

2. Sufficient Condition
   $H_0 = true => M_{EU} = M_{NAT} = M_{REG} =>$

$$y = \frac{1}{\frac{1}{n}\sum_{i=0}^{n-1} x(t-i)_{(g)}} * \frac{1}{n_G}\left(\sum_{j=0}^{n_G}\left(\frac{1}{n}\sum_{i=0}^{n-1} x(t-i)\right)_{(g-j)}\right)$$

(2.1)[1]

where

$y = DV = \{1 + / - Variance\}$:
In how far is this one DS, which I've just downloaded, in coherence with the other DSs from the other GL, regarding the same MSt, the same Time-Period/Interval?

$x = IV$:
Mean (Same as average, but henceforth, called mean) or MA (Depending on if it is GER or FRA DS), of this one DS, which I've just downloaded.

n = Nu*mber of time periods (e.g. 24 hours)

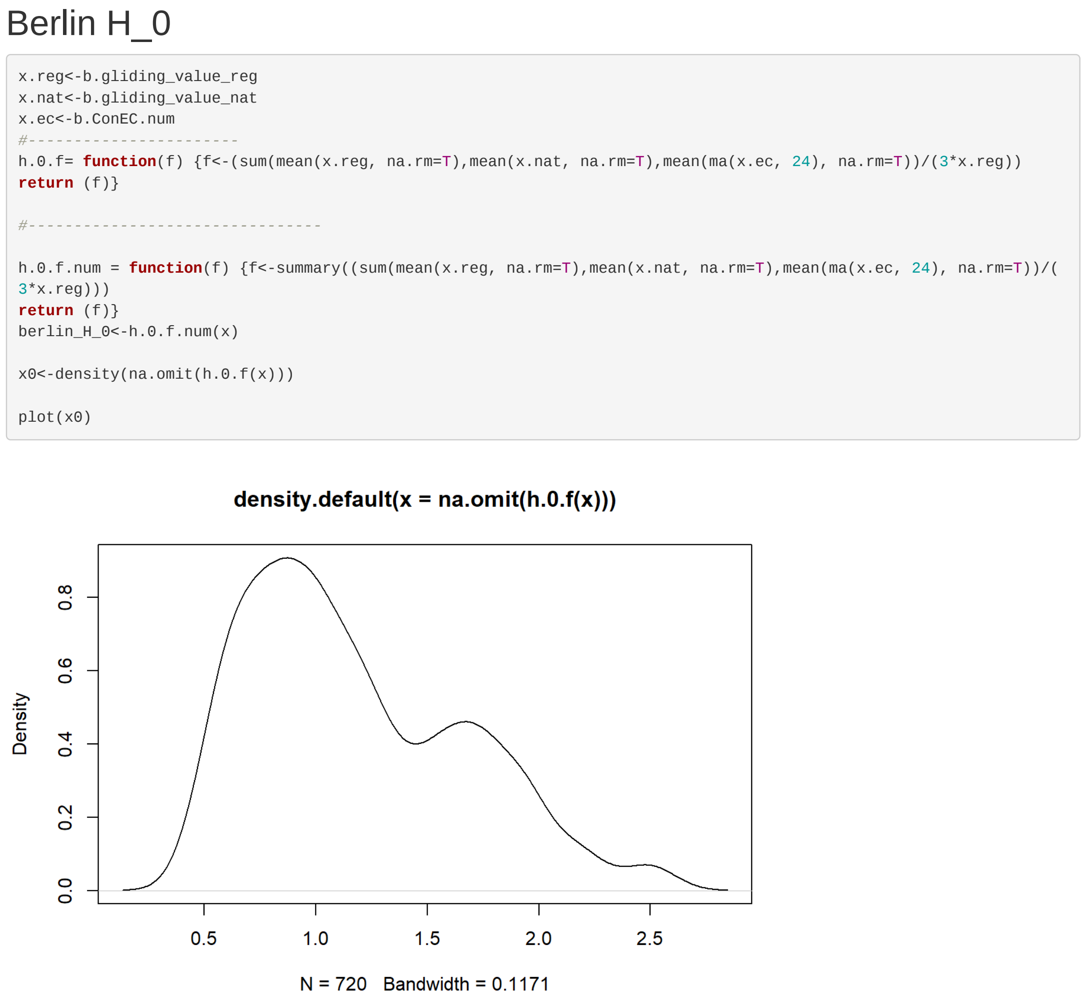
n_G = Number of GL

$t - i$ = Index number (24-0, 24-1 ... 24-23)

$g - j$ = Index number (3 GL-0, 3 GL-1, 3 GL-2)

In other words, dependant upon my selection of DSs (either 2 FRA DSs or 3 GER DSs), I sum up the MAs of the respective GLs and divide them by the product of the count of GLs with one of the MAs, so that if the MAs are exactly the same, the solution has to be 1 (one). See Figure 2.1 for example data from Berlin with the function translated to R (statistical programming language (R Core Team, 2019))

[1] This function can be easily refined in regard to the discrimination power by adding the square or logarithm to the nominator and the denominator.



Berlin H_0

```
x.reg<-b.gliding_value_reg
x.nat<-b.gliding_value_nat
x.ec<-b.ConEC.num
#----------------------
h.0.f= function(f) {f<-(sum(mean(x.reg, na.rm=T),mean(x.nat, na.rm=T),mean(ma(x.ec, 24), na.rm=T))/(3*x.reg))
return (f)}

#-------------------------------

h.0.f.num = function(f) {f<-summary((sum(mean(x.reg, na.rm=T),mean(x.nat, na.rm=T),mean(ma(x.ec, 24), na.rm=T))/(
3*x.reg)))
return (f)}
berlin_H_0<-h.0.f.num(x)

x0<-density(na.omit(h.0.f(x)))

plot(x0)
```

density.default(x = na.omit(h.0.f(x)))
Density
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
N = 720 Bandwidth = 0.1171

FIGURE 2.1: Density Berlin

What can be noted here is that the maximum likelihood estimator is very close to 1, but also that there are a few values, which are greater than 1, this shouldn't be the case. The reason for this is that the reference value reg (the dividend) is smaller than the values of the other two GLs

2.4 Integration of Data-Consistency-Schema

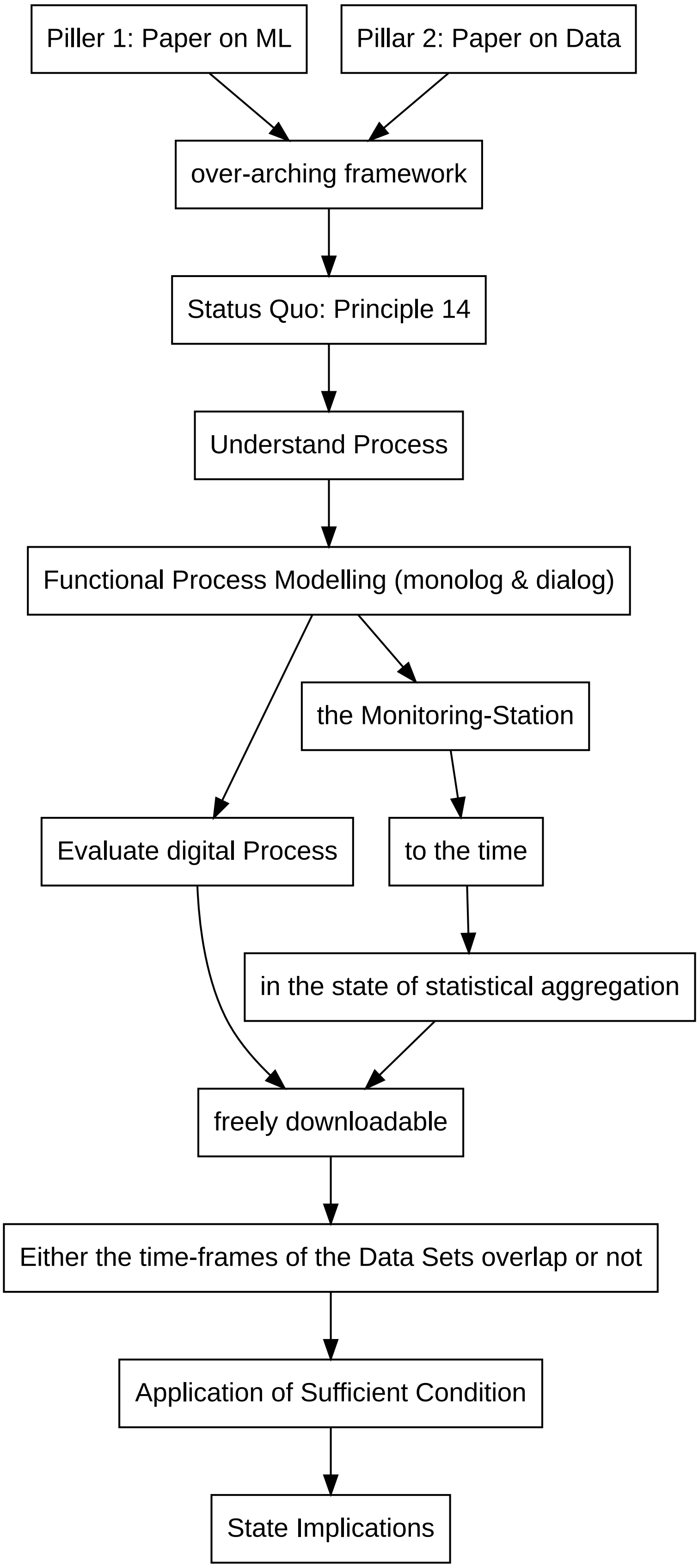


FIGURE 2.2

Chapter 3

Data-Pulling-Process

3.1 Chapter Objective

The objective of this chapter is to explain the data-procurement process. First, a general description of the procured data will be given (ex-post perspective), and then the process of how this outcome was achieved will be described in concrete terms.

3.2 Data sets Comparison

(a) Target population: All (past (Start date oriented at UBA: 2016) and present (time-frame of 4 years)) PM-10 DSs from GER and FRA published on the websites of one of the respective authorities (reg, nat, eu)

(b) N: 21 DC-Checks with 52 DSs

(c) Sample size:

[1] Group GER: 10 times by 3 DSs makes a total of 30 DS

[2] Group FRA: 11 times by 2 DSs, makes a total of 22 DSs (5.2)

(d) Sampling method:

stratified sample, where the strata would be the GL (3), see following Section.

Explanation:
The target population is the population of interest. The samples from FRA and GER are analysed in order to make an inference concerning the target population (see Inference Part 5.3). In total 21 DC-Checks were conducted, which makes a total of 52 DSs. The number of 52 DSs is caused by the different count of GL, which are for GER: 10 times by 3 DSs (from every GL one DS) makes a total of 30 DS, and for FRA: 11 times by 2 DSs (from every GL one DS), makes a total of 22 DSs (5.2). The sampling method is illustrated in the section 3.3.

3.3 Stratified Sampling

The Table 3.1 depicts the strata from which the data is procured. In the context of this work the table shows, furthermore, that the GL take the place of the Strata. The process of procuring the DSs from the Strata/ GL is explained in the following for each of the two Member States.

N^O	GL FRA	Air-Authority FRA	GL GER	Air-Authority GER
1	Departement	Non	Regierungsbezirk	Non
2	Region	Centrally organised: Atmo ,13 different (e.g.Lyon)	Bundesland	16 different (e.g. Bavaria (All Cities))
3	Republique de France	LCSQA	Bundes-Republik	UBA
4	eu	EEA	eu	EEA

TABLE 3.1: Stratified Sampling

(Own, functional depiction based on: La Constitution, éd. Points, 9e éd. (2009), Artikel 20 Absatz 1 GG)

3.3.1 FRA

In the case of FRA, an inner state comparison between the GLs makes no sense as the measurement is in the hand of the AASQA . The national agency LCSQA refers to the data of the AASQA: "La surveillance de la qualité de l'air ambiant en France est confiée aux 18 Associations Agréées de Surveillance de la Qualité de l'Air (AASQA)"*LCSQA* 2020. A translation of this signifies that the control of the data quality itself is left to the association of regional authorities.

Concrete Data pulling process FRA:

(a) From Atmo the names of the air-quality-authorities of the respective regions is procured

(b) Then the DS are downloaded from the homepages of the regional air-quality-authorities

(c) The final step consists out of the retrieving of the corresponding DS from the EEA

3.3.2 GER

In any "Bundesland" there is one environmental agency, that collects the data from the MSt.

Concrete Data pulling process GER:

(a) Download from the homepage of the respective "Bundesland" the DS (e.g. Bavaria (All Cities))

(b) Retrieve the corresponding DS from the UBA (nat authority)

(c) Attain the corresponding DS from the EEA

Chapter 4

Single Case Analysis

4.1 Chapter Objective

This chapter highlights a recurring problem that occurs when the data is analysed. Those readers who do not have exclusive (in the frame of a job) access to the data will eventually run into the same problem if applying a comparable analysis-schema as the process stated below.

4.2 Applying the Data-Consistency-Schema

The way the respective GL publish their data are followings:

N^O	GL/ Authority	State of statistical aggregation
1	Regional	the rounded non-gliding-average per hour
2	National (Only GER)	rounded 24h MA
3	Europe	not rounded, non-gliding Values

TABLE 4.1: Single Case

An example for the statistical aggregation state of the nat values in GER :

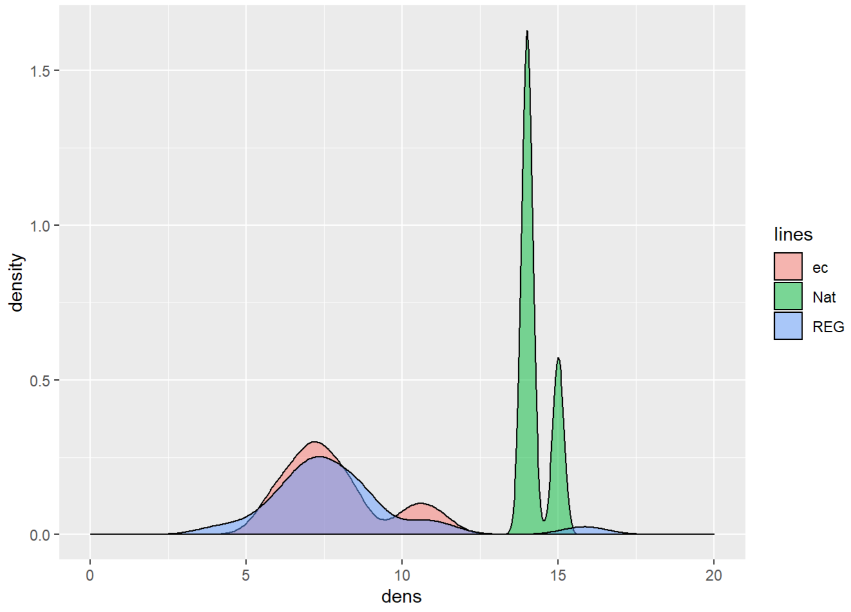
$$m^{(24)}_{MA} = \frac{1}{24}\sum\chi(24-0)+\chi(24-1)+\chi(24+2)...\chi(24-23) \quad (4.1)$$

This has important implications (4.2.1) for this work, as discussed in the next step.

4.2.1 Implications

On some homepages (e.g. Saarland , see "tab_200412_concat.3" to get all cases) of reg authorities in GER there is only the possibility to download the data for the preceding 24h. The problem here is that if these values are non-gliding values (meaning the mean per hour), then the comparison to the nat GL isn't possible. This is caused by the fact that all the PM10 Data, which can be procured from the GER nat GL is 24h non-centred gliding average(s). The problem lies within the time-frames: The person x who downloads the PM10 Data from the reg authority can take only the 24h MA over this time-frame but the nat authority takes it constantly, so that the researcher has to begin with the start value of the reg GL whereas the nat authority has a head-start of 24h. This causes the weighting of the values to be different. This means also that the distribution of the values are different, so that the comparison of densities doesn't help (4.1) to evaluate if the data has is the same/ has the same distributions. In such cases the Modus ponens (4.2) is applied:

Marburg Time-frame problematic



Marburg Modus Ponens applied

```
t.test.EC.NAT.mar<-t.test(PM10.nat.mar, ma24.PM10.ec.mar,paired=F,var.equal=T)
t.test.EC.NAT.mar
```

```
##
##  Two Sample t-test
##
## data:  PM10.nat.mar and ma24.PM10.ec.mar
## t = 0.069824, df = 17413, p-value = 0.9443
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2560482  0.2749643
## sample estimates:
## mean of x mean of y
##  17.57004  17.56058
```

FIGURE 4.1: Marburg Densities

The Table 4.2 shows the logical argumentation that if the 24h reg Values are equal to the 24h eu Values, and if these 24h eu Values are compared as part of a longer time-frame with the 24h gliding nat Values and these two longer time-frames proved to be the same (proved in the Figure 4.1 by the T-Test),

Modus ponens
reg (24h) = eu (24h) eu (24h) $\cap eu(xh) = nat(xh)$
reg (24h)$\cap nat(xh)$

TABLE 4.2: Modus ponens

then Conclusio: 24h reg Values are part of nat Values. But then, of course, the reg GL is under-represented. A re-weighting approach is here avoided by just applying the Sufficient Condition to (only) the nat and eu GL.

Chapter 5

Member State Comparison

5.1 Chapter Objective

This chapter is best described as statistical-analysis-chapter. Up till this point, this work had to establish an understanding of why and how the work was executed. Now, it is time to build on these pillars in order to answer the research question.

5.2 Descriptive Part

(a) General description
The intention behind making a table with the summaries (NA's, mean, median etc.) is to have a broad range of indicators to see irregularities. Concerning the NAs: They can't be valid as they were excluded from the function. Furthermore, negative values are illogical in this context (PM10-Data can't be negative): The most probable scenario is that a missing value is coded -999 or is coded in similar fashion. And, if this were a work about the overall data validity, this would be taken into consideration. However, since this work is concerned with the DC, such values are only considered in order to avoid that the mean is dragged towards an outlier. For such cases, the direct comparison with the median is useful. The quantiles are also a great help to see between which values the function for the H_0 validation (Sufficient Condition) move. Finally, those means, which had the values equal to "Infs"(e.g. $1e^{-6}$ (Infinitely small decimal number)) with the median close to 1 (one) were recoded as 1 (one).

(b) Case descriptions

[1] Case FRA: Out of a total of 18 Regions, 12 have fulfilled the Necessary Condition. For more information, which "Regions" dropped out, please consider "Drop_Out_Table_200411.xlsx" (see "READ_FIRST_200414" in the zip-folder). From this 12 cases, the Region Bourgogne-Franche-Comté, Departement Doubs was doubled. The reason for this doubling is stated in Section 1.4. In order not to weight one region more than other regions, from the 12 cases one of each pair is excluded. This leaves 11 cases for the Sufficient Condition. But as in GER there are only 9 cases for the sufficient condition, two further cases are excluded for the model to avoid an imbalance in the coefficient (see 5.3).

[2] Case GER: Out of a total of 16 "Bundesländer" (so called "Stadtstaaten" included), 11 passed the Necessary Condition. Using the same

approach as described above: In order not to weight one "Bundesland" more than other "Bundesland", two of the three cases from the "Bundesland: Nordrhein Westfalen" were cut out, leaving 9 in total for the Sufficient Condition.

5.2.1 Threshold Values

A threshold value of 1.0 is set for the Sufficient Condition. This isn't only justified by the definition of the Sufficient Condition (2.1) but also by the comparison of unrelated/ incoherent DSs:

```
> x.reg<-PM10.REG.DENW101.Dortmund.Steinstrasse
> x.nat<-PM10.NAT.DENW101.Dortmund.Steinstrasse
> x.ec<-PM10.EC.DENW008.Dortmund.Eving
> h.0.f.num = function(f)
{f<-summary((sum(mean(x.reg, na.rm=T),
mean(x.nat, na.rm=T),
mean(ma(x.ec, 24), na.rm=T))/(3*x.nat)))
+ return (f)}
> h.0.f.num(x)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2731  0.7623  1.0164  1.1613  1.4073  9.1476
```

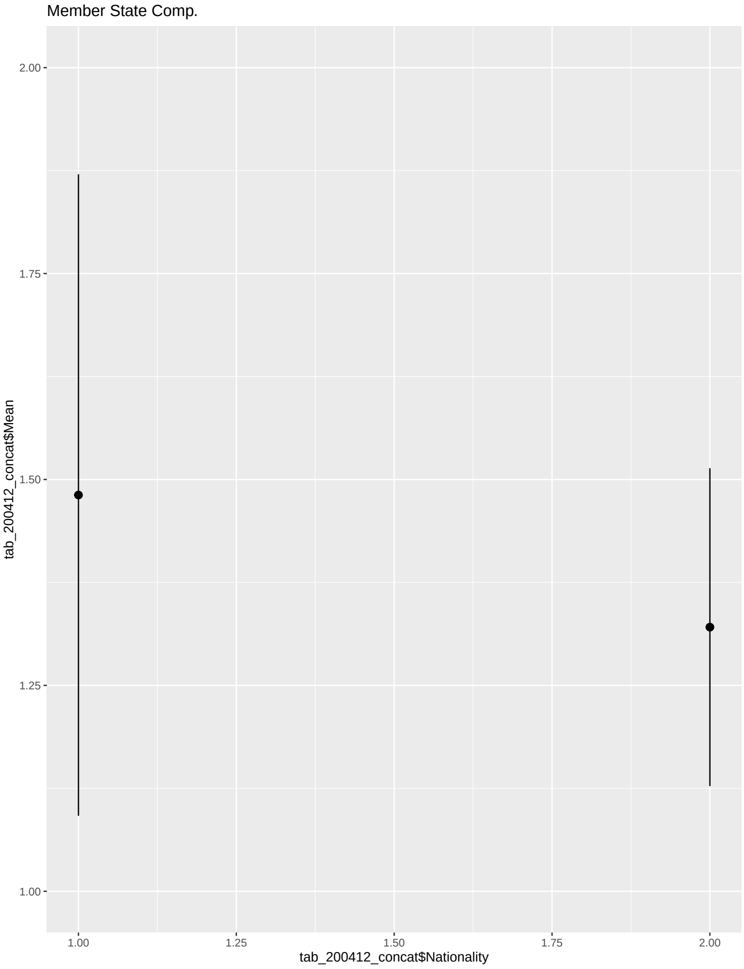


FIGURE 5.1: Member State Comp.

Following cases have fulfilled the threshold-value of 1.0:

(a) FRA Cases: 4

(b) GER Cases: 1

If we ignore the hard threshold-value of 1.0, the following can be procured from the plot 5.1: the FRA Values (see 1.0 x-Axis) scatter more than the GER Values (see 2.0 x-Axis). Furthermore, the mean of the FRA Values lies at 1.436, whereas the mean of the GER Values is 1.320. This finding is meaningful for the question whether a model could be needed: There is variance in the DC in both Member States. The next question to arise is if this variance can be explained by the suggested explanatory variable of the Count of GL.

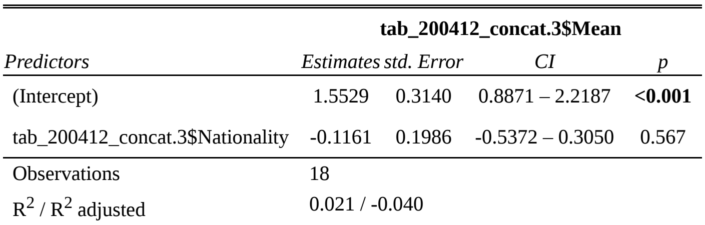
5.3 Inference Part

5.3.1 Model

The model applied here is an additive one, with only one IV, the Count of GL (cf. hypothetical independent variable). This qualifies the model as a bivariate analysis. The same number of FRA Values (originally 11) and GER Values (originally 9) is established by excluding two of the FRA Values in alphabetical order .

$$y = \beta_1 x_{GER/FRA} \quad (5.1)$$

The application of a general linear model to a bivariate question (DV: Data-Consistency, IV: Count of GL) might be somewhat irritating. Normally, a Pearson-, Spearman-, Kendall-correlation is applied to binary questions. The result would of a Pearson correlation is -0.144, which confirms, that there is only a weak negative correlation between the variable count of GL and the DC. However, a general linear model is used here for reasons of convenience. (5.2).



	tab_200412_concat.3$Mean			
Predictors	*Estimates*	*std. Error*	*CI*	*p*
(Intercept)	1.5529	0.3140	0.8871 – 2.2187	**<0.001**
tab_200412_concat.3$Nationality	-0.1161	0.1986	-0.5372 – 0.3050	0.567
Observations	18			
R^2 / R^2 adjusted	0.021 / -0.040			

FIGURE 5.2: Model

For the work at hand, the most important aspect of the general linear model summary is the estimate for the coefficient: "tab_200412_concat.3$Nationality". The interpretation of the coefficient "Nationality" is that the y-axis constant (= intercept) lies at 1.5529 whereas the coefficient is -0.1161. This means that an additional GL or the change from FRA to GER causes the mean to fall by -0.1161. This implies that the GER values are closer to fulfilling the Sufficient Condition than the FRA values. However, this analysis has to be seen relative to its statistical significance, which lies at 0.567, hence, about 10 times over the $p<0.05$ threshold value to consider the H_0 as endangered. The F-Statistic (F-statistic: 0.3417 on 1 and 16 DF) makes it even more clear that only two from a total of 18 DC-Checks (Total of 46 DS) can be predicted by the IV. Furthermore, the H_1 would have to be reformulated so that it wouldn't say that a higher count of GL causes a degeneration of the DC but rather an amelioration (for further information as well as a model based on the binomial distribution, please open the .Rmd file).

Chapter 6

Findings

6.1 Chapter Objective

This chapter is the summation of the work. The main findings and limitations of the work are presented.

6.2 Main Findings

(a) Concerning the H_0: Based on the results found, the null hypothesis of the thesis can't be rejected. This means that the stimulus of the count of GL doesn't cause a degeneration of the DC in the case of GER with the control group of FRA.

(b) In the following the conclusions in regard to Principle 14: a comparability is certainly possible between the different "files". Concerning the "joint use": this whole project was based on making joint use from different providers, as the different GL are different providers. However, in the 52 DSs analysed, a lot of inconsistencies were found which are reflected in the outcomes of the Sufficient Condition. Missed hours (lapses from 01:00 - 03:00 without indicating NA) were more the norm than the exception in case of the reg Values (see .Rmd Version: Case: "ANGOULEME_GAMBETTA_FR09106"). This made the search for the coherent value on the other GL time-consuming. However, it is also explicitly mentioned on some websites that the data comes directly from the MSt and isn't yet validated. Another problem was that some DSs didn't respect the row and columns nomenclature (e.g. DS Toulon). Lastly the analysis of FRA showed that the centralization of provider-competences AASQA makes it easier to find the respective websites of the reg providers. However, because of a temporary assignment of the button function, the download of the DSs couldn't be built into the R code directly. Here, the UBA and the EEA make it easier to build in the DS-downloads directly into the code.

(c) A relevant connection between the "over-arching framework" mentioned in the Section "State of the Art" (see 1.2) and the concrete insights found in this paper is easily made and illustrative. If the R-Code-Script (0.1) of this work is as any indication of how much time and effort is needed to clean the data in order to make basic statistical analysis, there is a problem. The problem consists of what was described in Section 2b. In order to solve this problem, the mentioned "over-arching framework" (*COM(2020) 66 final* 2020, p. 12) might help. The standardization of the DSs would be particularly important to ensure machine-readable formats and the accessibility to the DSs by the front-ends would enable an automatization of the analysis. And, the fact that the "A European strategy for data" (*COM(2020) 66 final* 2020) and the White Paper "On Artificial Intelligence" (*COM(2020) 65 final* 2020) were published on the very same day shows that the European commission has the goal to build the infrastructure for an automatization of the data analysis.

6.3 Limitations

6.3.1 General Evaluation:

Internal Validity The Internal Validity does not exist as there is no traceable causality. This, in turn, is based on the assumption that there are multiple sources of error for inconsistent data (cf. 2.2.2). Regarding the objective (2.1) of the work at hand, the internal validity is a non-suiting indicator to measure this non-experiment.

External Validity As the non-experiment at hand is conducted in the frame of given circumstances (1.4.1), the external validity is high. Meaning that if wanted one could have just proceeded with the DC-Checks until a full census would have been achieved. Hence, the application of habits of thought to fields different from those in which they have been formed" is avoided (cf. Von Hayek, 1989). They are avoided because the developed schema (the Necessary Condition and Sufficient Condition) is tailored for the given circumstances.

Reliability The non-experiment is, in its entirety, reproducible (0.1). This allows for verification of the results. Multiple loops regarding the reviewing of the DS were done, which should further improve the temporal stability of the results.

6.3.2 Concrete weaknesses of the thesis:

(a) The argumentation at the end of the section 1 "Nature of Experiment" can be remedied with the Data-Generating-Process. This means that there can always be mistakes (see 2.2.2) even though the "Bundesländer"/"Regions" apply the same measurement techniques concerning the cities and MSt under their authority. That would mean, in turn, that the sample isn't representative. This could be debated by pointing out that multiple samples from one city and "Bundesländer"/"Regions" were drawn to avoid this, and that all but one (MSt: "Dortmund Steinstrasse") proved to be consistent (see .Rmd Version of this thesis).

(b) Regarding the pseudo-experimental setup, one could argue that an IV of nationality is in this pseudo-experimental setup interchangeable with the suggested IV: Count of GL. Because one cannot distinguish between the category nationality/ different member state (FRA, GER) and Count of GL (2 GL for FRA, 3 GL for GER). The counter-argument focuses on the main result found, namely that there is no significant difference, so neither hypothetical independent variable loads on the IV: DC. Only if the working hypothesis could have been rejected, then the following question would have arisen: Does every member state/nation has a different level of DC or are there two groups, one for federal-states and one for presidential-states.

(c) Another relevant point is mentioned in Section 4.2.1. There are different time interval restrictions on the respective GL websites or values are inexplicitly repeated or left out so DSs had to be cut/shortened down until they could be compared. Comparisons were made from x <= one day - one year (see "READ_FIRST_200414" and "table tab_200412_concat.csv"). In the (smaller) one-day-strata-samples (1.4) outliers were more weighted. However, this again was anticipated for the summary (mean, median, NA's etc.) of the Sufficient Condition was taken for the final comparison between the GL of GER internally and GER vs. FRA. Hence, the outlier could be taken into consideration (5.2) by the median.

6.4 Further Findings

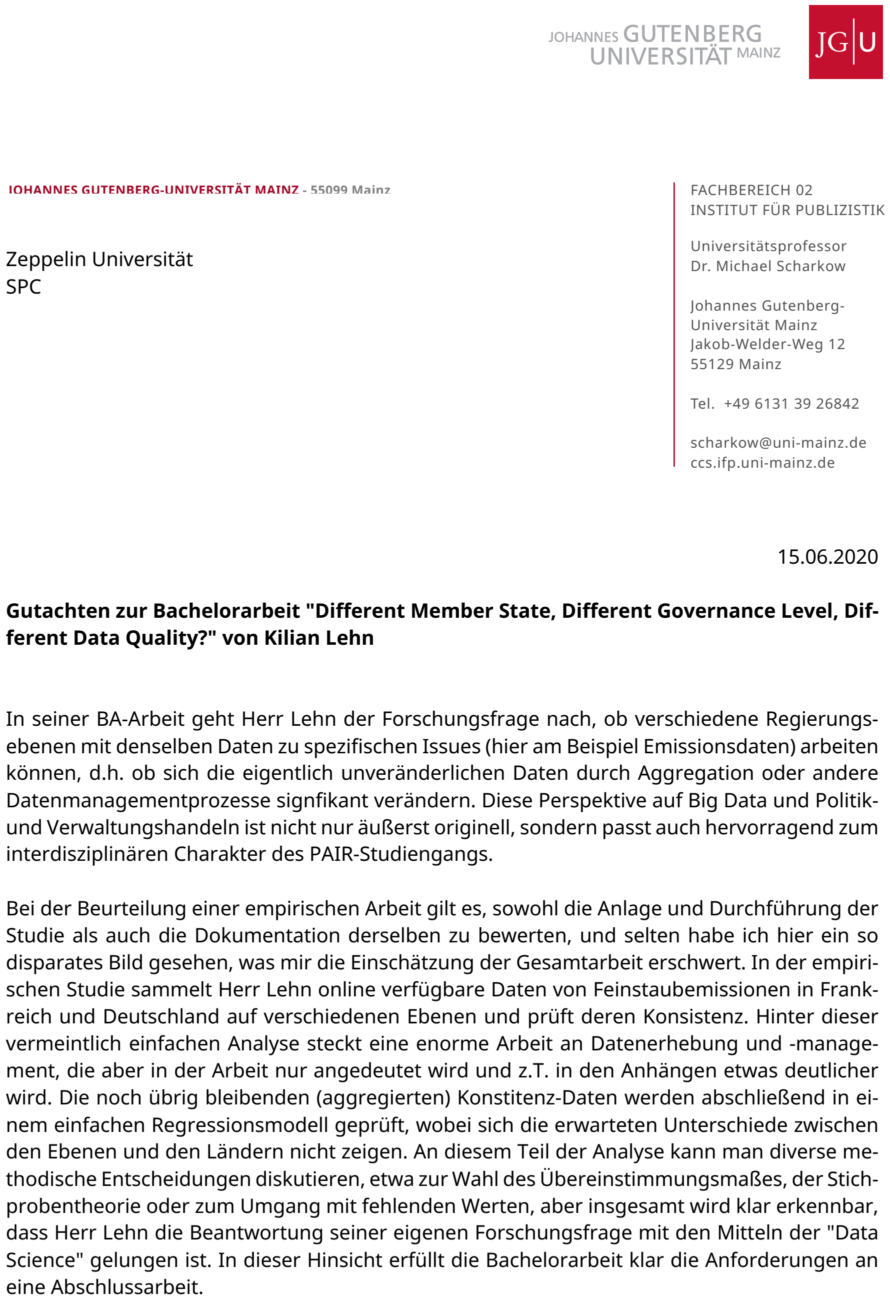
This work started out with the data source: EC Dashboard. However, the data source was later changed to the EEA for the following reasons:

(a) Spatial mapping would have been necessary to compare the EC Dashboard values with the nat and reg values because the EC Dashboard Values are in the Unit: Tons per Year. Whereas the nat and reg values are in the Unit: Concentration μ/m^3.

(b) Only through personal correspondence with the Joint Research Center Ispra, (Directorate B – Growth and Innovation, Territorial Development Unit (B3)*JRC* 2019) was the information obtained that the data displayed on the EC Dashboard website is simulated data generated by the GAINS-Model. This means that a comparison by spatial mapping with the simulated data would have been a review of the fit of the simulated data from the Joint Research Center Ispra to the real data but wouldn't have helped to answer the question about the DC.

6.5 Assessment

In the following, the assessment is attached.



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

JOHANNES GUTENBERG-UNIVERSITÄT MAINZ - 55099 Mainz

Zeppelin Universität
SPC

FACHBEREICH 02
INSTITUT FÜR PUBLIZISTIK

Universitätsprofessor
Dr. Michael Scharkow

Johannes Gutenberg-
Universität Mainz
Jakob-Welder-Weg 12
55129 Mainz

Tel. +49 6131 39 26842

scharkow@uni-mainz.de
ccs.ifp.uni-mainz.de

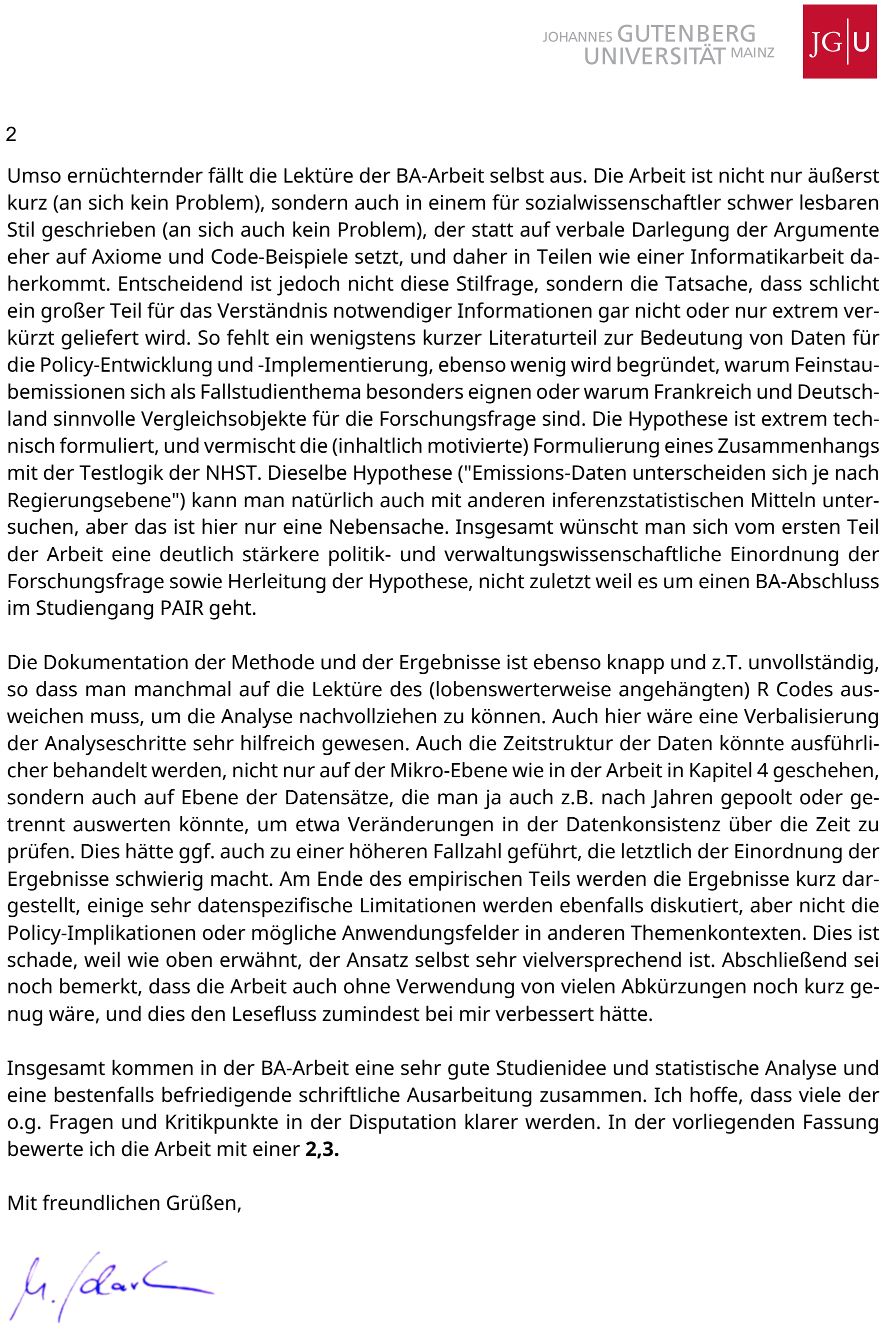
15.06.2020

Gutachten zur Bachelorarbeit "Different Member State, Different Governance Level, Different Data Quality?" von Kilian Lehn

In seiner BA-Arbeit geht Herr Lehn der Forschungsfrage nach, ob verschiedene Regierungsebenen mit denselben Daten zu spezifischen Issues (hier am Beispiel Emissionsdaten) arbeiten können, d.h. ob sich die eigentlich unveränderlichen Daten durch Aggregation oder andere Datenmanagementprozesse signfikant verändern. Diese Perspektive auf Big Data und Politik- und Verwaltungshandeln ist nicht nur äußerst originell, sondern passt auch hervorragend zum interdisziplinären Charakter des PAIR-Studiengangs.

Bei der Beurteilung einer empirischen Arbeit gilt es, sowohl die Anlage und Durchführung der Studie als auch die Dokumentation derselben zu bewerten, und selten habe ich hier ein so disparates Bild gesehen, was mir die Einschätzung der Gesamtarbeit erschwert. In der empirischen Studie sammelt Herr Lehn online verfügbare Daten von Feinstaubemissionen in Frankreich und Deutschland auf verschiedenen Ebenen und prüft deren Konsistenz. Hinter dieser vermeintlich einfachen Analyse steckt eine enorme Arbeit an Datenerhebung und -management, die aber in der Arbeit nur angedeutet wird und z.T. in den Anhängen etwas deutlicher wird. Die noch übrig bleibenden (aggregierten) Konstitenz-Daten werden abschließend in einem einfachen Regressionsmodell geprüft, wobei sich die erwarteten Unterschiede zwischen den Ebenen und den Ländern nicht zeigen. An diesem Teil der Analyse kann man diverse methodische Entscheidungen diskutieren, etwa zur Wahl des Übereinstimmungsmaßes, der Stichprobentheorie oder zum Umgang mit fehlenden Werten, aber insgesamt wird klar erkennbar, dass Herr Lehn die Beantwortung seiner eigenen Forschungsfrage mit den Mitteln der "Data Science" gelungen ist. In dieser Hinsicht erfüllt die Bachelorarbeit klar die Anforderungen an eine Abschlussarbeit.

FIGURE 6.1



JOHANNES GUTENBERG UNIVERSITÄT MAINZ JG|U

2

Umso ernüchternder fällt die Lektüre der BA-Arbeit selbst aus. Die Arbeit ist nicht nur äußerst kurz (an sich kein Problem), sondern auch in einem für sozialwissenschaftler schwer lesbaren Stil geschrieben (an sich auch kein Problem), der statt auf verbale Darlegung der Argumente eher auf Axiome und Code-Beispiele setzt, und daher in Teilen wie einer Informatikarbeit daherkommt. Entscheidend ist jedoch nicht diese Stilfrage, sondern die Tatsache, dass schlicht ein großer Teil für das Verständnis notwendiger Informationen gar nicht oder nur extrem verkürzt geliefert wird. So fehlt ein wenigstens kurzer Literaturteil zur Bedeutung von Daten für die Policy-Entwicklung und -Implementierung, ebenso wenig wird begründet, warum Feinstaubemissionen sich als Fallstudienthema besonders eignen oder warum Frankreich und Deutschland sinnvolle Vergleichsobjekte für die Forschungsfrage sind. Die Hypothese ist extrem technisch formuliert, und vermischt die (inhaltlich motivierte) Formulierung eines Zusammenhangs mit der Testlogik der NHST. Dieselbe Hypothese ("Emissions-Daten unterscheiden sich je nach Regierungsebene") kann man natürlich auch mit anderen inferenzstatistischen Mitteln untersuchen, aber das ist hier nur eine Nebensache. Insgesamt wünscht man sich vom ersten Teil der Arbeit eine deutlich stärkere politik- und verwaltungswissenschaftliche Einordnung der Forschungsfrage sowie Herleitung der Hypothese, nicht zuletzt weil es um einen BA-Abschluss im Studiengang PAIR geht.

Die Dokumentation der Methode und der Ergebnisse ist ebenso knapp und z.T. unvollständig, so dass man manchmal auf die Lektüre des (lobenswerterweise angehängten) R Codes ausweichen muss, um die Analyse nachvollziehen zu können. Auch hier wäre eine Verbalisierung der Analyseschritte sehr hilfreich gewesen. Auch die Zeitstruktur der Daten könnte ausführlicher behandelt werden, nicht nur auf der Mikro-Ebene wie in der Arbeit in Kapitel 4 geschehen, sondern auch auf Ebene der Datensätze, die man ja auch z.B. nach Jahren gepoolt oder getrennt auswerten könnte, um etwa Veränderungen in der Datenkonsistenz über die Zeit zu prüfen. Dies hätte ggf. auch zu einer höheren Fallzahl geführt, die letztlich der Einordnung der Ergebnisse schwierig macht. Am Ende des empirischen Teils werden die Ergebnisse kurz dargestellt, einige sehr datenspezifische Limitationen werden ebenfalls diskutiert, aber nicht die Policy-Implikationen oder mögliche Anwendungsfelder in anderen Themenkontexten. Dies ist schade, weil wie oben erwähnt, der Ansatz selbst sehr vielversprechend ist. Abschließend sei noch bemerkt, dass die Arbeit auch ohne Verwendung von vielen Abkürzungen noch kurz genug wäre, und dies den Lesefluss zumindest bei mir verbessert hätte.

Insgesamt kommen in der BA-Arbeit eine sehr gute Studienidee und statistische Analyse und eine bestenfalls befriedigende schriftliche Ausarbeitung zusammen. Ich hoffe, dass viele der o.g. Fragen und Kritikpunkte in der Disputation klarer werden. In der vorliegenden Fassung bewerte ich die Arbeit mit einer **2,3.**

Mit freundlichen Grüßen,

FIGURE 6.2

Bibliography

Balestracci, Davis (2006). "Data 'sanity': statistics and reality". In: *Quality in Primary Care* 14.1, pp. 49–53.

Banerjee, Abhijit and Esther Duflo (2013). *Poor Economics. A Radical Rethinking of the Way to Fight Global Poverty*.

Bryman, Alan (2008). "Of methods and methodology". In: *Qualitative Research in Organizations and Management: An International Journal*.

Camus, Albert (1957). *The Nobel Prize in Literature 1957*. [Online; accessed 1. Apr. 2020]. URL: https://www.nobelprize.org/prizes/literature/1957/camus/25232-albert-camus-banquet-speech-1957.

Carrete, Lorena et al. (2012). "Green consumer behavior in an emerging economy: confusion, credibility, and compatibility". In: *Journal of consumer marketing*.

Colquitt, Jason A and Cindy P Zapata-Phelan (2007). "Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal". In: *Academy of Management Journal* 50.6, pp. 1281–1303.

COM(2020) 65 final (2020). [Online; accessed 11. May 2020]. URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

COM(2020) 66 final (2020). [Online; accessed 11. May 2020]. URL: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf.

ESS Handbook (2018). [Online; accessed 31. Mar. 2020]. URL: https://ec.europa.eu/eurostat/cros/content/ess-handbook-methodology-data-validation-version-20-revision-2018_en.

Estat:eurostat, corporate-body. (2018). "European statistics code of practice : for the national statistical authorities and Eurostat (EU statistical authority)." In: *Publications Office of the European Union*. URL: https://op.europa.eu/en/publication-detail/-/publication/661dd8ef-7439-11e8-9483-01aa75ed71a1/language-en.

Eurostat (2020). [Online; accessed 31. Mar. 2020]. URL: https://ec.europa.eu/eurostat/web/quality/principle14.

Geddes, Barbara (2003). *Paradigms and sand castles: Theory building and research design in comparative politics*. University of Michigan Press.

Hynes, Nick, D Sculley, and Michael Terry (2017). "The data linter: Lightweight, automated sanity checking for ml data sets". In: *NIPS MLSys Workshop*.

Jacobson, Ivar (1993). *Object-oriented software engineering: a use case driven approach*. Pearson Education India.

JRC (2019). [Online; accessed 16. Apr. 2020]. URL: https://ec.europa.eu/info/sites/info/files/organisation_charts/organisation-chart-jrc_en.pdf.

Kaplan, Robert S and P David (2000). "Norton. 1996". In: *The balanced scorecard: translating strategy into action*.

Kolb, Robert W (2010). *Lessons from the financial crisis: Causes, consequences, and our economic future*. Vol. 12. John Wiley & Sons.

LCSQA (2020). URL: https://www.data.gouv.fr/fr/organizations/laboratoire-central-de-surveillance-de-la-qualite-de-lair/.

Morgan, Stephen L and Christopher Winship (2015). *Counterfactuals and causal inference*. Cambridge University Press.

PM10 (2020). [Online; accessed 29. Mar. 2020]. URL: https://www.eea.europa.eu/themes/air/air-quality/resources/glossary/pm10.

Polyzotis, Neoklis et al. (Dec. 2018). "Data Lifecycle Challenges in Production Machine Learning: A Survey". In: *SIGMOD Rec.* 47.2, 17–28. ISSN: 0163-5808. DOI: 10.1145/3299887.3299891. URL: https://doi.org/10.1145/3299887.3299891.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Radaelli, Claudio M (2017). *Technocracy in the European Union*. Routledge.

Schnell, Rainer, Paul B Hill, and Elke Esser (1999). *Methoden der empirischen Sozialforschung*. R. Oldenbourg Muenchen ua.

Von Hayek, Friedrich August (1989). "The pretence of knowledge". In: *The American Economic Review* 79.6, pp. 3–7.

Weizenbaum, Joseph (1976). *Computer power and human reason: From judgment to calculation.* WH Freeman & Co.