December 2020 (update 2021)

Methodological documentation of the revision of the gap filling exercise

*ETC/ATNI consortium partners:*   
NILU – Norwegian Institute for Air Research, Aether Limited, Czech Hydrometeorological Institute (CHMI), EMISIA SA, Institut National de l’Environnement Industriel et des risques (INERIS), Universitat Autònoma de Barcelona (UAB), Umweltbundesamt GmbH (UBA-V), 4sfera Innova,  
Transport & Mobility Leuven NV (TML)

Authors:

Jaume Fons-Esteve, Francisco Domingues, Maria José Ramos (UAB)

Cover design: XXX

Cover photo © xxxxxxxxxxxx xxxxxxxxxxxxx

Layout: Xxxxxx Xxxxxxx

**Legal notice**

The contents of this publication do not necessarily reflect the official opinions of the European Commission or other institutions of the European Union. Neither the European Environment Agency, the European Topic Centre on Air pollution, transport, noise and industrial pollution nor any person or company acting on behalf of the Agency or the Topic Centre is responsible for the use that may be made of the information contained in this report.

**Copyright notice**

© European Topic Centre on Air pollution, transport, noise and industrial pollution, 2020

Reproduction is authorised, provided the source is acknowledged.

Information about the European Union is available on the Internet. It can be accessed through the Europa server (www.europa.eu).

*The withdrawal of the United Kingdom from the European Union did not affect the production of the report.*

*Data reported by the United Kingdom are included in all analyses and assessments contained herein, unless otherwise indicated.*

**Author(s)**

Jaume Fons Esteve, Francisco Domingues, Maria José Ramos (UAB)

***DISCLAIMER***

This ETC/ATNI Working Paper has not been published and therefore no reference can be made to it in any other formal or published document or document meant for publication.

This document has not been subjected to any review by the European Environment Agency (EEA) or EEA’s member countries. It does not represent the formal views of the EEA and its ETC/ATNI.

European Topic Centre on Air pollution,

transport, noise and industrial pollution

c/o NILU – Norwegian Institute for Air Research

P.O. Box 100, NO-2027 Kjeller, Norway

Tel.: +47 63 89 80 00

Email: [etc.atni@nilu.no](mailto:etc.atni@nilu.no)

Web : <https://www.eionet.europa.eu/etcs/etc-atni>

**Meta information on this EIONET Report**

**Task reference**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AP** | **Task #** | **Task Title** | **ETC Task leader** | **EEA contact Persons** |
| 2020 | 1.1.5.1 | Noise data – operational compilation and management | Núria Blanes | Eulàlia Peris |

**Type, titles and series number of report** *(EIONET report no. and ISBN-no. will be given by NILU)*

|  | **EIONET- ETC/ATNI Working Paper 2020** |
| --- | --- |
| **Title:** | Methodological documentation of the revision of the gap filling exercise |
| **Sub title:** |  |
| **Number: (\*)** |  |

**(\*)** Obtaining a series number: See instructions at the Eionet Forum IG “ETC/ATNI Consortium”:

<https://forum.eionet.europa.eu/etc-atni-consortium>

**Authors**

|  | **Name** | **Institute** | **ORCID-iD (https://orcid.org/)** |
| --- | --- | --- | --- |
| **Lead Author:** | Jaume Fons | UAB |  |
| **Co-authors:** | Francisco Domingues | UAB |  |
|  | Maria José Ramos | UAB |  |
|  |  |  |  |

**Document history**

|  |  |
| --- | --- |
| Lead author to confirm:  I hereby confirm that I have familiarised myself with the EEA Writing Manual  (to be found at: <https://forum.eionet.europa.eu/etc-atni-consortium/library/general-etc-atni>). | YES:  X |

Make sure to deliver completed drafts and versions to Eionet Forum IG “ETC/ATNI Consortium” in its appropriate pre-defined IP- and Task-subfolder at <https://forum.eionet.europa.eu/etc-atni-consortium/library>

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Authors/Reviewers | Date | Comments |
| 1.0 | J. Fons Esteve, F. Domingues, M. J. Ramos | 18.12.2020 | Issue the report |
| 2.0 | J. Fons Esteve, F. Domingues, M. J. Ramos | 30/04/21 | Updated version with additional documentation included. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Be aware that this document history list will be removed from the final version of the report.

**REQUIRED INFORMATION NEEDED FOR THE PUBLICATION :**

**SUMMARY** (used in item listings and search results)**:**

|  |
| --- |
| The experience of the 3 reporting cycles of the END has demonstrated substantial delays on the delivery of the data. For that reason, methodologies to estimate missing data have been developed as early as 2013 to have a more complete overview of the extent of population exposed to various noise sources. This report focusses on specific aspects of the current methodology to estimate missing data. In particular it provides a systematic approach when regression is required (e.g. estimation of population exposed to agglomerations roads based on number of inhabitants), and explore a new approach for major airports based on estimating the noise contour band when this information is not provided. |

**TAGS** (Tags are commonly used for ad-hoc organisation of content)**:**

|  |
| --- |
| Noise exposure data, gap filling methodology |

**ABSTRACT :**

|  |
| --- |
| The experience of the 3 reporting cycles of the END has demonstrated substantial delays on the delivery of the data. For that reason, methodologies to estimate missing data have been developed as early as 2013 to have a more complete overview of the extent of population exposed to various noise sources. This report focusses on specific aspects of the current methodology to estimate missing data. In particular it provides a systematic approach when regression is required (e.g. estimation of population exposed to agglomerations roads based on number of inhabitants), and explore a new approach for major airports based on estimating the noise contour band when this information is not provided. |

Contents

[Summary 6](#_Toc64915422)

[Acknowledgements 7](#_Toc64915423)

[1 Objectives 8](#_Toc64915424)

[2 Input data 8](#_Toc64915425)

[3 END agglomerations data: gap filling 10](#_Toc64915426)

[3.1 Gap filling method for agglomerations- road 10](#_Toc64915427)

[3.1.1 Overview 10](#_Toc64915428)

[3.1.2 Regression 14](#_Toc64915429)

[3.1.3 Identify outliers and adjust agglomerations with > 100% people exposed 14](#_Toc64915430)

[3.1.4 Visual inspection of the relationship between people exposed and number of inhabitants 15](#_Toc64915431)

[3.1.5 Data subsetting for regression analysis and validation 17](#_Toc64915432)

[3.1.6 Test models 18](#_Toc64915433)

[3.1.7 Validation 20](#_Toc64915434)

[3.1.8 Estimate values for missing data 21](#_Toc64915435)

[3.1.9 Calculate the total people exposed in Europe 21](#_Toc64915436)

[3.1.10 Estimate the people distributed by noise bands 21](#_Toc64915437)

[3.2 Gap filling method for railway noise, aircraft noise and industrial noise inside agglomerations 21](#_Toc64915438)

[4 END major roads and major railways exposure data outside agglomerations: gap filling 26](#_Toc64915439)

[4.1 Gap filling method 26](#_Toc64915440)

[5 END Major airport exposure outside agglomerations: gap filling 30](#_Toc64915441)

[5.1 Overview 30](#_Toc64915442)

[5.2 A methodology based on estimating the noise contour around runaways 33](#_Toc64915443)

[5.2.1 Overview 33](#_Toc64915444)

[5.2.2 Data requirements 34](#_Toc64915445)

[5.2.3 Delineation of the runaways and noise contours 34](#_Toc64915446)

[5.2.4 Detailed description of the methodology 36](#_Toc64915447)

[5.2.5 Outcome of the test 37](#_Toc64915448)

[6 Conclusions 40](#_Toc64915449)

[References 41](#_Toc64915450)

[Annex I. Summary of statistics 42](#_Toc64915451)

[Annex II. R script 43](#_Toc64915452)

[Annex III. Reported and delineated contour for Lden 55 dB (major airports) 56](#_Toc64915453)

Summary

The experience of the 3 reporting cycles of the END has demonstrated substantial delays in delivering the data. For that reason, methodologies to estimate missing data have been developed as early as 2013 to have a complete overview of the extent of the population exposed to various noise sources. This report focusses on specific aspects of the current methodology to estimate missing data. In particular, it provides a systematic approach when regression is required (e.g. estimation of the population exposed to agglomerations roads based on the number of inhabitants), and explore a new method for major airports based on estimating the noise contour band when this information is not provided.

Acknowledgements

Eulàlia Peris, the EEA project manager, supported the work and provided valuable comments on the report.

# Objectives

The experience of the 3 reporting cycles of the END (2007, 2012, 2017) has demonstrated substantial delays on the reporting of MS, consequence of a learning process dealing with the complexity of the reporting and the END requirements. Because that, since the starting of the reporting of the END the option to estimate missing data has been considered. Such estimation allows to have a more realistic and complete picture of people potentially exposed to different noise sources.

The latest update of the methodology was reported by Ramos (2019), where the need for some improvements were identified -and out of the scope of the work at that time. The current report tackles the following issues in order to improve the methodology and systematise the estimation of the missing data:

* Provide a systematic approach to test alternative regression models when it is pertinent (e.g. population exposed to roads inside agglomerations).
* Specify the methodology for propagation of errors which will lead to provide figures of total people exposed with the corresponding confidence interval.
* Improvements on the estimation of people exposed to noise from major airports based on the estimation of the noise contour for Lden 55 dB.

This methodological report summarises the steps followed to obtain estimated results of a complete noise exposure covering the END sources. Steps are exemplified with selected data.

The process followed is described in the next sections:

* Chapter 2: Input data
* Chapter 3: END agglomerations – road noise exposure data: gap filling exercise
* Chapter 4: END major roads and major railways noise exposure data
* Chapter 5: END major airports, test of the new approach

# Input data

This report is based on the review of the methodology developed in 2019 (Ramos, 2019). Therefore, the same data source have been used, for comparability reasons, when data was required to test some improvements. The data covers the data reported until 01/01/2019[[1]](#footnote-1).

In order to facilitate the implementation of the workflow for estimating missing data, the minimum data requirements are specified in Table 1, Table 2, and Table 3 for the different noise sources. Therefore, a preliminary step is to extract the needed information from the internal noise database(s) which compile the reported data. The most critical part is the identification of completeness for major roads and major rails. The objective of this report is out of the scope of the preliminary data selection -details are available at Ramos (2019). Moreover, changes on the data model implemented in Reportnet 3.0 may facilitate this step on selecting the needed data. For that reason, Table 1, Table 2, and Table 3 are provided as reference for futures use.

It should be noted that the methodology also requires data from the previous reporting cycle. The notation used in this report, to generalise the methodology, is detailed as follows:

* t2, data from the current reporting cycle
* t1, data from the previous reporting cycle

Table 1. Structure of the input data for people exposed to noise from roads inside agglomerations.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type of data** | **Comment** |
| Country | String |  |
| Country2 | String |  |
| Agglomeration\_name | String |  |
| RLID | String |  |
| UniqueAgg\_ID | String |  |
| Year | Integer | Reference year of the reported data |
| Population | Integer |  |
| Lden per noise band | Integer | -1 not applicable, -2 not available (reported), -9999 not reported |
| Lnight per noise band | Integer | -1 not applicable, -2 not available (reported), -9999 not reported |

Table 2. Structure of the input data for people exposed to noise from major airports.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type of data** | **Comment** |
| Country | String |  |
| Mair\_name | String |  |
| ICAO\_code | String |  |
| Year | Integer | Reference year of the reported data |
| Lden per noise band | Integer | -1 not applicable, -2 not available (reported), -9999 not reported |
| Lnight per noise band | Integer | -1 not applicable, -2 not available (reported), -9999 not reported |

Table 3. Structure of the input data for people exposed to noise from major roads and major rails.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type of data** | **Comment** |
| Country | String |  |
| Country2 | String |  |
| Unique identifier(s) | *String (Several fields)* | All needed fields that provide the unique link to the noise database |
| Completeness | String | Complete, partial, not applicable, not provided |
| Year | Integer | Reference year of the reported data |
| Lden per noise band | Integer |  |
| Lnight per noise band | Integer |  |

# END agglomerations data: gap filling

## Gap filling method for agglomerations- road

### Overview

This method is based on the working paper “Establishing a methodological proposal to interpolate a complete coverage on noise exposure at theEU level” (ETC/ACM, 2015), with some improvements described by Ramos (2019).

Figure 1 provides an overview of the process for estimating missing data adapted from Ramos (2019). In each step, the decision tree uses the best approach according to the available ancillary data:

* Use data from the previous reporting period if available.
* When this information is also missing, the following steps are taken:
  + People exposed is estimated with regression analysis, being the number of inhabitants the independent variable.
  + Finally, estimate the distribution of the population exposed by noise bands applying the European average.

A more detailed explanation and the basis for the methodology applied in each step is provided in Table 5.

The gap filling methodology also includes those few cases where data is missing for some noise bands. Then, **partial gap filling** is conducted as follows:

* Use data from the previous reporting period for the missing noise band(s).
* If data from the previous reporting period is not available use the European average of % of the population exposed by noise band to estimate the missing data. Table 4 provides an example of a partial gap filling when data from the previous period is not available.

The following sections focus on those aspects that were not documented on previous reports:

* Selection of the appropriate regression model
* Estimation of the missing data with the corresponding error (error of the prediction).
* Calculation of the people exposed in Europe as sum of all the agglomerations (reported and gap filled). Calculation of the propagation of error as a result of adding individual values with their own error (error of the prediction).
* Estimate the distribution of total noise exposure by noise bands and associated error -error of the function’s sum and error to distribute the total by noise bands.

All these steps have been implemented in R to ensure the traceability and replicability of the whole process, including the assessment of different regression models.

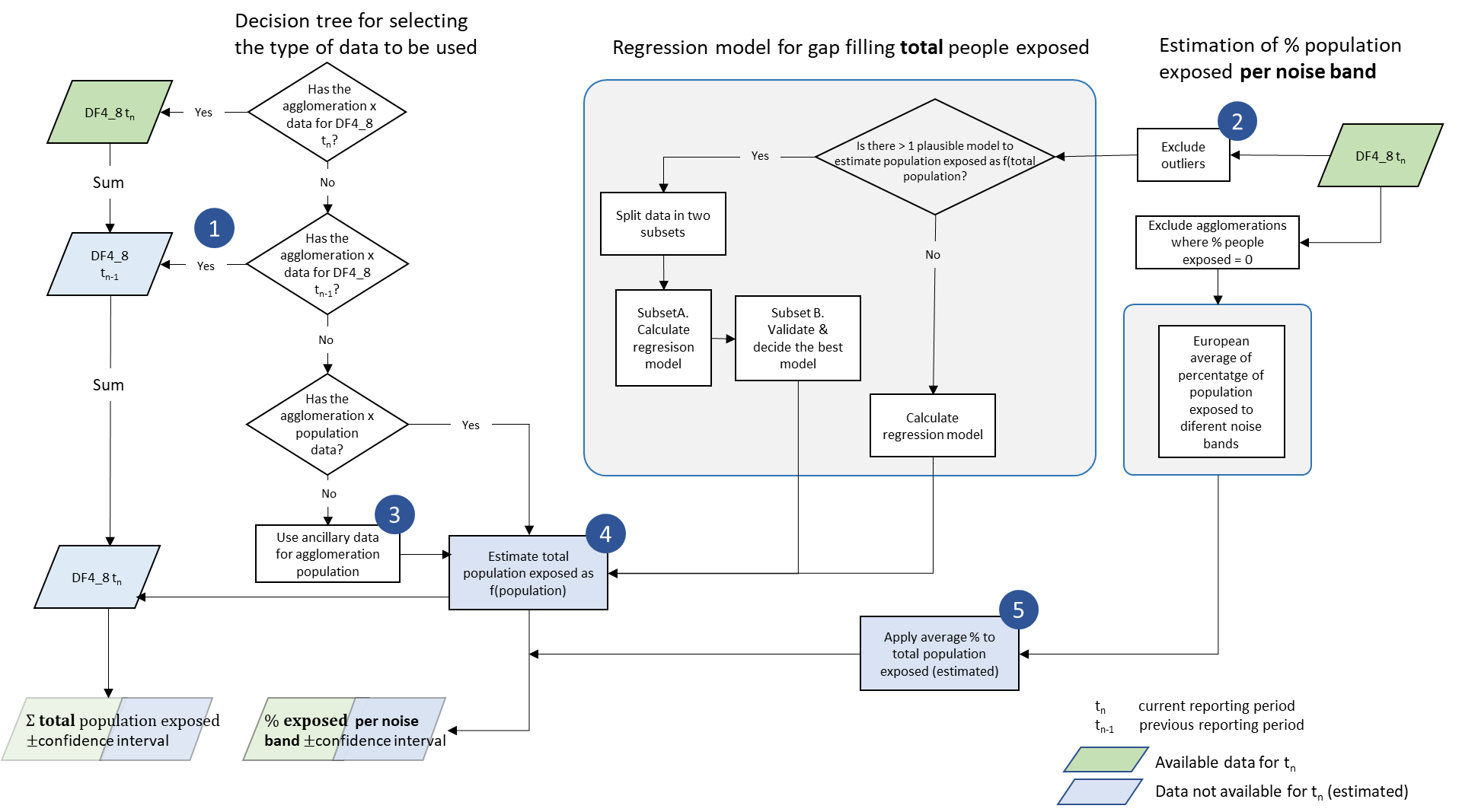
Table 4. Example of partial gap filling when data is missing for some noise bands. N.d., no data reported (missing data).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lden dB bands** | **55-59** | **60-64** | **65-69** | **70-74** | **>75** |
| Reported data | 11.7680 | 6.000 | 1.500 | *n.d.* | *n.d.* |
| % of people exposed distributed by noise band (European average) | 45,8 | 28,3 | 18,3 | 7,0 | 0,6 |
| Reported + gap filled data (italics) | 11.7680 | 6.000 | 1.500 | *9.500* | *800* |

Table 5. Methods used on the various steps of gap filling people exposed to noise from roads inside agglomerations. Numbers refer to the steps highlighted in Figure 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Method for gap filling** | **Explanation** | **Comment** |
| 1 | Use the data delivered in the previous reporting period for the same agglomeration, if available. | A comparative analysis described in Fons et al. (2017) concluded that this is the method with the smaller error. | The method is constrained to the availability of the data on the previous reporting cycle. |
| 2 | Exclude outliers from the reported data to be used for the regression (step 4) | Outliers from the percentage of change between tn-1 and tn based on the interquartile range (IQR, the difference between 3rd and 1st quartile). The exact threshold is ± 1,5 IQR. | Previous assessments have identified some extreme population changes exposed to noise between two reporting periods (Ramos, 2019). Therefore, the estimation of missing data in 2019 (Ramos, 2019), adopted a methodology to exclude outliers. |
| 3 | Estimate the total population from ancillary data. | If the country reported the agglomeration's delineation, the population could be derived from Urban Atlas described in Fons et al. (2015). The average error of the estimation based on Urban Atlas is 3%, ranging from 1 to 10%. If the delineation has not been reported, data from Eurostat can be used. The agglomeration population is the independent variable of the regression (step 3) to estimate population exposed when data from the previous reporting cycle is not available. | The reporting cycle of the Urban Atlas is always one year later than the END. |
| 4 | Estimate the population exposed from the regression between population exposed and population of the agglomeration. | The regression and correlation analysis between population exposed and potential predictors (total population, area,…) is documented in Fons et al. (2015). Later on, Fons et al. 2017) demonstrated that, if available, using data from the previous reporting period (step 1) is more accurate than the regression approach. | The current report provides a detailed description of the metrics to evaluate the best regression model, including estimating the error and confidence interval in the final aggregation of data (EU figures). There is no a priory regression model to be applied each time that the gap filling is developed. The regression model needs to be checked each time since the relationship is strongly dependent on the data included for the regression. |
| 5 | Estimate the % of the population exposed per noise bands (%) from the European average | Based on the total number of people exposed per each reported agglomeration, we calculate the percentage that each noise band represent versus the total number of people exposed, for Lden and for Lnight. Then we derive the mean at European level. This approach is discussed in depth in Fons et al. (2015). | Initially, the % was calculated on a country basis when were enough agglomerations reported. The assessment was done in Fons et al. (2017) highlights that using the country average has a similar error than the European average. Therefore, it was decided to use in all cases the European average for simplicity. |

Figure 1. Overview of the process for estimating the population exposed to roads inside agglomerations when data is not available. The methodology only applies to agglomerations that have to report according to END requirements. Numbers indicate specific methods for gap filling depending on available ancillary data -details are described in Table 5. Source: updated from Ramos (2019).



### Regression

The regression procedure could be summarised as follows:

1. Identify outliers from the percentage of change of the people exposed between current reporting period (t2) and previous reporting cycle (t1).
2. Plot Lden against the number of inhabitants to visually inspect the most suitable regression model. Since it is not always obvious which is the best model, the most plausible ones are retained and tested.
3. Transform the data if it improves the linearity (most common transformation in previous assessments was log transformation).
4. Divide the data in two groups to test different regression models: one group is used to run the regressions, and the second group is used to validate the models.
5. Calculate the regression model and analyse the corresponding statistics.
6. Estimate the number of people exposed on the validation subset and compare results between different models.
7. Apply the selected model to the missing data
8. Calculate the total people exposed in Europe with the confidence interval corresponding to the estimated data.

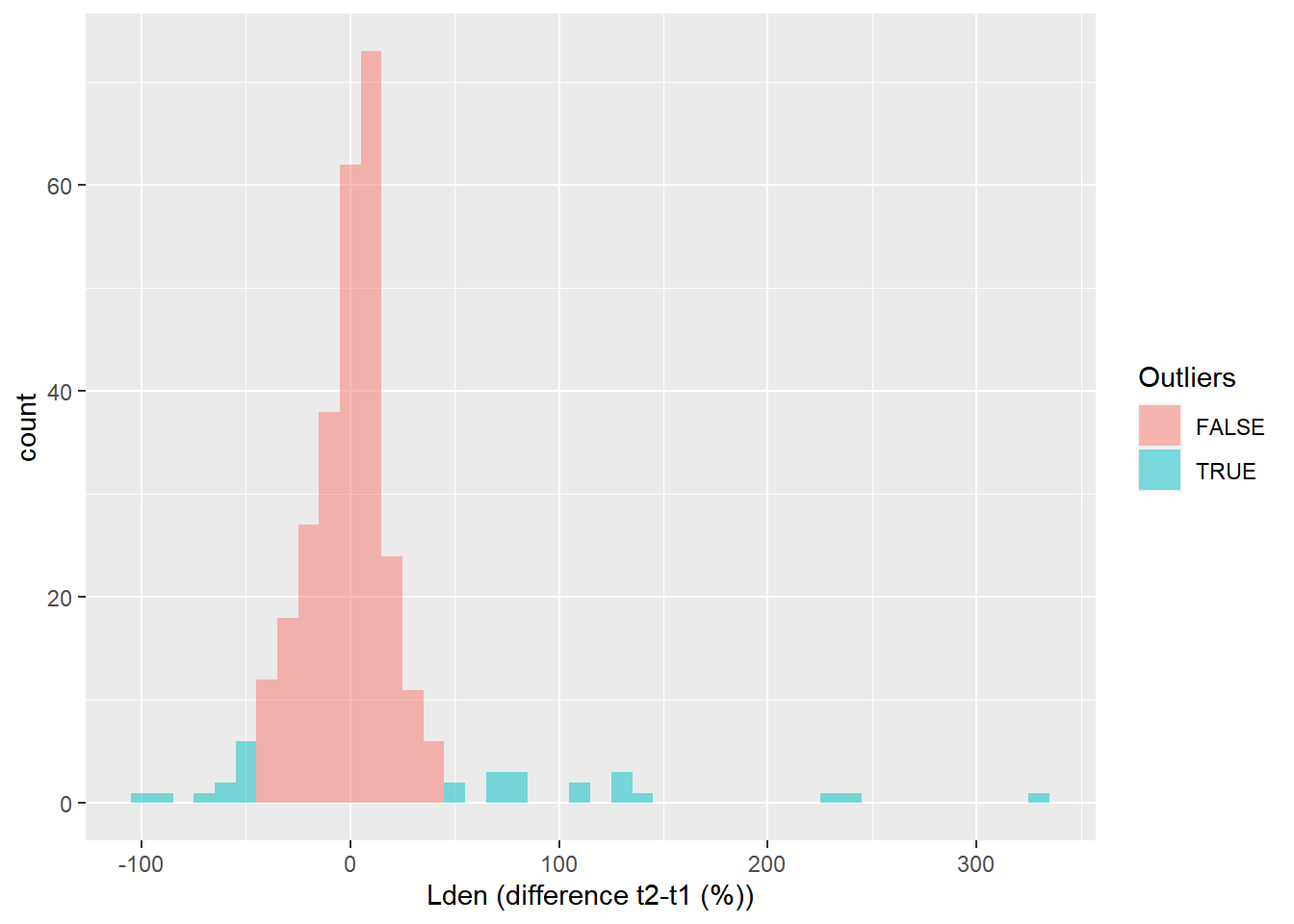
Each step is further described in the following sections. We have used the estimation of population exposed to Lden to illustrate the methodology. The same approach could be followed in other sources and Lnight when required.

### Identify outliers and adjust agglomerations with > 100% people exposed

Previous assessments have identified some extreme cases of change of population exposed to noise between two reporting periods (Ramos, 2019). Therefore, the estimation of missing data in 2019 (Ramos, 2019), adopted a methodology to exclude outliers. This methodology is based on the interquartile range (difference between 3rd and 1st quartile).

Figure 2 illustrates the distribution of both outliers and non-outliers for the percentage of change of population exposed to Lden from roads inside agglomerations equal or greater than 55 dB. In that case about 10% of the data were identified as outliers.

Figure 2. Histogram of the percentage of change of people exposed between t1 and t2 (roads inside agglomerations). The colour differentiate outliers from non-outliers. N = 299 agglomerations.



A small number of agglomerations declared more than 100% of the total population exposed. This may be possible due to the rounding to the nearest hundred of people exposed. Therefore, rounding may exceed by 250 people the agglomeration population (50 people per each noise band).

When the population exposed exceeds 100% of agglomeration population, we have adjusted the population exposed to the agglomeration population.

### Visual inspection of the relationship between people exposed and number of inhabitants

The first step for conducting the regression is to look at the scatter plot between the two involved variables: people exposed and number of inhabitants (Figure 3). At first sight it seems to fit a perfect linear regression. However, given the skewed distribution of both variables, a log-log transformation shows a better approximation to a normal distribution of both variables (Figure 4).

In that case we will test the following models:

People\_exposed = a + b\*Number\_inhabitants

People\_exposed = a + b\*Number\_inhabitants + c\*Number\_inhabitants2

Log(People\_exposed) = a + b\*log(Number\_inhabitants)

The second model correspond to a polynomial regression of order 2, which is useful when there is a small bending on the relationship between the two variables. Although it is not evident that it would be useful in that case, we will include it as illustration that it can be easily implemented and tested.

Zero values in the log transformation are problematic since log(0) is not a real number. Therefore, in case of zero values we should add a small amount to all zeros: 0,1. This does not have an impact on the total number of people exposed while keeping the agglomeration in the regression analysis. This is relevant for Lnight. As it is logical, no zero values have been observed in people exposed to Lden equal or greater to 55 dB.

Figure 3. Scatterplot of number of inhabitants and people exposed to noise from roads inside agglomerations (Lden equal or greater than 55 dB). N = 329 agglomerations.

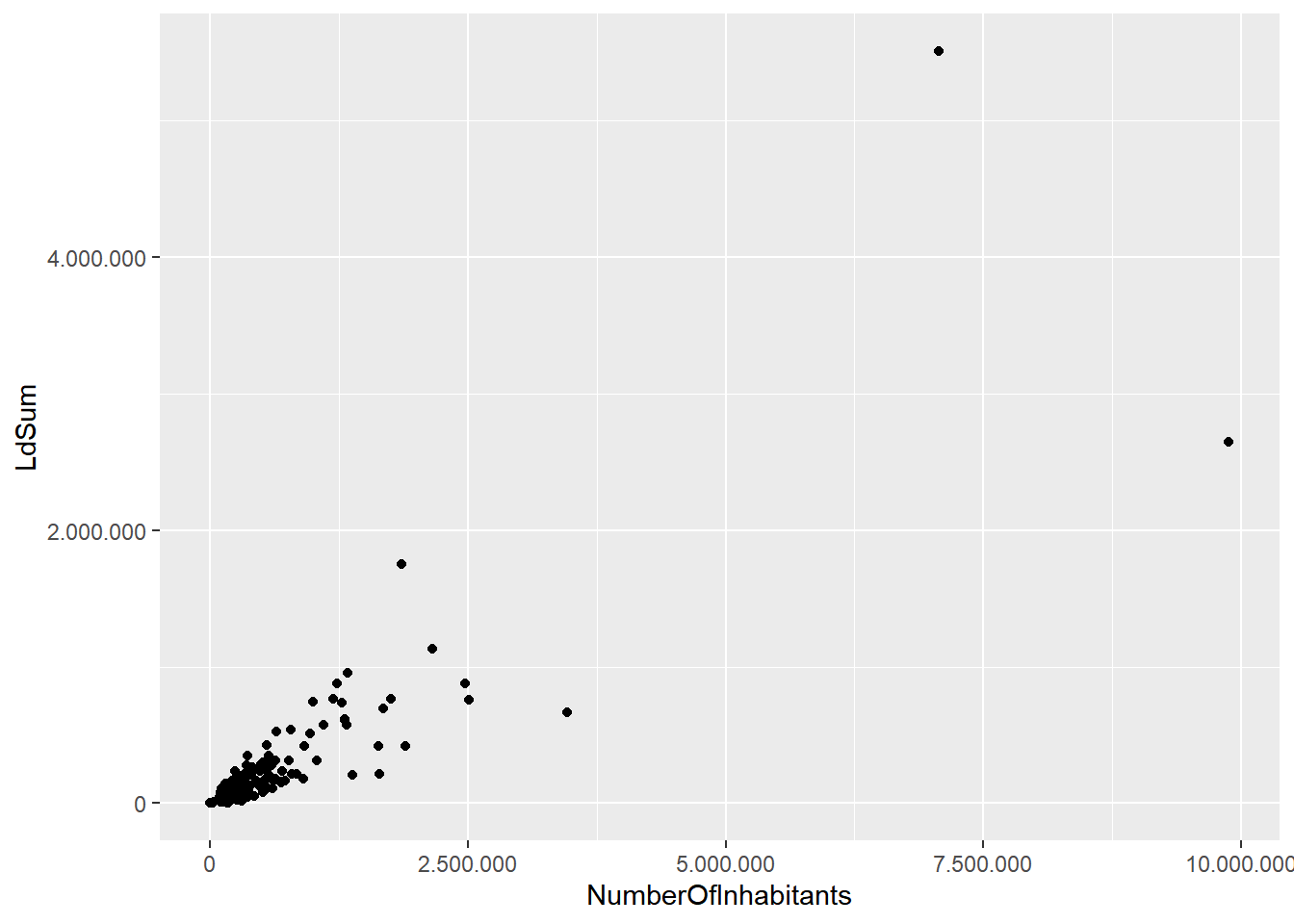
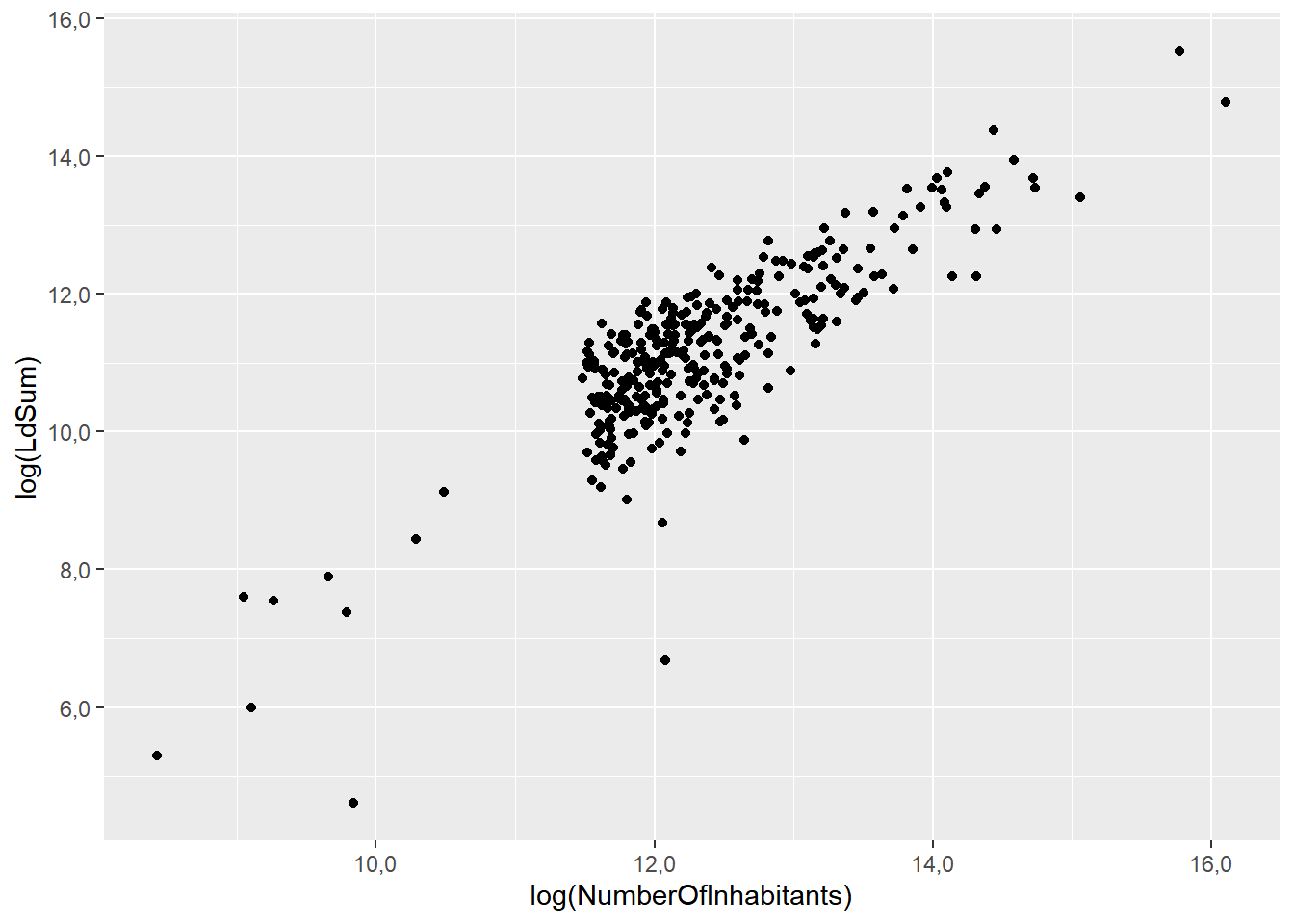


Figure 4. Log-log transformation of number of inhabitants and people exposed to exposed to noise from roads inside agglomerations (Lden equal or greater than 55 dB). N = 329 agglomerations.



### Data subsetting for regression analysis and validation

Data subsetting refers to divide the agglomerations where data has been reported in two groups: one for estimating the regression parameters and the other one to validate the regression. Then, we apply the regression model to the second subset, validation, which is independent from the data used to estimate the model. Finally, the outcome can be compared with the original data.

The following requirements are needed for subsetting:

* Minimum number of samples (agglomerations). The total number of (complete) data should follow the rule (Snee, 1977)

N > 2\*(nr of independent parameters) +25

In our case, we only include one parameter resulting in 27 as the minimum number of agglomerations required for a valid splitting.

* As a general rule data is splitted by a 70:30 ratio, being 70 for estimating the model and 30 for validation (Snee, 1977).

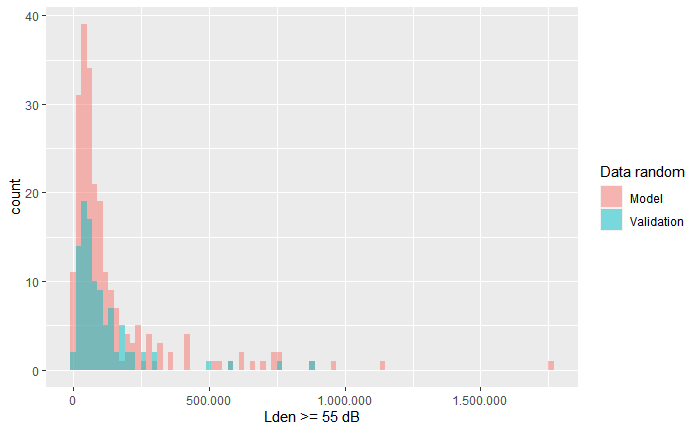
According to these rules, 226 agglomerations where data is reported -outliers excluded, have been randomly divided as follows:

* 226 agglomerations for estimating the regression model
* 103 agglomerations for validation

It is important that both subsets follow the same distribution. The Kolmogorov–Smirnov test is a nonparametric test of the equality of continuous one-dimensional probability distributions. The distribution of both model and validation subset are depicted in Figure 5. In that case the probability of the Kolmogorov–Smirnov statistic is 0,47. Therefore, the null hypothesis that both data sets have the same distribution is not rejected.

When the distributions significantly differ, a new random subset needs to be selected until the condition of the same distribution is met.

Figure 5. Distribution of two subsets of agglomerations: agglomerations used to estimate the regression model and agglomerations used for validation.



### Test models

The three models described in step 2 have been calculated on the subset of 226 agglomerations selected for that purpose:

* Model 1. Linear. People\_exposed = a + b\*Number\_inhabitants
* Model 2. Polynomic. People\_exposed = a + b\*Number\_inhabitants + c\*Number\_inhabitants2
* Model 3. Log-log. Log(People\_exposed) = a + b\*log(Number\_inhabitants)

From the performance perspective, model 3 has the highest R2, followed by model 2, and model 1 (Table 6). The other statistics, related to the model accuracy, can only be compared between model 1 and model 2 since are scale dependent -model 3 has been log transformed. In all cases, the lower of sigma, AIC, and BIC, the better. In that case we see that model 2 has lower values than model 1.

Sigma measures the average error performed by the model in predicting the outcome (Table 6). Therefore, sigma could be read as the error on estimating the people exposed to noise: there is a small difference (about 6.300 people) between model 1 (229.022 people) and model 2 (222.724 people).

Finally, the F statistic p value, which measures the statistical significance of the regression, is in line with the previous statistics: model 3 is more significant than model 2, and model 2 more than model 1 (Table 6).

In addition to the accuracy, diagnostic plots are relevant to identify possible weakness of the regression model. Figure 6 provides three of the most common diagnostic plots:

* Residual versus fitted. This plot shows if residuals have non-linear patterns. All models have certain deviation: model 1 and model 2 have a strong deviation on agglomerations with higher exposure (right side of the figure). Model 3 has a smaller deviation on both extremes. This indicates that other factors may be relevant to predict people exposed, and the number of inhabitants is only one factor -probably the main factor given the high level of prediction.
* Normal Q-Q. This plot shows if residuals are normally distributed. In that case models 1 and 2 show that the extremes are problematic, while model 3 has a better fit to the line (normal distribution).
* Residuals versus leverage. This plot helps us to find influential cases if any. Agglomerations that are outside the dotted red lines (Cook’s distance) are influential in the model, i.e. these agglomerations have more weight on defining the model, compared with the other agglomerations. Model 1 and model 2 have three agglomerations with a strong impact on the model (three points outside the Cook’s distance represented by the dotted line). The log transformation in model 3 was effective in removing the strong influence of extreme values.

Details of the output are provided in Annex I.

Table 6. Statistics of the three tested regression models.



Model 1

Model 2

Model 3





Figure 6. Diagnostic plots for the three regression models: residuals versus fitted (first row), normal Q-Q (second row), and residuals versus leverage (third row). Number are identifiers of agglomerations. Red line, trend of the plot. Dotted lines, Cook’s distance (0,5 and 1).

|  |  |  |
| --- | --- | --- |
| Linear | Polynomial | Log-log |
|  |  |  |
|  |  |  |
|  |  |  |

### Validation

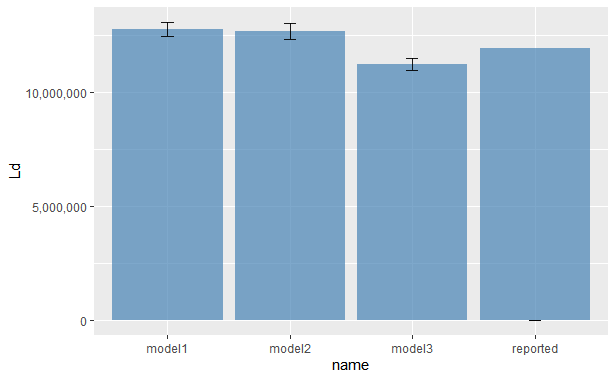
In the previous step we have seen that model 3 looked better because of higher R2 and better performance on the diagnostics.

Table 7 and Figure 7 show the results of applying the 3 regression models to the subset selected for validation (step 3). Model 1 and model 2 overestimate the population exposed, while model 3, closer to the reported values, underestimate the population exposed. The % of difference between reported data and the estimates from the three models are very close, being model 3 the one with lower percentage (5,9%). Also the confidence interval for model 3 is lower. Therefore, model 3 (log-log transformation) will be used to estimate the missing data.

Table 7. Results of the validation of the three regression models. N = 103 agglomerations.

|  |  |  |  |
| --- | --- | --- | --- |
|  | People exposed | % of difference | Confidence  interval |
| Reported | 11.951.291 |  |  |
| Model 1 | 12.777.297 | 6,9 | 323.394 |
| Model 2 | 12.689.144 | 6,2 | 363.808 |
| Model 3 | 11.245.629 | -5,9 | 279.311 |

Figure 7. Validation of the three regression models to estimate people exposed to noise from roads inside agglomerations (Lden equal or greater than 55 dB). Confidence interval (95%) provided for the regression models. n = 103 agglomerations.



### Estimate values for missing data

Once the regression model has been selected, the following steps are taken:

1. For each agglomeration where the data is missing, estimate the population exposed by applying the regression model. In that particular case, since, we selected a log-log regression we have to transform back the estimated population exposed (anti log).
2. Calculate the SE for each estimated value.

### Calculate the total people exposed in Europe

1. Sum all the estimated values
2. We need to calculate the SE of the sum which is obtained by *quadrature* ofthe individual SE
3. Finally, calculate the confidence interval.

### Estimate the people distributed by noise bands

Once the estimated total number of people exposed is calculated (previous step), we distribute the population between the different noise bands.

1. Based on the total number of people exposed per each agglomeration reported by Member States, we calculate the percentage that each noise band represent versus the total number of people exposed, for Lden and for Lnight, and then we derive the mean at European level. It needs to be taken into consideration that the percentage values have been obtained discarding the agglomerations providing 0 people exposed in all noise bands (Lden and Lnight, or Lden, or Lnight). Due to the rounding process, 0 could mean 0 to 49 people exposed; therefore, multiple combinations are possible with the same outcome of 0 people exposed. Consequently, an equal attribution of 20% of people exposed to each noise band only represents one of the multiple possible combinations. For that reason, agglomerations with 0 people exposed are excluded.
2. We apply the percentages to the agglomerations where we have estimated the total population, with the corresponding error of the estimate.
3. Finally, we aggregate all the agglomerations at European level, with the corresponding estimation of the confidence interval.

## Gap filling method for railway noise, aircraft noise and industrial noise inside agglomerations

This section summarises the method applied to gap fill exposure information for railways noise, aircraft noise and industrial noise inside agglomerations.

Figure 8 provides an overview of the process for estimating missing data, as Ramos (2019) described.

In each step, the decision tree uses the best approach according to the available ancillary data:

* Use data from the previous reporting period if available.
* When data from the previous reporting cycle is not available, data is estimated with the European average of the % of population exposed inside the agglomeration. In that case, no significant correlation was found between population exposed and other predictor parameters (e.g., the agglomeration population, area of the agglomeration); therefore, the European average is the best alternative (Fons et al., 2015).
* Based on the total number of people exposed per each agglomeration reported by the Member States, we calculate the percentage that each noise band represent versus the total number of people exposed, for Lden and for Lnight. Then we derive the mean at European level. It needs to be taken into consideration that the percentage values have been obtained discarding the agglomerations providing 0 people exposed in all noise bands (Lden and Lnight, or Lden, or Lnight). Due to the rounding process, 0 includes figures ranging from 0 to 49 people exposed; therefore, multiple combinations are possible with the same outcome of 0 people exposed. Consequently, an equal attribution of 20% of people exposed to each noise band only represents one of the several possible combinations. For that reason, agglomerations with 0 people exposed are excluded.
* We apply each noise band’s percentages to the agglomerations where we have estimated the total population, with the corresponding error of the estimate.
* Finally, we aggregate all the agglomerations at European level, with the corresponding estimation of the confidence interval.

A more detailed explanation and the basis for the methodology applied in each step is provided in Table 12.

The gap filling methodology also includes those few cases where data is missing for some noise bands. Then, **partial gap filling** is conducted as follows:

* Use data from the previous reporting period for the missing noise band(s).
* If data from the previous reporting period is not available use the European average of % of the population exposed by noise band to estimate the missing data. Table 4 provides an example of a partial gap filling when data from the previous period is not available.

Each step that involves the estimation of data, corresponding error is calculated. Finally, these errors are propagated to the aggregated European figures as described in sections 3.19 and 3.1.10.

Table 8. Methods used on the different steps of gap filling people exposed to noise from railways, airports and industry inside agglomerations. Numbers refer to the steps highlighted in Figure 18.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **Method for gap filling** | **Explanation** | **Comment** | |
| 1 | Exclude the agglomeration if it was reported as -1 (not applicable) at tn-1 | An agglomeration reporting -1 for a source in the previous reporting cycle (tn-1), meant that that source was not applicable according to the END specifications. Therefore, we retain the non-applicability at tn (Fons et al., 2016). Check done in previous gap-filled data (2016, 2017, 2018 and 2019) demonstrates that the assumption was correct in 90% of cases. Therefore, the potential underestimation of the European figure (data excluded) is more accurate than the overestimation, when all these agglomerations are gap filled. | This step only applies if data is not reported, and “not applicable” is not explicitly mentioned at tn. | |
| 2 | Exclude outliers from the reported data to be used for the European average (step 5) | Outliers from the change percentage between tn-1 and tn based on the interquartile range (IQR, the difference between 3rd and 1st quartile). The exact threshold is ± 1,5 IQR. | Previous assessments have identified some extreme population changes exposed to noise between two reporting periods (Ramos, 2019). Therefore, the estimation of missing data in 2019 (Ramos, 2019), adopted a methodology to exclude outliers. | |
| 3 | Use the data delivered in the previous reporting period for the same agglomeration, if available. | A comparative analysis described in Fons et al. (2017) concluded that this is the method with the smaller error. | The method is constrained to the availability of the data on the previous reporting cycle. | |
| 4 | Estimate the total population from ancillary data. | If the country reported the agglomeration's delineation, the population could be derived from Urban Atlas described in Fons et al. (2015). The average error of the estimation based on Urban Atlas is 3%, ranging from 1 to 10%. If the delineation has not been reported, data from Eurostat can be used. The agglomeration population is the used in step 5. | | The reporting cycle of the Urban Atlas is always one year later than the END. |
| 5 | Estimate the population exposed by multiplying the population of the agglomeration with the European average of the % of people exposed. | No significant correlation was found between population exposed and other predictor parameters (e.g., the agglomeration population, area of the agglomeration); therefore, the European average is the best alternative (Fons et al, 2015). |  | |
| 6 | Estimate the % of the population exposed per noise bands (%) from the European average. | Based on the total number of people exposed per each reported agglomeration, we calculate the percentage that each noise band represent versus the total number of people exposed, for Lden and for Lnight. Then we derive the mean at European level. This approach is discussed in depth in Fons et al. (2015). | Initially, the % was calculated on a country basis when were enough agglomerations reported. Fons et al. (2017) concluded that using the country average has a similar error than the European average. Therefore, the European average is used for simplicity. | |

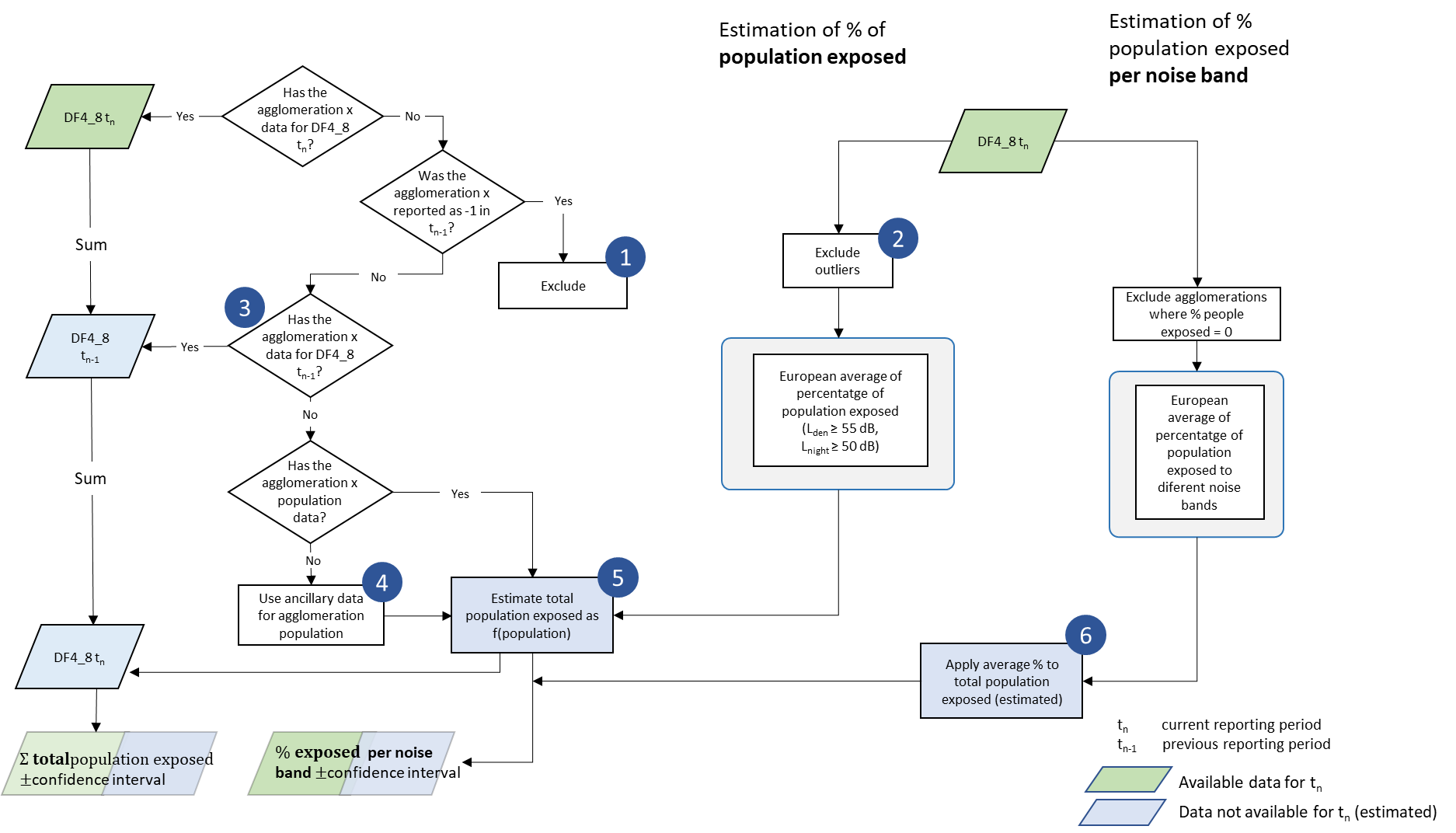


Figure 8. Overview of the process for estimating the population exposed to noise from rails, airports or industry inside agglomerations when data is not available. The methodology only applies to agglomerations that have to report according to END requirements. Numbers indicate specific methods for gap filling depending on available ancillary data -details are described in Table 7. Source: updated from Ramos (2019).

# END major roads and major railways exposure data outside agglomerations: gap filling

## Gap filling method

This method is based on the working paper establishing a methodological proposal to interpolate a complete coverage on noise exposure at EU level (ETC/ACM, 2015).

Figure 9 provides an overview of the process for estimating missing data, as Ramos (2019) described.

In each step, the decision tree uses the best approach according to the available ancillary data:

* Partial gap filling if reported data is not complete. Data completeness can only be evaluated if exposure has been delivered by the road and rail segments. Then network segments are linked to DF1\_5 dataflow to match segments to be reported with the actual data reported. Missing segments are gap filled with the regression between people exposed and the length of the transport network (the procedures is the same as described for roads inside agglomerations. It should be noted that the length of the transport network also includes major source inside agglomerations. However, the data reported on people exposed refers only to people outside agglomerations. When the exposure information has been delivered as one single value for the entire network, the codes are supplied as -1 or -2 or the codes between dataflows (DF1\_5 and DF4\_8) do not match, then the comparison of the code is not possible, and the dataset is assumed as complete.
* If data is not reported, use data from the previous reporting period if available and complete (same procedure as explained in the above bullet point to evaluate completeness).
* When data from the previous reporting cycle is not available or not complete, the regression between the number of people exposed and the transport network's length has been calculated. Then the regression has been applied to estimate missing data. In that case, complete gap filling is applied, i.e. even if some data (incomplete) is available from the previous period, the regression is applied to the full extent of the transport network. It should be noted that the length of the transport network also includes major source inside agglomerations. However, the data reported on people exposed refers only to people outside agglomerations.
* Once the total number of people exposed is estimated, we distribute the total population exposed to the different noise bands based on the European average of the population exposed by noise bands. The European average is calculated with all the available data, even if it is incomplete for a certain region of the country. The European average discards the countries or regions providing 0 people exposed in all noise bands. Due to the rounding process, 0 could mean 0 to 49 people exposed. Therefore attributing 20% to each noise band would not be accurate since other options would also be feasible.

Table 9. Methods used on the different steps of gap filling people exposed to noise from roads inside agglomerations. Numbers refer to the steps highlighted in Figure 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Method for gap filling** | **Explanation** | **Comment** |
| 1 | Use the data delivered in the previous reporting period, if available. | A comparative analysis described in Fons et al. (2017) concluded that this is the method with the smaller error. | The method is constrained to the availability of the data on the previous reporting cycle, and the data is complete. |
| 2 | When the data is not complete for a certain country, estimate the missing data with the regression between people exposed and the length of the transport network. | The methodology for partial gap filling is described in Fons et al. (2015).  The regression and correlation analysis between population exposed and potential predictors (country area, length of transport network) is documented in Fons et al. (2015). Later on, Fons et al. 2017) demonstrated that, if available, using data from the previous reporting period (step 1) is more accurate than the regression approach. | The current report provides a detailed description of the metrics to evaluate the best regression model, including estimating the error and confidence interval in the final aggregation of data (EU figures). There is no a priory regression model to be applied each time that the gap filling is developed. The regression model needs to be checked each time since the relationship is strongly dependent on the data included for the regression. The approach described in section 3.1 is also valid here. |
| 3 | Estimate the population exposed from the regression between population exposed and km of road or rail. | The regression and correlation analysis between population exposed and potential predictors (country area, length of transport network) is documented in Fons et al. (2015). Later on, Fons et al. 2017) demonstrated that, if available, using data from the previous reporting period (step 1) is more accurate than the regression approach. | The current report provides a detailed description of the metrics to evaluate the best regression model, including estimating the error and confidence interval in the final aggregation of data (EU figures). There is no a priory regression model to be applied each time that the gap filling is developed. The regression model needs to be checked each time since the relationship is strongly dependent on the data included for the regression. The approach described in section 3.1 is also valid here. |
| 4 | Estimate the % of the population exposed per noise bands (%) from the European average | Based on the total number of people exposed, we calculate the percentage that each noise band represent versus the total number of people exposed, for Lden and for Lnight, and then we derive the mean at European level. This approach is discussed in depth in Fons et al. (2015). |  |

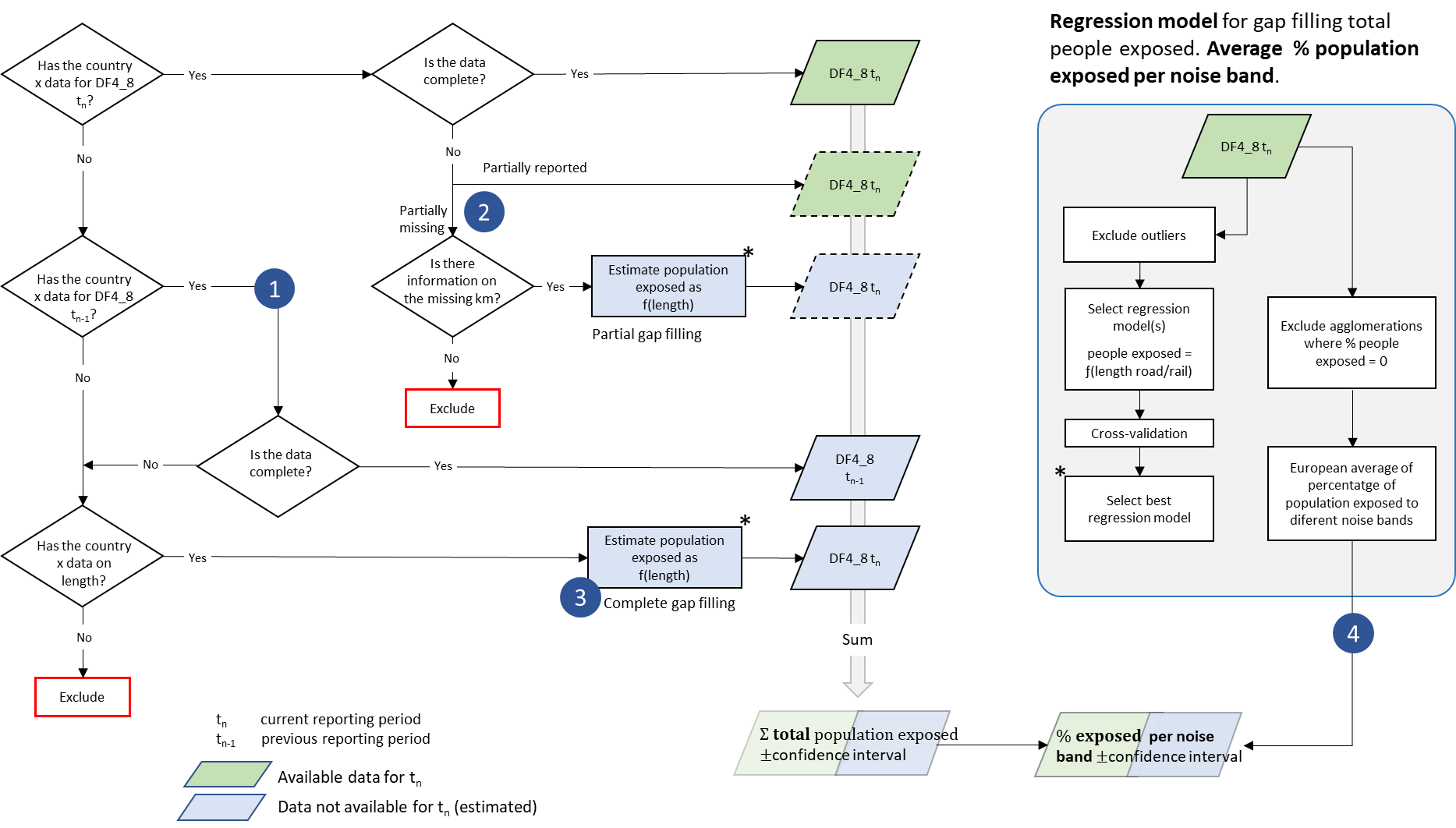


Figure 9. Overview of the process for estimating the population exposed to noise from major roads and major rails when data is not available. The methodology only applies to countries that have to report according to END requirements. Numbers indicate specific methods for gap filling depending on available ancillary data -details are described in Table 9. Dotted lines: data partially reported, and data partially gap filled. Source: updated from Ramos (2019).

# END Major airport exposure outside agglomerations: gap filling

## Overview

The current methodology to estimate missing data on the population exposed to major airports outside agglomerations is presented in Figure 1 (Ramos, 2019). In this case, we start with the calculation of the average relative change of population exposed between the current reporting period (t2) and previous cycle (t1) as described in Jones (2013):

Where *n* is the number of major airports with reported data for the period t1 and t2, being t2 the current reporting cycle, and *i* is the *i*th major airport where data is available.

In practical terms this relative difference can be expressed as ratio and directly applied to those major airports where data for the current reporting period is missing, but available for the previous period:

Then, this ratio It should be noted that the use of the ratio of change could be extended back up to two reporting cycles. For example, in the case that data for a certain major airport is only available for 2007, then we calculate the ratio of change for the period 2007 – 2017 and apply this ratio to the data of 2007.

In this case outliers have also to be excluded from the calculation as explained in the case of roads inside agglomerations (10).

The use of the relative difference is based on the fact that it provides better estimates than any other predictor, e.g. number of annual flights (Fons et al., 2016). The low correlation between people exposed and the number of flights is explained by the fact that exposure to aircraft noise is strongly dependent on local conditions, including specific operational measures (take-off and landing routes, time of the day,…), meteorological conditions, or land use planning. However, the current approach has its limitations since it assumes a homogenous change ratio in all agglomerations. Figure 11 and Figure 12 illustrate the distribution of the relative difference between 2012 and 2017. The value ranges from -1 (100% decrease on exposure) to 1,5 (150% increase of population exposed). About 9 out of 40 major airports are outliers (23%).

Figure 10. Overview of the workflow for estimating the population exposed to major airports outside agglomerations when data is not available. Source: Ramos (2019).



Figure 11. Box plot of the relative change of population exposed to noise from major airports between 2012 and 2017 (Lden ≥ 55 dB). Outliers are indicated in red. The lower and higher limits of the box correspond to the 1st and 3rd quartile, respectively. The horizontal line inside the box is the median.

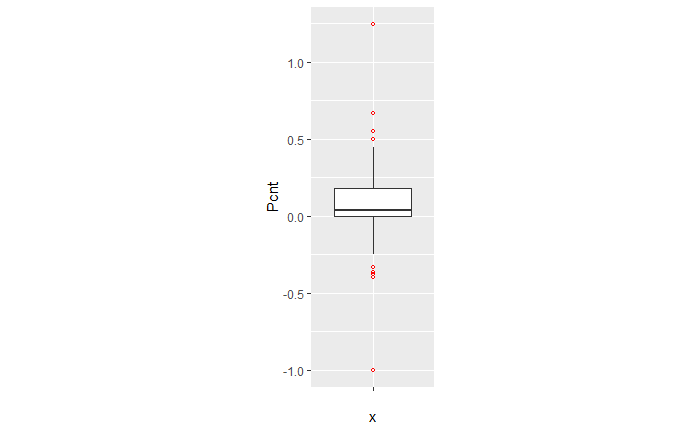
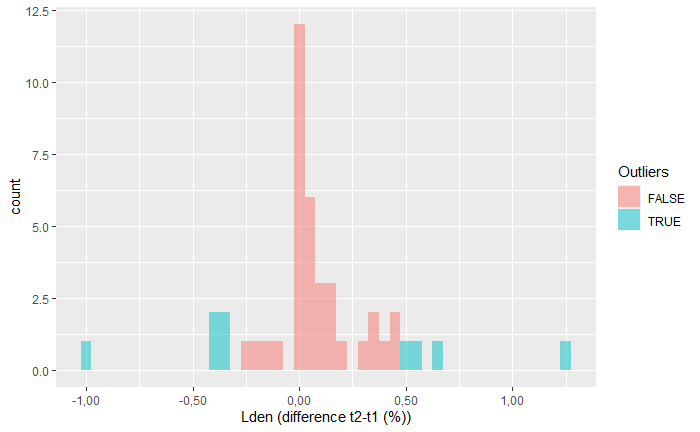


Figure 12. Distribution of the population relative change exposed to noise from major airports between 2012 and 2017 (Lden ≥ 55 dB). Outliers in green.



Lden Relative change (2017 – 2012)

Since the outliers are equally distributed around the mean, there is no major impact on the average ratio if we include them in the calculation (Table 10). However, the standard error decreases by 40% when outliers are excluded. Moreover, if outliers are not excluded, the estimated confidence interval of the change ratio ranges from a small decrease (-0,03) to 0,18 increase because the standard error (0,10) is higher than the average (0,08), as presented in Table 10.

Table 10. Mean, upper and lower boundaries of the confidence interval (95%), and standard error (SE) of the change of population exposed to noise from major airports between 2012 and 2017 (Lden ≥ 55 dB). Data is presented for the complete set of available major airports and for the subset without outliers. N is the number of major airports.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Lower boundary** | **Average** | **Upper boundary** | **SE** | ***n*** |
| **With outliers** | -0,03 | 0,08 | 0,18 | 0,10 | 40 |
| **Without outliers** | 0,03 | 0,08 | 0,14 | 0,06 | 31 |

Given all these uncertainties, the current report provides an alternative method and test its suitability and potentially higher performance. Our hypothesis is that delineating an approximated noise contour band in those major airports that have not reported data will better estimate the population exposed than the current method. We assume that with this approach, we better capture local conditions with a reasonable effort of computation.

The methodology has been tested in 10 airports where all the information has been reported (Table 11)

Table 11. Airports selected to test the proposed methodology.

|  |  |  |  |
| --- | --- | --- | --- |
| **Airport** | **ICAO code** | **Annual traffic** | **Runaways** |
| Berlin-Tegel | EDDT | 182200 | 2 runways (side-by-side) |
| Berlin-Schönefeld | EDDB | 70324 | 2 parallel runways |
| Copenhagen | EKCH | 251799 | 3 runways cross model |
| Hamburg | EDDH | 153876 | 2 runways cross model |
| Helsinki-Vantaa | EFHK | 168704 | 3 runways, 2 parallel, one crossing (almost perpendicular) |
| Lisbon | LPPT | 159795 | 2 runways cross diagonal disposal (45 degrees) |
| Milano-Malpensa | LIMC | 166509 | 2 parallel runways |
| Napoli | LIRN | 64712 | 1 runway |
| Wien | LOWW | 226811 | 2 runways diagonal disposal |
| Budapest Ferihegy | LHBP | 96705 | 2 runways diagonal disposal |

## A methodology based on estimating the noise contour around runaways

### Overview

The methodology to estimate the population exposed to major airports (Lden) is synthesised in Table 12. The method for Lnight would be similar.

As can be seen, the methodology has two major elements of uncertainty:

* Delineation of the noise contour around the major airports. These contours depend on several factors: number and length of runaways, local regulations, meteorological conditions, or orography, just to name some of them.
* Assuming that all the population living inside the contour is exposed to noise. Therefore, existing measures like building insulation are not considered since it is not feasible to introduce this component at European scale.

Table 12. Methodology to estimate population exposed to major airports (inside and outside agglomerations). The

|  |  |  |  |
| --- | --- | --- | --- |
| **Steps** | | **Output** | **Comments** |
| 1. Delineation of the noise contour for Lden 55dB (lower boundary) | | Noise contour for Lden 55dB (lower boundary) |  |
|  | 1. Identify runaway(s) |  | From satellite imagery draw a simple line representing the full length of each runaway |
|  | 1. Delineation of the contour for Lden 55dB around runaways |  | Delineation of the contour based on a certain buffer around the runaways |
| 1. Cross the contour of Lden 55dB with the delineation of the agglomeration | | In case of the presence of one or more agglomerations: contour outside and contour inside agglomerations(s) | This step is needed to differentiate the people exposed **outside** and **inside** the agglomeration (if the contour intersects with one or more agglomerations) |
| 1. Calculate the population inside the noise contour (inside and outside the agglomeration) | | Distribution of the peoples exposed by noise bands | Cross the areas of the previous step with the population grid. The obtained value will be an estimate of the population exposed to a major airport (inside and outside agglomerations). |
| 1. Distribute the total population exposed to major airport by noise bands | |  | Apply the European average of % of population exposed distributed by noise bands. |

### Data requirements

The following data has been used to test the methodology:

* Reference image for runaways: Google Maps
* Delineation of agglomerations as provided by the information reported by countries according to END specifications
* Noise contour bands reported by countries
* Population. Outside agglomerations, GEOSTAT population at 1 km grid[[2]](#footnote-2)

### Delineation of the runaways and noise contours

The most critical issue is the delineation of the noise contour. During the testing phase, it was taken into consideration the use of wind maps and its monthly/year direction means as usually aircraft depart and arrive counter wind. As well ATS routes, traffic maps and the use of waypoints and fixes to determine routes were analysed to determine the most common path of planes in each airport, but unfortunately, no relevant data was found. Many airports implement noise mitigation techniques (noise abatement procedures), that increases the uncertainty when creating a common delineation noise model, including the following:

* Defining noise abatement procedures that avoid residential areas as far as possible and avoid over-flying sensitive sites such as hospitals and schools
* Using continuous descent approaches and departure noise abatement techniques
* Ensuring that the optimum runway(s) and routes are used as far as conditions allow
* Avoiding unnecessary use of auxiliary power units by aircraft on-stand
* Building barriers and engine test-pens to contain and deflect noise
* Towing aircraft instead of using jet engines to taxi
* Limiting night operations
* Limiting the number of operations or the extent of a critical noise contour
* Providing noise insulation for the most severely affected houses
* Applying different operational charges based on the noisiness of the aircraft
* Monitoring individual noise levels and track keeping and penalising any breach

However, since this approach is quite an effort consuming, it was decided to take one airport as a model and replicate the noise contour on the airports without data, adapted to the length of the runaways. For the one runway setup airports, the replication of the Tegel delineation is enough to have a strong approach close to the reality on the gap filled airports. When talking about two or three runway setup, the complexity is higher. Despite that, taking into account what has been reported on contour maps, the following table shows the type of artificial delineations created to calculate the population exposed and gap fill missing data:

In relation to the generated contour maps for gap filling, the best way to replicate the possible track of the aircraft is to look at what is happening in the airports, there is information reported. Depending on many factors such as frequent wind flows, orography, local regulations, etc, an aircraft creates one type of noise path or another. On perfect conditions, the most common is that the plane creates a noise track similar to what was reported in the Berlin-Tegel airport, a very smooth pattern of noise where small bellies, corresponding to the end of the runway, where aircraft reaches its maximum power in corresponding to the end of the runway, where aircraft reaches its maximum power in the ground, arises in the middle of this ellipse form.

For the one runway setup airports the replication of the Tegel delineation is enough to have a strong approach close to the reality on the gap filled airports. When talking about two or three runway setup the complexity is higher. Despite that, taking into account what have been reported on contour maps the following table shows the type of artificial delineations created to calculate the population exposed and gap fill missing data.

Table 13.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of runway** | **Delineation model** | **Rationale** | **Approaches** |
| 1 runway | Tegel ellipse | the most common type of aircraft noise track | for shorter runways scale the Tegel ellipse the proportional difference of the runway length in relation with the Tegel main runway length (3km) |
| 2 runways cross | Tegel ellipse both runways | hard to determine which of the two runways have more movements than the other producing larger noise tracks. In general main runways are built on an east-west line and the secondary track to fit local conditions of winds, orography, etc. The solution found was to decrease Tegel ellipse to half of its area in the “secondary” (not main) runway oriented north-south. | The secondary approach is to use the length of the runway as a factor to determine the size of the Tegel ellipse. Check Wien case study which has 3,5km each runway and Lisbon having the main runway with 3,8km, and the secondary runway with 2,4km. Precise runway length available under external links (3) |
| 2 runways parallel | Tegel ellipse both runways | normally very similar noise behaviour for both runways. | Tegel ellipse seems to fit most of the cases when replicated twice in this case. Use the length of runways as an indicator of the Tegel ellipse size. Scale accordingly. |
| 3 runways, 2 parallel, 1 crossing | Tegel ellipse parallel runways and 1/3 size Tegel ellipse for crossing runway | In most cases, the third runway is shorter having less movements being used mainly for local flights or when weather conditions are rough. The proposal is to reduce the Tegel ellipse to 1/3 of its area approximately. In general very few movements/year are produced in that runway. | Use the length of runways as an indicator of the Tegel ellipse size. Scale accordingly. |
| 3 runways (or more) other setup | Tegel ellipse parallel runways and 1/3 size Tegel ellipse for crossing runway | The approach in this case is to use Tegel ellipse for the larger runways and decrease the ellipse to 1/3 if the runway is smaller. | Use the length of runways as an indicator of the Tegel ellipse size. Scale accordingly. |

### Detailed description of the methodology

The methodology could d be described as follows:

* copy model contour map delineation into a new feature
* move the central vertex into the central part of the runway and replicate it to all the airports being analysed
* rotate the delineation so it fits the orientation of the runway
* scale delineation into 1/3 of the model for secondary runways generally with less traffic
* dissolve delineations into one single area. Only in the case of more than 1 runway
* calculate area by hectares for all airports and all areas, reported contours and generated contours for gap filling
* use identity to print the agglomeration border into the model delineation so the areas are separated into inside or outside agglomeration
* to calculate the population exposed use “extract by mask” tool and select the working zone as the mask and the population raster as the data to extract
* convert the output raster of population exposed into integer values with the “int” tool
* sum the values of population for each specific raster zone

*Maps to compare reported contours and generated contours for gap filling:*

Arcmap file:

S:\Common workspace\Noise\2020\Mairports\Testing\_01.mxd

Images:

S:\Common workspace\Noise\2020\Mairports\img

### Outcome of the test

Table 13 provides an overview of the result of two methodologies to estimate people exposed to noise from major airports when this information has not been reported:

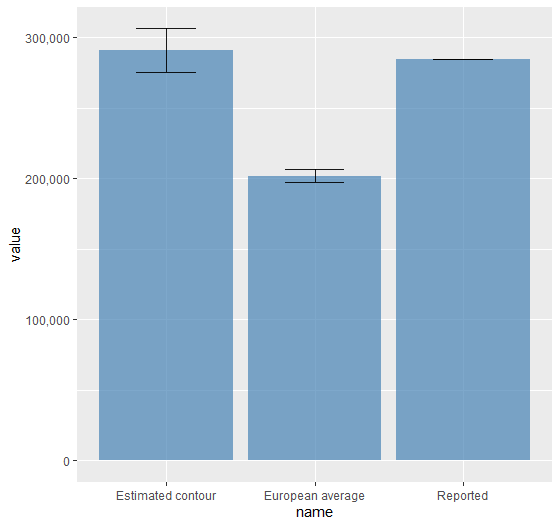
* Using a European average with the reported data
* Delineating the noise contour for Lden 55 dB, and calculate the population inside.

As can be seen, the gap filling with the European average is less accurate but has higher precision, while the outcome of delineating the noise contour is the opposite: it is very accurate, although three times less precise. The implication on the final values can also be seen in Figure 13. In that case, it seems that is preferable a more accurate estimation (closer to the reality) with a reasonable precision of 7%, which means that real value will be ± 20.000 people.

Table 14. Accuracy and precision of the two tested methodologies to gap fill population exposed to major airports outside agglomerations. Precision values provided as number of people exposed are illustrative and valid for a reported value of 284.000 people exposed (9 major airports).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methodology** | **Description** | **Accuracy** | **Precision** | |
| % | People exposed |
| European average | Very easy to apply. Just calculate the average ratio of change between two periods | 29% | 2,2% | 4.500 |
| Delineation of the contour for Lden 55 dB | Calculation intensive. Needs to identify the runaways, delineate the contour, exclude the area inside the agglomeration, and attribute the population. | 2,3% | 6,9% | 20.000 |

Figure 13. Reported and estimated values for people exposed to noise from 9 major airports (Lden ≥ 55 dB). Reported, reported data. Estimated contour, people exposed to noise has been estimated by delineating an approximated contour band. European average, people exposed has been estimated by applying to the data from the previous reporting cycle the European average of change of population exposed.



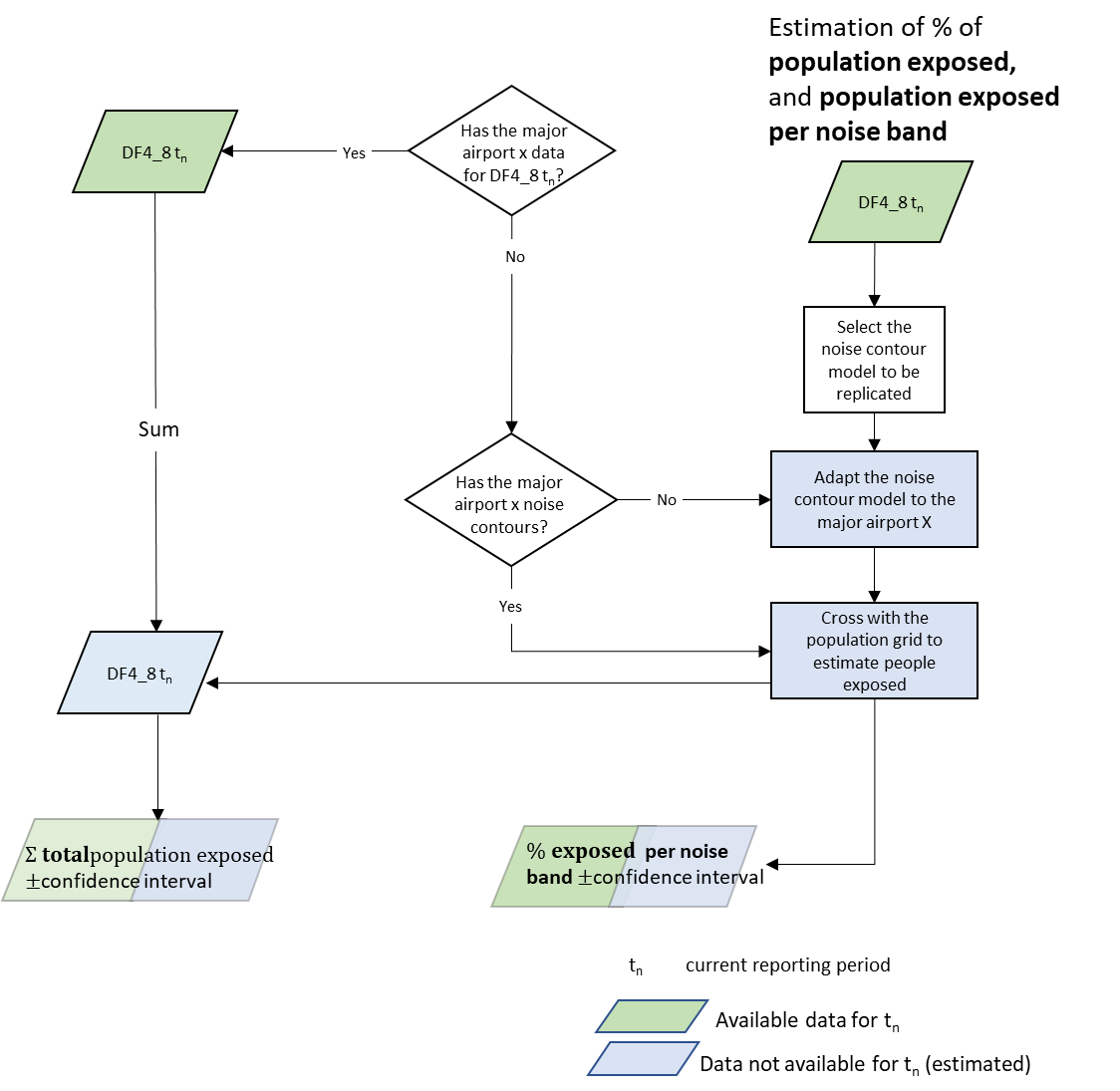
## Summary of the proposed methodology

The approach described in the previous sections is summarised in Figure 14. In that case, we don’t consider to use data from the last reporting period since it produces a higher error.

In summary, when data not reported:

* If noise contours are provided, cross them with the population grid and calculate the population exposed.
* If contours are not provided, delineate the contour based on the reference contour (see section5.2.3), and then cross with the population grid.

Figure 14. Overview of the process for estimating the population exposed to noise from major airports when data is not available.



# Conclusions

This report provides some improvements on gap filling that could be implemented in the next reporting of END data:

* The selection of the best fit of regression has been systematised and can be replicated with the R script developed in this task. It also provides the results with the corresponding confidence interval, by integrating the propagation of errors as consequence of several steps involved in the final calculation. This is relevant since this process may be very tedious without proper design of the workflow. It also facilitate to complete the workflow on the same environment without the need to change between applications.
* The error of applying European averages has been calculated, being able to assess its accuracy.
* The proposed method for major airports improves the estimate of the people exposed, at the cost of a more intensive processing.

# References

Fons, J., Sáinz, M., Blanes, N., Houthuijs, D. 2015. Methodological proposal to interpolate a complete coverage on noise exposure at EU level. ETC/ACM Working paper.

Fons, J., Blanes, N., Ramos, M., 2017. END gap filling data 2012: outcomes’ evaluation. Comparison between gap filled data results in 2016 and 2017. ETC/ATM Working paper.

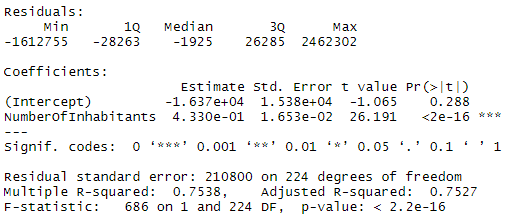
Jones, N., 2013. Forecasting ENDRM DF4\_8 data to 2020, 2030 and 2050. Extrium P058.

Snee, 1977. Validation of regression models: Methods and Examples. Technometrics, 19(4): 422.

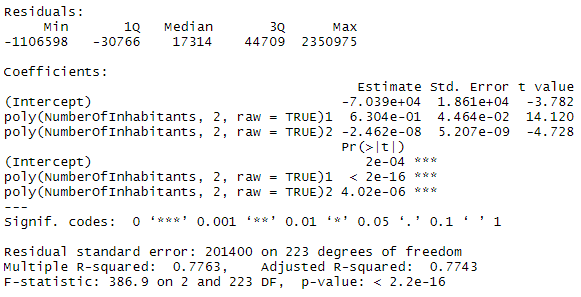
# Annex I. Summary of statistics

Summary of statistics of the regression between people exposed to noise from roads inside agglomerations (Lden  >= 55 dB) and Number of inhabitants (see section 1.1.2.4 Validation).

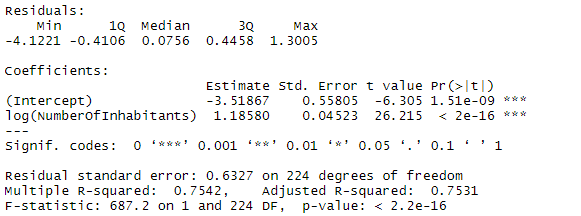
**Model 1. Lineal**



**Model 2. Polynomial (2)**



**Model 3. Log-log**



# Annex II. R script

R Script (Rnotebook) that process interactively the following steps :

* Import of the data to be processed
* Calculate and exclude outliers
* Builds the total people exposed in Europe
* Select the data reported
* Gap fill with data reported in previous period if available
* Regression for the remaining data
* Subset reported data in test and validation
* Define the models (interactive)
* Test the models
* Select the model (interactive)
* Apply the regression
* Calculate the total with confidence interval

Script

---

title: "Gap filling of people exposed to road inside agglomerations"

author: "Jaume Fons-Esteve"

date: 04/12/20

output:

html\_document:

df\_print: paged

toc: true

toc\_depth: 2

toc\_float: true

word\_document: default

html\_notebook: default

---

```{r setup, echo=FALSE}

library(knitr)

knitr::opts\_chunk$set(echo = FALSE)

```

# Overview

This notebook process data to estimate missing values of people exposed to road

noise inside agglomerations.

The methodology follows the one described in the report

E:\Universitat Aut?noma de Barcelona\INTERFASE - Documentos\1.Current\EEA\_ETCATNI\_2020\WP 1.1.5.1 Noise data\ST3. Gap filling\Documentation\EionetRep\_ETCATNI\_2019\_1\_Noise indicators

under the EnvNoiseDirective FINAL.pdf

The methodology has been updated and available at the following link

LINK TO THE REPORT 2020

Summary of the process

\* year1 clean data

\* year2 clean data (current year)

\* Subset A -> data available

\* Subset B -> missing data, but available at t-1

\* Subset C -> estimate from regression

\* At the end merge A+B+C

“`{r load}

# Load packages -------------

library(janitor) #needed for names(data) list variables in datset. Reformat names

library(dplyr) #select a subset of columns

library ("data.table")

library(readxl)

library(pivottabler) #pivot tables

library(ggplot2) #graphics

library(scales) #modify scales and format of numbers

library(modelr) #helper functions for computing regression model performance metrics

library(broom) #creates a tidy data frame containing the model statistical metrics

library(caret) #Classification And REgression Training

# Name of the source to be processed. This facilitate recycling the scripts.

source = 'aggl\_rd'

```

# 1.Import and prepare the data

Data is imported from Excel. PASTE THE PATH TO FILE .

Creates t1 = data from time 1; t2 = data for time 2

```{r import, echo=FALSE}

t1 <- read\_excel("E:/IDrive-Sync/UAB/1.Current/2020\_ETC-ATNI/1151\_Gap\_filling/Data/END\_DataReported\_2012\_F180912\_T190101.xlsx",

sheet = "Agg\_road\_reported")

t2 <- read\_excel("E:/IDrive-Sync/UAB/1.Current/2020\_ETC-ATNI/1151\_Gap\_filling/Data/END\_DataReported\_2017\_F180912\_T190101.xlsx",

sheet = "Agg\_road\_reported")

```

Rename noise bands

\* Add a column with the name of the source (Agg\_road)

\* Calculate Lden>=55 (Exclude -2 -1 -9999)

\* Calculate Lnight>=50 (Exclude -2 -1 -9999)

\* LdRprt = -2, no data; -1 not applicable; 1 reported

\* Ld\_More100Pcnt = TRUE, FALSE LdSum > NrOfInhabitants

\* Create a new table t1.cln, t2.cln

```{r clean, echo=FALSE}

# t1 -> t1.cln cln=clean

# New variables: LdSum, LdRport (reported), Ld\_More100Pcnt (LdSum > Inhabitants)

# New variables: LnSum, LnRport (reported), Ln\_More100Pcnt (LnSum > Inhabitants)

t1.cln <- t1 %>%

rename(Ld55 = nLden5559, Ld60 = nLden6064, Ld65 = nLden6569,

Ld70 = nLden7074, Ld75 = nLden75, Ln50 = nLnight5054,

Ln55 = nLnight5559, Ln60 = nLnight6064, Ln65 = nLnight6569,

Ln70 = nLnight70) %>%

mutate(Src = source) %>% #Add a column with the bame of the source

mutate(LdSum = ifelse(Ld55 > -1, Ld55+Ld60+Ld65+Ld70+Ld75, NA)) %>%

mutate(LnSum = ifelse(Ln50 > -1, Ln50+Ln55+Ln60+Ln65+Ln70, NA)) %>%

mutate(LdRprt = case\_when(Ld55 < -1 ~ -2, #-2 & -9999 assigned -2

Ld55 == -1 ~ -1, #-1 assigned -1

Ld55 > -1 ~ 1)) %>%

mutate(LnRprt = case\_when(Ln50 < -1 ~ -2, #-2 & -9999 assigned -2

Ln50 == -1 ~ -1, #-1 assigned -1

Ln50 > -1 ~ 1)) %>%

mutate(Ld\_More100Pcnt = if\_else(LdSum > NumberOfInhabitants ,TRUE, FALSE)) %>%

mutate(Ln\_More100Pcnt = if\_else(LnSum > NumberOfInhabitants ,TRUE, FALSE))

# t1 Reorder columns, exclude extra info

t1.cln <- t1.cln[c("Src","Ctry","Ctry2", "ReferenceYear", "RLID", "AggloNameEn",

"EU28", "NumberOfInhabitants", "Ld55", "Ld60", "Ld65", "Ld70",

"Ld75", "LdSum", "LdRprt", "Ld\_More100Pcnt", "Ln50", "Ln55", "Ln60",

"Ln65", "Ln70", "LnSum", "LnRprt", "Ln\_More100Pcnt")]

# t2

t2.cln <- t2 %>%

rename(Ld55 = nLden5559, Ld60 = nLden6064, Ld65 = nLden6569, Ld70 = nLden7074,

Ld75 = nLden75, Ln50 = nLnight5054, Ln55 = nLnight5559,

Ln60 = nLnight6064, Ln65 = nLnight6569, Ln70 = nLnight70) %>%

mutate(Src = source) %>% #create a column with the name of the source

mutate(LdSum = ifelse(Ld55 > -1, Ld55+Ld60+Ld65+Ld70+Ld75, NA)) %>%

mutate(LnSum = ifelse(Ln50 > -1, Ln50+Ln55+Ln60+Ln65+Ln70, NA)) %>%

mutate(LdRprt = case\_when(Ld55 < -1 ~ -2, #-2 & -9999 assigned -2

Ld55 == -1 ~ -1, #-1 assigned -1

Ld55 > -1 ~ 1)) %>%

mutate(LnRprt = case\_when(Ln50 < -1 ~ -2, #-2 & -9999 assigned -2

Ln50 == -1 ~ -1, #-1 assigned -1

Ln50 > -1 ~ 1)) %>%

mutate(Ld\_More100Pcnt = if\_else(LdSum > NumberOfInhabitants ,TRUE, FALSE)) %>%

mutate(Ln\_More100Pcnt = if\_else(LnSum > NumberOfInhabitants ,TRUE, FALSE))

# t2 Reorder columns, exclude extra info

t2.cln <- t2.cln[c("Src","Ctry","Ctry2", "ReferenceYear", "RLID", "AggloNameEn",

"EU28", "NumberOfInhabitants", "Ld55", "Ld60", "Ld65", "Ld70",

"Ld75", "LdSum", "LdRprt", "Ld\_More100Pcnt", "Ln50", "Ln55", "Ln60",

"Ln65", "Ln70", "LnSum", "LnRprt", "Ln\_More100Pcnt")]

```

## Output: countries with nr of aggl with reported data (pt=pivot table)

```{r pvitt\_reported, echo=TRUE}

t1.Ld.pt <- PivotTable$new() #Pivot table with counts of missing data

t1.Ld.pt$addData(t1.cln)

t1.Ld.pt$addColumnDataGroups("LdRprt")

t1.Ld.pt$addRowDataGroups("Ctry")

t1.Ld.pt$defineCalculation(calculationName="Total", summariseExpression="n()")

t1.Ld.pt$evaluatePivot()

t1.Ln.pt <- PivotTable$new() #Pivot table with counts of missing data

t1.Ln.pt$addData(t1.cln)

t1.Ln.pt$addColumnDataGroups("LnRprt")

t1.Ln.pt$addRowDataGroups("Ctry")

t1.Ln.pt$defineCalculation(calculationName="Total", summariseExpression="n()")

t1.Ln.pt$evaluatePivot()

t2.Ld.pt <- PivotTable$new() #Pivot table with counts of missing data

t2.Ld.pt$addData(t2.cln)

t2.Ld.pt$addColumnDataGroups("LdRprt")

t2.Ld.pt$addRowDataGroups("Ctry")

t2.Ld.pt$defineCalculation(calculationName="Total\_Ld", summariseExpression="n()")

t2.Ld.pt$evaluatePivot()

t2.Ln.pt <- PivotTable$new() #Pivot table with counts of missing data

t2.Ln.pt$addData(t2.cln)

t2.Ln.pt$addColumnDataGroups("LnRprt")

t2.Ln.pt$addRowDataGroups("Ctry")

t2.Ln.pt$defineCalculation(calculationName="Total\_Ln", summariseExpression="n()")

t2.Ln.pt$evaluatePivot()

t1.Ld.pt

t1.Ln.pt

t2.Ld.pt

t2.Ln.pt

```

## Output: agglomerations where Ldn | Ln > NrInhabitants

```{r MoreThan100, echo=TRUE}

t1.cln %>% filter(Ld\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LdSum) %>%

arrange(Ctry, AggloNameEn)

t1.cln %>% filter(Ln\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LnSum) %>%

arrange(Ctry, AggloNameEn)

t2.cln %>% filter(Ld\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LdSum) %>%

arrange(Ctry, AggloNameEn)

t2.cln %>% filter(Ln\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LnSum) %>%

arrange(Ctry, AggloNameEn)

```

# 2. Gap filling

There are three steps

\* Subset A. Data available for t2 (current phase)

\* Subset B. Data not availabel for t2, but available from t1

\* Subset C. None of the previous cases apply. Regression

## 2.1.A.Subset of available data for t2

Output: t2A.Ld, t2A.Ln

t2 = current phase

A = data reported

New column DataSrc = Indicate the source of the data (current, previous

reporting period, regression).

```{r subsetA}

t2A.Ld <- t2.cln %>%

filter(LdRprt == 1) %>% #data reported for t2

mutate(DataSrc = "t2") %>% #create column DataSrce DataSource

select(-17:-24) #exclude Ln

t2A.Ln <- t2.cln %>%

filter(LdRprt == 1) %>% #data reported for t2

mutate(DataSrc = "t2") %>% #create column DataSrce DataSource

select(-9:-16) #exclude Ld

#Agglomerations where People exposed > NrInhabitants

t2A.Ld %>% filter(Ld\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LdSum) %>%

arrange(Ctry, AggloNameEn)

t2A.Ln %>% filter(Ln\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LnSum) %>%

arrange(Ctry, AggloNameEn)

```

## 2.2.B.Agg without data for t2 and data available for t1: use data from t1 if available

Output: t2B.Ld, t2B.Ln

t2 = current phase

B = data from t1 (previous reporting cycle)

New column DataSrc = Indicate the source of the data (current, previous

reporting period, regression).

```{r subsetB, echo=TRUE}

#Lden

tmp1.Ld <- t1.cln %>% #Select from t1 reported data

filter(LdRprt == 1) %>%

select(-15, -17:-24) #Exclude columns from Ln, LdRprt

tmp2.Ld <- t2.cln %>%

filter(LdRprt == -2) %>% #select from t2 missing data

select("RLID", "LdRprt")

t2B.Ld <- inner\_join(tmp1.Ld, tmp2.Ld, by= "RLID") %>%

mutate(DataSrc = "t1") #create column DataSrce DataSource

#Lnight

tmp1.Ln <- t1.cln %>% #Select from t1 reported data

filter(LdRprt == 1) %>%

select(-9:-16, -23) #Exclude columns from Ld, LnRprt

tmp2.Ln <- t2.cln %>%

filter(LnRprt == -2) %>% #select from t2 missing data

select("RLID", "LnRprt")

t2B.Ln <- inner\_join(tmp1.Ln, tmp2.Ln, by= "RLID") %>%

mutate(DataSrc = "t1") #create column DataSrce DataSource

#Agglomerations where People exposed > NrInhabitants

t2B.Ld %>% filter(Ld\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LdSum) %>%

arrange(Ctry, AggloNameEn)

t2B.Ln %>% filter(Ln\_More100Pcnt == TRUE) %>%

select(Ctry, AggloNameEn, NumberOfInhabitants,LnSum) %>%

arrange(Ctry, AggloNameEn)

```

## 2.3 C.Regression

Data not available (t1 neither t2)

Regression

### 231 Outliers from % change population exposed

#### a.Join t1.cln2 & t2.cln2 (cln = clean)

```{r join}

# join

t1t2 <- inner\_join(t1.cln, t2.cln, by = "RLID", suffix = c(".t1", ".t2"))

# create difference and % of difference [100\*(t2-t1)/t1]

t1t2 <- t1t2 %>%

mutate(ld\_t2\_t1 = LdSum.t2 - LdSum.t1) %>%

mutate(ln\_t2\_t1 = LnSum.t2 - LnSum.t1) %>%

mutate(ld\_pcnt = 100 \* ld\_t2\_t1 / LdSum.t1) %>% #percentage of difference

mutate(ln\_pcnt = 100 \* ln\_t2\_t1 / LnSum.t1)

```

#### b.Outliers t2-t1 (%)

Create columns Outl\_Ld, Outl\_Ln (TREU, FALSE)

```{r outl}

# Statistics for Lden

tmpt1t2 <- t1t2 %>% #exclude SumLden > NInhabitants

filter(Ld\_More100Pcnt.t1 == FALSE & Ld\_More100Pcnt.t2 == FALSE) %>%

select(ld\_pcnt)

Ld.summary <- summary(tmpt1t2$ld\_pcnt)

iqrLd <- Ld.summary[[5]] - Ld.summary[[2]] # Estimate interquartile range (3rd - 1st)

lower\_bound\_Ld <- Ld.summary[[2]] - (1.5 \* iqrLd) # Identify bounds for outliers

upper\_bound\_Ld <- Ld.summary[[5]] + (1.5 \* iqrLd)

# create column outlier

t1t2 <- t1t2 %>%

mutate(Outl\_Ld = if\_else(ld\_pcnt > upper\_bound\_Ld | ld\_pcnt <lower\_bound\_Ld, TRUE, FALSE))

# Statistics for Lnight

tmpt1t2 <- t1t2 %>%

filter(Ln\_More100Pcnt.t1 == FALSE & Ln\_More100Pcnt.t2 == FALSE) %>%

select(ln\_pcnt)

Ln.summary <- summary(tmpt1t2$ln\_pcnt)

iqrLn <- Ln.summary[[5]] - Ln.summary[[2]] # Estimate interquartile range (3rd - 1st)

lower\_bound\_Ln <- Ln.summary[[2]] - (1.5 \* iqrLn) # Identify bounds for outliers

upper\_bound\_Ln <- Ln.summary[[5]] + (1.5 \* iqrLn)

# create column outlier

t1t2 <- t1t2 %>%

mutate(Outl\_Ln = if\_else(ln\_pcnt > upper\_bound\_Ln | ln\_pcnt < lower\_bound\_Ln, TRUE, FALSE))

# add the columns Outl\_Ld, Outl\_Ln to t2.cln

tmpt1t2 <- t1t2 %>% select("RLID", "Outl\_Ld", "Outl\_Ln") #select only relevant columns

t2.cln <- left\_join(t2.cln, tmpt1t2, by="RLID") #add columns to t2.cln

```

####c.Summary of statistics

```{r outl\_output, echo=TRUE}

Ld.summary

iqrLd

lower\_bound\_Ld

upper\_bound\_Ld

Ln.summary

iqrLn

lower\_bound\_Ln

upper\_bound\_Ln

```

####d.Visualization. Overlaid histograms

```{r histogram, evalu=TRUE}

tmpt1t2 <- t1t2 %>%

filter(Ld\_More100Pcnt.t1 == FALSE &

Ld\_More100Pcnt.t2 == FALSE ) #exclude Lden > NrInhabitants

ggplot(tmpt1t2, aes(x=ld\_pcnt, fill=Outl\_Ld)) +

geom\_histogram(binwidth=10, alpha=.5, position="identity") +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) + #change scientific notation

xlab("Lden (difference t2-t1 (%))") + #x label

labs(fill = "Outliers") #title of legend

tmpt1t2 <- t1t2 %>% filter(Ln\_More100Pcnt.t1 == FALSE &

Ln\_More100Pcnt.t2 == FALSE ) #exclude Ln > NrInhabitants

ggplot(tmpt1t2, aes(x=ln\_pcnt, fill=Outl\_Ln)) +

geom\_histogram(binwidth=10, alpha=.5, position="identity") +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) + #change scientific notation

xlab("Lnight (difference t2-t1) (%)") + #x label

labs(fill = "Outliers") #title of legend

```

### 232 Outliers t2 SumLden >= 55 dB, SumLn >= 50 dB

t2 exclude >100% exposed

t2 identify outliers Ld Ln

```{r OutlRgr}

#Lden Statistics

tmp2.Ld <- t2A.Ld %>% #exclude SumLden > NInhabitants

filter(Ld\_More100Pcnt == FALSE) %>%

select(LdSum)

Ld.summary <- summary(tmp2.Ld$LdSum)

iqrLd <- Ld.summary[[5]] - Ld.summary[[2]] # Estimate interquartile range (3rd - 1st)

lower\_bound\_Ld <- Ld.summary[[2]] - (1.5 \* iqrLd) # Identify bounds for outliers

upper\_bound\_Ld <- Ld.summary[[5]] + (1.5 \* iqrLd)

#Lden create column outlier

t2A.Ld <- t2A.Ld %>%

mutate(Outl\_Sum = if\_else(LdSum > upper\_bound\_Ld | LdSum <lower\_bound\_Ld, TRUE, FALSE))

#Lnight Statistics

tmpt2.Ln <- t2A.Ln %>%

filter(Ln\_More100Pcnt == FALSE) %>%

select(LnSum)

Ln.summary <- summary(tmpt2.Ln$LnSum)

iqrLn <- Ln.summary[[5]] - Ln.summary[[2]] # Estimate interquartile range (3rd - 1st)

lower\_bound\_Ln <- Ln.summary[[2]] - (1.5 \* iqrLn) # Identify bounds for outliers

upper\_bound\_Ln <- Ln.summary[[5]] + (1.5 \* iqrLn)

#Lnight create column outlier

t2A.Ln <- t2A.Ln %>%

mutate(Outl\_Sum = if\_else(LnSum > upper\_bound\_Ln | LnSum < lower\_bound\_Ln, TRUE, FALSE))

```

#### a.Output

```{r outl\_output\_Rgr, echo=TRUE}

Ld.summary

iqrLd

lower\_bound\_Ld

upper\_bound\_Ld

Ln.summary

iqrLn

lower\_bound\_Ln

upper\_bound\_Ln

```

#### b.Visualization. Overlaid histograms

```{r histogram\_Rgr, evalu=TRUE}

tmp2.Ld <- t2A.Ld %>%

filter(Ld\_More100Pcnt == FALSE) #exclude Lden > NrInhabitants

ggplot(tmp2.Ld, aes(x=LdSum, fill=Outl\_Sum)) +

geom\_histogram(binwidth=50000, alpha=.5, position="identity") +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) + #change scientific notation

xlab("Lden (People exposed Lden >= 55)") + #x label

labs(fill = "Outliers") #title of legend

tmp2.Ln <- t2A.Ln %>%

filter(Ln\_More100Pcnt == FALSE) #exclude Ln > NrInhabitants

ggplot(tmp2.Ln, aes(x=LnSum, fill=Outl\_Sum)) +

geom\_histogram(binwidth=50000, alpha=.5, position="identity") +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) + #change scientific notation

xlab("Lnight (People exposed Lnight >= 50)") + #x label

labs(fill = "Outliers") #title of legend

```

### 233 Regression Lden

plot with and without outliers

regression with outliers

iterate for each model

regression

statistics of regression

plot residuals

plot outcome of different model

estimate for missing data with chosen model

#### a.Plot Lden ~ Nr of inhabitants

look at the scatter plot to decide alternative models

```{r plotLdPop}

#Lden Plot with outliers

ggplot(data = t2A.Ld, mapping = aes(x = NumberOfInhabitants, y = LdSum)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Lden Plot without outliers

tmp2.Ld <- t2A.Ld %>% filter(Ld\_More100Pcnt == FALSE &

Outl\_Sum == FALSE)

ggplot(data = tmp2.Ld, mapping = aes(x = NumberOfInhabitants, y = LdSum)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

#### b.transformations

Transform LdenSum to log

```{r log\_trans}

#Transform if it is needed

t2A.Ld <- t2A.Ld %>%

mutate(log.LdSum = log(LdSum))

ggplot(data = t2A.Ld, mapping = aes(x = NumberOfInhabitants, y = log.LdSum)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

transform sqrt

```{r sqrt\_trans}

#Transform if it is needed

t2A.Ld <- t2A.Ld %>%

mutate(sqrt(LdSum))

ggplot(data = t2A.Ld, mapping = aes(x = NumberOfInhabitants, y = sqrt(LdSum))) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

Transform loglog

```{r loglog\_trans}

#Transform if it is needed

t2A.Ld <- t2A.Ld %>%

mutate(log(NumberOfInhabitants))

ggplot(data = t2A.Ld, mapping = aes(x = log(NumberOfInhabitants), y = log(LdSum))) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

#### c.Random selection of a subset for verification

Output: t2A.Ld.model to run the regression; t2A.Ld.validation to validate

```{r selectvalidation}

#Select 65% of all countries for regression, the rest for validation

#The selection is proportional to the distributin of agglomerations within ctries

split <- createDataPartition(t2A.Ld$Ctry, p = .65, list = F)

validation <- t2A.Ld[-split,]

t2A.Ld.validation <- left\_join(validation, t2A.Ld, )

t2A.Ld.model <- anti\_join(t2A.Ld, t2A.Ld.validation)

```

### 2331.Regression model1

Linear model y = a + b\*x

Data source: t2A.Ld.model

Look at

\* heterostadicity. Residuals vs fitted

\* Normal Q-Q. Follow the line

\* residuals within leverage (ouside = strong influence on refression)

```{r regression.model1}

#Calculate regression

Ld.Rgr1 <- lm(LdSum ~ NumberOfInhabitants, data=t2A.Ld.model)

#Assessing model quality

glance(Ld.Rgr1) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr1)

sigma(Ld.Rgr1)/mean(t2A.Ld.model$LdSum)

mean(t2A.Ld.model$LdSum)

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr1) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

##### Validation model1

```{r validation.model1}

#Predict with t2A.Ld.validation

predict <- predict(Ld.Rgr1, newdata = t2A.Ld.validation, interval = "confidence")

Ld.Predict.1 <- cbind(t2A.Ld.validation, predict)

#Plot fitted vs data

ggplot(data = Ld.Predict.1, mapping = aes(x = LdSum, y = fit)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Regression fitted vs original data

Ld.Rgr1.predict <- lm(fit ~ LdSum, data=Ld.Predict.1)

#Assessing model quality

glance(Ld.Rgr1.predict) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr1.predict)

sigma(Ld.Rgr1.predict)/mean(t2A.Ld.validation$LdSum)

mean(t2A.Ld.validation$LdSum)

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr1.predict) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#calculate confidence interval of the sum

# sqrt[ sum(confidence\_interval\_individual^2) ]

Ld.Predict.1 <- Ld.Predict.1 %>%

mutate(CI = (upr - lwr)/2) %>%

mutate(sqr.CI = CI^2)

meanCI = mean(Ld.Predict.1$CI)

CI = sqrt(sum(Ld.Predict.1$sqr.CI))

#Compare total values reported vs estimated + CI

sum(Ld.Predict.1$LdSum)

sum(Ld.Predict.1$fit)

CI

meanCI

```

### 2332.Regression model2

Polynomial

Linear model y = a + b\*x + c\*x^2

Data source: t2A.Ld.model

Look at

\* heterostadicity. Residuals vs fitted

\* Normal Q-Q. Follow the line

\* residuals within leverage (ouside = strong influence on refression)

```{r rgr.model2}

#Calculate regression

Ld.Rgr2 = lm(LdSum ~ poly(NumberOfInhabitants, 2, raw=TRUE), data = t2A.Ld.model)

#Assessing model quality

glance(Ld.Rgr2) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr2)

sigma(Ld.Rgr2)/mean(t2A.Ld.model$LdSum)

mean(t2A.Ld.model$LdSum)

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr2) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

##### Validation model2

```{r validation.model2}

#Predict with t2A.Ld.validation

predict <- predict(Ld.Rgr2, newdata = t2A.Ld.validation, interval = "confidence")

Ld.Predict.2 <- cbind(t2A.Ld.validation, predict)

#Plot fitted vs data

ggplot(data = Ld.Predict.2, mapping = aes(x = LdSum, y = fit)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Regression fitted vs original data

Ld.Rgr2.predict <- lm(fit ~ LdSum, data=Ld.Predict.2)

#Assessing model quality

glance(Ld.Rgr2.predict) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr1.predict)

sigma(Ld.Rgr2.predict)/mean(t2A.Ld.validation$LdSum)

mean(t2A.Ld.validation$LdSum)

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr2.predict) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#calculate confidence interval of the sum

# sqrt[ sum(confidence\_interval\_individual^2) ]

Ld.Predict.2 <- Ld.Predict.2 %>%

mutate(CI = (upr - lwr)/2) %>%

mutate(sqr.CI = CI^2)

meanCI = mean(Ld.Predict.2$CI)

CI = sqrt(sum(Ld.Predict.2$sqr.CI))

#Compare total values reported vs estimated

sum(Ld.Predict.2$LdSum)

sum(Ld.Predict.2$fit)

CI

meanCI

```

### 2333.Regression model3

log-log

Linear model log(y) = a + b\*log(x)

Data source: t2A.Ld.model

Look at

\* heterostadicity. Residuals vs fitted

\* Normal Q-Q. Follow the line

\* residuals within leverage (ouside = strong influence on refression)

```{r rgr.model3}

#Calculate regression

Ld.Rgr3 = lm(log(LdSum) ~ log(NumberOfInhabitants), data = t2A.Ld.model)

#Assessing model quality

glance(Ld.Rgr3) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr3)

exp(sigma(Ld.Rgr3))/log(mean(t2A.Ld.model$LdSum))

log(mean(t2A.Ld.model$LdSum))

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr3) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

##### Validation model3

```{r validation.model3}

#Predict with t2A.Ld.validation

predict <- (predict(Ld.Rgr3, newdata = t2A.Ld.validation, interval = "confidence"))

Ld.Predict.3 <- cbind(t2A.Ld.validation, predict)

Ld.Predict.3 <- Ld.Predict.3 %>%

mutate(fit\_tr = exp(fit)) %>% #exponential to transform log log regression

mutate(lwr\_tr = exp(lwr)) %>%

mutate(upr\_tr = exp(upr))

#Plot fitted vs data

ggplot(data = Ld.Predict.3, mapping = aes(x = LdSum, y = fit\_tr)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Regression fitted vs original data

Ld.Rgr3.predict <- lm(fit\_tr ~ LdSum, data=Ld.Predict.3)

#Assessing model quality

glance(Ld.Rgr3.predict) %>%

dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

summary(Ld.Rgr1.predict)

sigma(Ld.Rgr3.predict)/mean(t2A.Ld.validation$LdSum)

mean(t2A.Ld.validation$LdSum)

#par(mfrow = c(2, 2)) # Split the plotting panel into a 2 x 2 grid

plot(Ld.Rgr3.predict) + scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Compare total values reported vs estimated

sum(Ld.Predict.3$LdSum)

sum(Ld.Predict.3$fit\_tr)

```

#### e.Plot Lnight \_ NrOfInhabitants

```{r}

#Ln Plot with outliers

ggplot(data = t2A.Ln, mapping = aes(x = NumberOfInhabitants, y = LnSum)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

#Ln Plot without outliers

tmp2.Ln <- t2A.Ln %>% filter(Ln\_More100Pcnt == FALSE &

Outl\_Sum == FALSE)

ggplot(data = tmp2.Ln, mapping = aes(x = NumberOfInhabitants, y = LnSum)) +

geom\_point() +

scale\_x\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ",")) +

scale\_y\_continuous(labels = comma\_format(big.mark = ".",

decimal.mark = ","))

```

# Annex III. Reported and delineated contour for Lden 55 dB (major airports)

|  |
| --- |
| Berlin Schonefeld (EDDB) |
| Mapa  Descripción generada automáticamente |
| Budapest (LHBP) |
| Mapa  Descripción generada automáticamente |

|  |
| --- |
| Copenhagen (EKCH) |
| Mapa  Descripción generada automáticamente |
| Hamburg (EDDH) |
| Mapa  Descripción generada automáticamente |

|  |
| --- |
| Helsinki – Vantaa (EFHK) |
| Mapa  Descripción generada automáticamente |
| Lisbon (LPPT) |
| Mapa  Descripción generada automáticamente |

|  |
| --- |
| Milano Malpensa (LIMC) |
| Imagen que contiene árbol  Descripción generada automáticamente |
| Napoli (LIRN) |
| Mapa  Descripción generada automáticamente |

|  |
| --- |
| Vienna (LOWW) |
| Mapa  Descripción generada automáticamente |



The European Topic Centre on Air pollution, transport, noise and industrial pollution (ETC/ATNI) is a consortium of European institutes under a framework partnership contract to the European Environment Agency.

European Topic Centre on Air pollution,   
transport, noise and industrial pollution

c/o NILU – Norwegian Institute for Air Research

P.O. Box 100, NO-2027 Kjeller, Norway

Tel.: +47 63 89 80 00

Email: [etc.atni@nilu.no](mailto:etc.atni@nilu.no)

Web : https://www.eionet.europa.eu/etcs/etc-atni

1. <https://forum.eionet.europa.eu/etc-atni-consortium/library/subvention-2019/task-deliveries-action-plan-2019/task-1.1.5.1-noise-data-operational-compilation-and-management/subtask-1.1-update-database-cws/datasets> [↑](#footnote-ref-1)
2. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat [↑](#footnote-ref-2)