



DataMan



Guidelines for statistical analysis of the Dataman databases

Version 1.0

September 2020

Prepared by

Alasdair Noble, Tony van der Weerden, Christian Ammon, Barbara Amon, Nicholas Hutchings, Cecile de Klein, Anja Hansen, Ignacio Beltran, Federico Dragoni, Gültac Cinar, Latifa Ouatahar

Table of Contents

1. Purpose of guidelines	3
2. Dataman database structure	3
3. Types of questions being asked of the DataMan databases.....	4
4. Limitations of the Databases	4
5. Approach to data analysis	5
6. Zero, negative and missing observations	5
7. Means vs replicate data.....	6
8. Field database	7
<i>Calculation of emission factors</i>	<i>7</i>
<i>Variables.....</i>	<i>8</i>
<i>Transformations, distributions and heteroscedasticity</i>	<i>8</i>
9. Storage and Housing databases.....	10
<i>Calculation of emission rates and emission factors</i>	<i>10</i>
<i>Standardised unit for emission rates.....</i>	<i>10</i>
<i>Derivation of emission factors.....</i>	<i>10</i>
<i>Variables.....</i>	<i>11</i>
<i>Transformations, distributions and heteroscedasticity</i>	<i>11</i>
10. Reporting Field, Housing and Storage data analysis	11
11. Acknowledgements.....	12
12. References	12

1. Purpose of guidelines

The purpose of these guidelines is to assist researchers wishing to conduct statistical analysis of the Dataman databases. There are three databases, each relating to a different stage of the manure management system: animal **housing**, manure **storage** and **field**-based emissions. These databases will expand over time, with data currently being collated and included as part of the DataMan project (<https://dataman.azurewebsites.net/>) and the Mitigating Emissions from Agricultural Systems (MELS) project (<https://www.mels-project.eu/>). The DataMan project was created to build a publicly-free global database of methane (CH₄), nitrous oxide (N₂O) and ammonia (NH₃) emissions (plus relevant activity and ancillary data) relating to livestock housing, storage and manure application to land, including excreta deposited during grazing. These databases, representing housing, storage and field emissions, provide an opportunity to identify possible variables influencing gas emissions from the manure management system. The aims of the DataMan and MELS projects are to provide researchers and policy makers alike with the most up-to-date knowledge on methods for managing GHG and NH₃ emissions from manure. We will refer to the databases as the DataMan databases, given these were initiated within the DataMan project.

Please note that these are only guidelines and help to highlight some of the issues with the Dataman databases. All the analyses are carried out on subsets of the databases and any subset may have different challenges. Individual researchers may have alternative approaches that they can defend.

These guidelines may also be applied to other datasets and databases. But please note that your dataset structure, content and the purpose of the analysis may differ from that of the Dataman project.

2. Dataman database structure

Relevant information (emission factors, and biotic and abiotic factors) was collated from published peer-reviewed research, theses, conference papers and existing databases. Additional data will be entered during the MELS project. As noted above, there are three databases, each relating to a different stage of the manure management system: animal **housing**, manure **storage** and **field**-based emissions. The latter includes both manure applied to land and dung and urine deposited during livestock grazing. Please consult the Ramiran Glossary of terms on livestock and manure management (Pain and Menzi, 2011) for definitions relating to manures.

The **housing** database includes more than 150 variables which were grouped into six categories: "General", "Gas measurement", "Animal", "Housing", "Manure" and "Climate". Data exists for CH₄, N₂O, NH₃, CO₂, H₂S, VOC, NO_x, odour and other gaseous emissions.

A wide range of units are used in the Housing database – when first released there were more than 140 different gas emission units. A small number of data have included emission factors for N₂O and NH₃ (kg N emitted/ kg N excreted or stored); these have been either supplied directly within publications or calculated using auxiliary data from publications. In the case of CH₄, emission factors (kg CH₄/kg volatile solids, as per the emission factor unit of the Intergovernmental Panel on Climate Change (IPCC)) could often not be calculated due to insufficient information on volatile solids (VS).

The **storage** database includes 140 variables which were grouped into six categories: "General", "Gas measurement", "Animal", "Manure" and "Climate". Data exists for CH₄, N₂O, NH₃, CO₂, H₂S, and other gaseous emissions.

As for Housing, a wide range of units is used in the Storage database: more than 110 different gas emission units are included. The number of data that include emission factors for N₂O and NH₃ (kg N emitted/kg N excreted) and CH₄ (kg CH₄/kg VS excreted) is limited.

The **field** database includes 94 variables which were grouped into six categories: “General”, “Gas measurement”, “Animal”, “Manure”, “Land” and “Climate”. This database is dominated by N₂O and NH₃ studies, although there are several CH₄ studies included. For the purposes of the Dataman and MELS projects, we are limiting our analysis to N₂O and NH₃ emissions, where the IPCC and UNECE (United Nations Economic Commission for Europe) emission factor units are adopted (kg N emitted/kg N excreted).

The animal housing and manure storage database contain primarily treatment-level data (i.e. mean values from multiple replicates of the same treatment), as the majority of data was derived from published research. In contrast, the majority of data (*ca* 75%) contained in the field section was sourced as replicate-level data.

3. Types of questions being asked of the DataMan databases

Below is a list of the type of questions being addressed in our analysis.

For each gas/source combination we can investigate:

1. What animal and manure types can we and/or should we group?
2. How do we calculate revised emission factors for a given group?
3. What are the significant variables influencing these emission factors?
4. How do emission factors and significant variables differ by region or country?
5. Can we develop a predictive model for these emission factors, and how robust is this model?
6. Can we quantify the effectiveness of different mitigation strategies aiming at reducing emission factors (e.g. storage practices, manure treatment, method of land application, use of inhibitors).

4. Limitations of the Databases

There are several key limitations associated with the Dataman databases. The two most important relate to an imbalanced dataset and the influence of institution on the predictive models.

As the dataset is a collection of experiments, each with their own objectives, there is no overarching design so it is highly imbalanced. For example, the Field database is highly imbalanced with respect to Climate Zones, with more than 90% of the observations obtained from studies conducted in temperate wet climates. Therefore, care is required when making conclusions from the data analysis. In addition, the dataset has been collated from individual studies where variables that may be of interest for an analysis have not necessarily been measured. The small numbers of observations resulting from some analyses may mean that statistical significance cannot be achieved for variables that are very likely to be important based on theoretical or biological principles.

A second limitation is the influence of institute or experiment on the variance in emission factors. This was noted by Hafner et al (2018) when analysing a large NH₃ emission dataset for pig and cattle slurry applied to land, where it was noticed that the variable ‘institute’ (i.e. the research institute that conducted the measurements) had a large influence on the modelled emission factors. This was

thought to be due to experimental methodologies potentially being unique to individual institutes but also can be due to site variables such as soil and climatic parameters that have not been included in the databases. Some countries or regions may only be represented in the databases by a single institute and in those cases the 'institute' and any other effects may be confounded.

5. Approach to data analysis

Guidelines on statistical analysis of the data is divided into (1) Field and (2) Housing and Storage. Housing and Storage are grouped together because the characteristics of these two databases are similar (mainly treatment-level data, limited number of emission factors available). Field data is treated separately due to the dominance of replicate-level data, and the high proportion of emission factor data within the database.

Many statistical modelling approaches could be used for data from the database. These include, but are not limited to, simple linear models, generalised linear models, linear models with random effects, machine learning techniques such as random forests and generalised additive models. These could all be applied in a frequentist or Bayesian paradigm and weights could be applied to account for different numbers of replicates.

As a general approach, where there is doubt as to the best analysis to perform, all analysis options should be considered and the conclusions to be drawn from each analysis should be assessed. If these are consistent with each other, then there is no problem and the results of the most favoured analysis can be reported, but a sentence should be included detailing the alternatives and that they all gave similar results. If there are inconsistencies, then further work will be needed to understand why and to choose the better approach.

Two areas of caution are noted with respect to approaches to data analysis:

- 1) The Dataman databases contain a large amount of data, with many variables that can be investigated. This has the potential to become an exercise in looking for "significant variables" and with the number of variables available some will almost certainly be found. In statistical circles this is termed 'p-hacking' or 'fishing' and users of the databases should be wary of this possibility. Any significant variables should not be accepted at face value but should be considered in the context of where the data come from and any theoretical or biological basis for relationships found.
- 2) If the intention is to use the databases to confirm or extend previously published results there is a risk that the data from the original study is included in the data base and it is possible that, in working with a subset of the data, the only remaining observations are those from that study or they dominate the remaining data. This then simply confirms the original publication without strengthening it in any way.

6. Zero, negative and missing observations

Negative emission factors and cumulative emissions exist in the field database but are rare in the housing and storage databases. Negative EFs and cumulative emissions from soils can be real and therefore some adjustment of the data is often necessary before a log transformation can be performed. However, setting negative values to zero is not recommended, even if they are within

the minimum detection limit of the method of measurement, as this introduces a bias into any analysis carried out.

If they exist in the housing and storage databases, they may be a measurement artefact due to upwind gas concentrations being greater than those from the manure storage or housing facility being sampled. Normally, if such a situation occurs during measurements, the data should have been excluded from further analysis: this is an important step for emission measurements with natural ventilation.

If genuine negative emission factors do exist (i.e. not a result of a measurement artefact), some adjustment of the data is often necessary before a log transformation can be performed. However, it is not recommended to set negative values to zero, or to delete those observations, even if they are within the minimum detection limit of the method of measurement. The reason for not setting these values to zero is that this introduces a bias into the dataset.

Another consideration in relation to detection limits is how data may be entered into the database. Where data was below the detection limit (either positive or negative), it is possible that values were entered as '< detection limit'. In this case, these entries can be replaced by a value that is midpoint between the detection limit (or the lowest positive value) and zero to ensure the results are included in the analysis.

Zeros are rare as they require zero cumulative emissions or exactly the same control and treatment emissions which is unlikely. It is assumed the data has been collated correctly, but if you are suspicious of any 'zero' results, check the original source of the data e.g. published article.

If the frequency of zeros or negative values is low, it is unlikely to impact on the overall result. This can be tested by including and excluding zeros and/or negatives in your analysis to determine their importance.

Missing values may occur in both explanatory variables and the emission of interest. It may be worthwhile to impute values if there are only a few. Various techniques can be used but they all include strong assumptions and should be used with caution and after consulting statisticians.

7. Means vs replicate data

The majority of data (*ca* 75%) contained in the Dataman **field** database are individual replicate-level data, with the remaining 25% of the database containing treatment means from different studies. For treatment mean data, the majority of field NH₃ data is likely to be reported as arithmetic means rather than back transformed means. Most of the field N₂O data is likely to be reported as geometric (or back-transformed) means. Where log transformations are performed, the back transformed means are often not clearly reported in terms of whether a bias correction was made. Bias corrections are necessary to ensure the back transformed means reasonably reflect the true mean of the data on the original scale.

How do we combine replicate data with mean data? Do we pretend means are single replicate experiments and ignore the problem?

- For NH₃ data, we can convert replicate level data to arithmetic means by assuming the majority of mean data has been calculated by arithmetic means. Mean data should then be

weighted according to the number of replicates if the number of replicates vary over a wide range (e.g. 3 to 12).

- For N₂O data, this is a little more difficult, as most studies will have reported a back-transformed mean (with or without a bias correction). Here, we suggest carrying out alternative analyses and assessing the conclusions.

The method of data analysis is also strongly influenced by the objective of the analysis. If the aim is to compare one dataset with other, the best statistical practice (residual analysis and transformations/distribution/heteroscedasticity used appropriately) should be applied to each data set. However, for estimating emission factors for a given source of GHG e.g. all sheep urine deposited onto New Zealand pastures, arithmetic means are most likely best. Either way, the variance of a mean emission factor should be reported as confidence intervals and not standard errors: this also aligns with the IPCC default emission factors.

If a transformation is applied to correct for skewness in the data then the back transformed predictions will be biased. The amount of bias can be assessed by comparing the back transformed mean with the simple mean of the data. Where the differences are large we suggest discussions with a statistician.

The **housing** and **storage** databases contain primarily treatment means, especially as many housing studies are conducted on a single house i.e. number of replicates = 1. Where a study includes more than one replicate for a treatment, the means are most common presented as arithmetic means. Only a small number of studies have individual replicate-level values entered in the databases.

If you are working with a subset of the housing and storage databases and there are only a small number of studies with replicate data it is probably best to calculate means for the few studies with replicates and weight by the number of replicates if known.

8. Field database

Calculation of emission factors

Field-based N₂O emission factors are calculated using the following equation (de Klein et al. 2020):

$$EF_3 = \frac{\text{Manure N}_2\text{O} - \text{Control N}_2\text{O}}{\text{N load}} \times 100\% \quad (1)$$

Where, EF₃ is the emission factor of a manure N source (manure, dung or urine) (N₂O-N lost as % of N source applied); Manure N₂O is the cumulative N₂O loss (kg N₂O-N/ha) from the N source (manure, dung or urine); Control N₂O is the cumulative N₂O loss from the control treatment (kg N₂O-N/ha) and N load is the amount of N applied with the N source (manure, dung or urine in kg N/ha). Recent guidance on statistical considerations relating to N₂O emission factors for chamber-based measurements were published by the GRA (de Klein et al. 2020): these should be read together with this report for N₂O data analysis.

In contrast with N₂O emissions measures in closed chambers, NH₃ is typically measured using dynamic chambers, wind tunnels or micrometeorological methods. This means that the cumulative emission is not measured as such but estimated through integration of several measurement readings over time. These measurements, corrected for background NH₃ concentrations, are typically integrated over time to derive net cumulative emissions.

Variables

The Dataman database contains many explanatory variables. Some may be structural (e.g. animal category, manure type) while others may be explanatory (e.g. seasons, slope). The latter are often continuous variables that have been collapsed into categories.

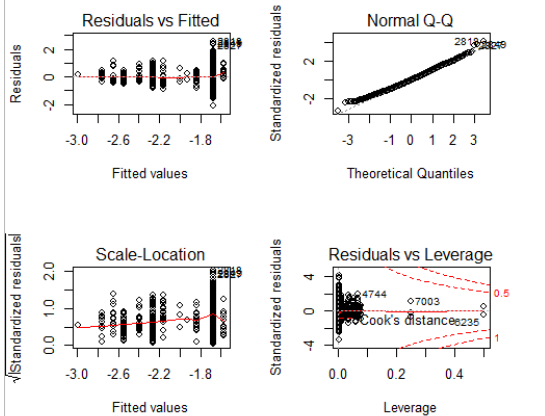
The fewer categories in explanatory variables, the better, because if interactions are included, a high number of parameters are generated. Therefore, aim to combine similar categories where possible. A useful rule of thumb is to have at least of ten observations per variable (or parameter in the model).

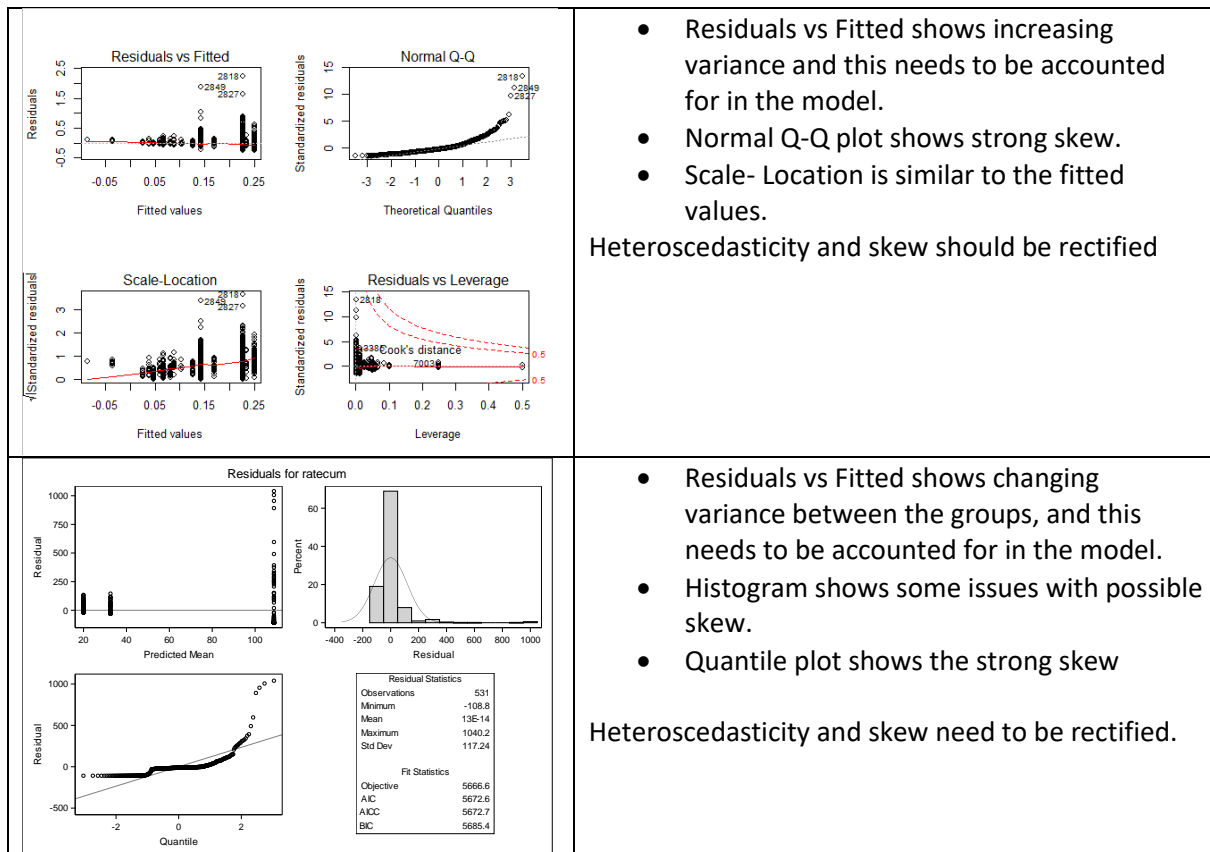
For analyses that are aimed at improving inventories, it is important to choose driving variables that can be obtained at the national scale. This may mean categorical variables (e.g. slurry tank or slurry lagoon) rather than continuous variables (e.g. surface area of storage) might be better, even if the latter would be more closely aligned with the processes driving emissions.

Transformations, distributions and heteroscedasticity

These terms are all closely related. Groups of data with unequal variance and strong skew are important issues that need to be resolved for statistical inference. Transformation of data to an approximately normal distribution may be a solution. Transformation options include log, square root, cube root and Box Cox. Another option is to adopt alternative distributions such as Poisson or negative binomial. In a Poisson distribution, only one value is reported that described both the mean and the shape of the distribution. Negative binomials are more flexible, because the Poisson distribution is constrained so the mean and standard deviation are equal but the negative binomial distribution relaxes this constraint.

If the residuals are consistent with the assumptions then there is no need to be overly concerned about the distribution of the data. This can be checked in collaboration with a person who has extensive experience in statistical data analysis. Below are some example plots with comments. If you are unsure, always seek statistical advice.

Residual Plot	Comment
 <p>The figure contains four diagnostic plots for a regression model. 1. Residuals vs Fitted: Shows residuals on the y-axis (ranging from -2 to 2) against fitted values on the x-axis (ranging from -3.0 to -1.8). There is a clear upward trend in the spread of residuals as fitted values increase, indicating heteroscedasticity. 2. Normal Q-Q: Shows standardized residuals on the y-axis (ranging from -2 to 2) against theoretical quantiles on the x-axis (ranging from -3 to 3). The points follow a straight line, suggesting the residuals are normally distributed. 3. Scale-Location: Shows the square root of the absolute value of standardized residuals on the y-axis (ranging from 0.0 to 2.0) against fitted values on the x-axis (ranging from -3.0 to -1.8). The spread increases with fitted values, similar to the Residuals vs Fitted plot. 4. Residuals vs Leverage: Shows standardized residuals on the y-axis (ranging from -4 to 4) against leverage on the x-axis (ranging from 0.0 to 0.4). Points 4744, 7003, and 235 are highlighted. A red dashed line indicates a Cook's distance of 0.5 for point 235.</p>	<ul style="list-style-type: none"> • Residuals vs Fitted shows possible increasing variance though this could be due to a factor that has not been accounted for, or a larger amount of data at that fitted value resulting in more variability. • Normal Q-Q plot is good. • Scale- Location is similar to the fitted values. • Leverage is getting a bit technical – talk to a statistician <p>Further investigation of larger variance at fitted value around -1.7 indicated but generally ok.</p>



- Residuals vs Fitted shows increasing variance and this needs to be accounted for in the model.
- Normal Q-Q plot shows strong skew.
- Scale- Location is similar to the fitted values.

Heteroscedasticity and skew should be rectified

- Residuals vs Fitted shows changing variance between the groups, and this needs to be accounted for in the model.
- Histogram shows some issues with possible skew.
- Quantile plot shows the strong skew

Heteroscedasticity and skew need to be rectified.

If the confidence intervals give non-sensible answers e.g. negative values, or values very close to zero, then the analysis needs to be reviewed.

The recommended procedure is as follows:-

- Graph the data
- Fit a model
- Look at the residuals
- Assess the model fit – how good is the prediction?
- Consider modelling subsets separately.

Heteroscedasticity refers to unequal variability in the data mostly due to different variability in different groups, for example “animal or manure types”. To properly assess the problem the model should be fitted and the residuals inspected. The solution may be modelling groups separately or applying a heteroscedastic model. It is probably safest to assume heteroscedasticity until you have shown otherwise.

Where possible, transformations should be considered as part of the scientific process and thus have some basis in the data collection process. A log transform can be considered to relate to percentage changes and other transformations may be justified in other ways.

9. Storage and Housing databases

Calculation of emission rates and emission factors

Most storage GHG and NH₃ experiments involve the use of laboratory- (< 500 L volume) and pilot-scale (> 500 L volume). Pilot-scale experiments would typically use experimental vessels located outdoors, with or without a shelter and submitted to ambient climatic conditions. Some experimenters take advantage of the presence of forced ventilation systems, and measure gas concentrations in the outlet stream of animal houses. Background gas concentrations sampled upstream are often included in measurement methods such as micrometeorological, wind tunnel or dynamic enclosure methods. The data are used to calculate emission factors for manure storage (N₂O and NH₃: kg N emitted/kg N excreted; CH₄: kg CH₄/kg VS excreted).

Housing experiments often include sampling of ambient gas concentrations outside and upwind of buildings. Data are often used to calculate emission rates rather than a cumulative emission, and rarely include emission factors (N₂O and NH₃: kg N emitted/kg N excreted; CH₄: kg CH₄/kg VS excreted).

Standardised unit for emission rates

As noted earlier, the housing and storage databases include a wide range of emission units. A statistical analysis of key drivers of emissions first requires the need to select a standard emission unit. Is there a “best” standard unit for emissions to analyse potential drivers?

It has been proposed that the suitable standard unit is:

Mass of gas / animal /time

For example, mg NH₃-N/cow/day

This unit also aligns with the IPCC methodology guidelines.

An alternative to ‘per animal’ is ‘per 500kg livestock unit (LU)’. The databases contain some emission data using this unit. Where available, data on animal age and/or weight can help with interpreting emissions associated with ‘per animal’ units, and conversion between ‘per animal’ and ‘per 500 kg LU’.

Derivation of emission factors

One of the key objectives of the Dataman project was to derive revised EF values for countries aiming to adopt alternative values to the IPCC default emission factors. One of the final activities of the Dataman project will be to submit revised EF values to the UNFCCC emission factor database. Associated with this is the need to understand the key drivers influencing the emission factors.

As such, it is necessary to convert as much of the housing and storage emission data to emission factors. A proportion of studies included in the database contain sufficient information to convert emission rates or cumulative emissions to emission factors, and thus this has been done where possible.

However, in many instances, there is insufficient information supplied for converting units to emission factors, resulting in the need to use alternative sources of additional data for this conversion. If using alternative data sources, there may be higher uncertainty in the resulting EF. Thus, it may be necessary to weight the data to account for potentially lower certainty, as performed with the analysis of survey data. If you are unsure how to weight the data, consult a statistician.

The alternative sources of data for deriving EF values will depend on the type of data required, but may include the following:

- Amount of excreta/animal/day
- N content of excreta/animal
- VS content of excreta/animal

Sources of data include scientific review articles (e.g. Hou et al. 2016), national GHG inventories, IPCC methodology guidelines and direct contact with animal nutrition scientists.

Since most of the emissions we consider here would be expected to vary throughout the year, according to temperature (and sometimes other variables), then this needs to be considered. For housing and storage, one option is to make measurements at periods throughout the year. Another is to relate the emissions to the climate variables and then use the latter to estimate the annual emissions.

Variables

See section 7 (**Field data**)

Transformations, distributions and heteroscedasticity

See section 7 (**Field data**)

10. Reporting Field, Housing and Storage data analysis

When reporting your results, include a description of any simple linear model used, either an ANOVA or regression model, with the response and explanatory variables detailed.

However, if a more complex analysis is carried out the following should be included where appropriate:

- What type of transformation was used?
- What did the distributions look like?
- What type of modelling was performed? E.g. linear model, fixed effect model, etc.
- Were subsets of the data used for the analysis?
- What software package was used? Provide a reference.

11. Acknowledgements

Funded by the New Zealand Government to support the objectives of the Livestock Research Group of the Global Research Alliance on Agricultural Greenhouse Gases and financially supported by the German Federal Ministry of Food and Agriculture (BMEL) through the Federal Office for Agriculture and Food (BLE), grant number 2819ERA10A (MELS project, funded under the Joint Call 2018 ERA-GAS, SusAn and ICT-AGRI on “Novel technologies, solutions and systems to reduce the greenhouse gas emissions in animal production systems”).

12. References

de Klein, C.A.M., Alfaro, M.A., Giltrap, D., Topp, C.F.E., Simon, P.L., Noble, A.D.L., van der Weerden, T.J., 2020. Global Research Alliance N₂O chamber methodology guidelines: Statistical considerations, emission factor calculation, and data reporting. *Journal of Environmental Quality*.

<https://doi.org/10.1002/jeq2.20127>

Hafner, S.D., Pacholski, A., Bittman, S., Burchill, W., Bussink, W., Chantigny, M., Carozzi, M., Géniermont, S., Häni, C., Hansen, M.N., Huijsmans, J., Hunt, D., Kupper, T., Lanigan, G., Loubet, B., Misselbrook, T., Meisinger, J.J., Neftel, A., Nyord, T., Pedersen, S.V., Sintermann, J., Thompson, R.B., Vermeulen, B., Vestergaard, A.V., Voylokov, P., Williams, J.R., Sommer, S.G., 2018. The ALFAM2 database on ammonia emission from field-applied manure: Description and illustrative analysis. *Agricultural and Forest Meteorology* 258, 66-79.

Hou, Y., Z. Bai, J. P. Lesschen, I. G. Staritsky, N. Sikirica, L. Ma, G. L. Velthof and O. Oenema . 2016. Feed use and nitrogen excretion of livestock in EU-27. *Agriculture, Ecosystems & Environment* 218: 232-244.

Pain, B and Menzi, H. 2011. Ramiran Glossary of terms on livestock and manure management. Ramiran. http://ramiran.uvlf.sk/doc11/RAMIRAN%20Glossary_2011.pdf