
THE USE OF DATA-DRIVEN TRANSFORMATIONS AND THEIR APPLICABILITY IN SMALL AREA ESTIMATION

Prof. Dr. Natalia Rojas-Perilla

↘ **Keywords:** *Data-driven transformations – small area estimation – poverty mapping – generalized regression models*

ABSTRACT

In general, researchers have been using data transformations as a go-to tool to assist scientific work under the classical and linear mixed regression models instead of developing new theories, applying complex methods or extending software functions. However, transformations are often automatically and routinely applied without considering different aspects on their utility. This work summarizes the main findings from the paper by the author (Rojas-Perilla, 2018), which presents a unified theory of data-driven transformations for linear and linear mixed regression models that includes applications to small area prediction and the development of open source software.

↘ **Schlüsselwörter:** datengetriebene Transformationen – Small-Area-Schätzung – Armutsabbildung – verallgemeinerte Regressionsmodelle

ZUSAMMENFASSUNG

Im Allgemeinen nutzen Forscherinnen und Forscher Daten-Transformationen als Hilfsmittel für die wissenschaftliche Arbeit unter den klassischen und linearen gemischten Regressionsmodellen, anstatt neue Theorien zu entwickeln, komplexe Methoden anzuwenden oder Softwarefunktionen zu erweitern. Allerdings werden Transformationen oft automatisch und routinemäßig angewendet, ohne verschiedene Nutzenaspekte zu betrachten. Diese Arbeit fasst die wichtigsten Erkenntnisse aus der Dissertation der Autorin (Rojas-Perilla, 2018) zusammen, die eine einheitliche Theorie der datengetriebenen Transformationen für lineare und lineare gemischte Regressionsmodelle vorstellt, die Anwendungen im Bereich Small Area Estimation (SAE)-Verfahren und die Entwicklung von Open-Source-Software enthält.



Prof. Dr. Natalia Rojas-Perilla

ist Juniorprofessorin für angewandte Statistik an der Freien Universität Berlin. Sie forscht zu den Schwerpunkten Small Area Verfahren, Gemischte Modelle, Räumliche Verfahren und Poverty Mapping. Für ihre Dissertation „The use of data-driven transformations and their applicability in small area estimation“ wurde sie mit dem Gerhard-Fürst-Preis 2020 des Statistischen Bundesamtes ausgezeichnet und stellt diese im vorliegenden Beitrag vor.

“Everything should be made as simple as possible, but not simpler.”

– Albert Einstein

1

Introduction

The literature of transformations in theoretical statistics and practical case studies in different research fields is rich and most relevant results were published during the early 1980s. More sophisticated and complex techniques and tools are available nowadays to the applied statistician as alternatives to using transformations. However, simplification is still a gold nugget in statistical practice, which is often the case when applying suitable transformations within the working model.

One of the goals of data analysts is to establish relationships between variables using regression models. Standard statistical techniques for linear and linear mixed regression models are commonly associated with interpretation, estimation, and inference. These techniques rely on basic assumptions underlying the working model, listed below:

- › Normality: Transforming data to create symmetry in order to correctly use interpretation and inferential techniques
- › Homoscedasticity: Creating equality of spread as a means to gain efficiency in estimation processes and to properly use inference processes
- › Linearity: Linearizing relationships in an effort to avoid misleading conclusions for estimation and inference techniques

Different options are available to the data analyst when the model assumptions are not met in practice. Researchers could formulate the regression model under alternative and more flexible parametric assumptions. They could also use a regression model that minimizes the use of parametric assumptions or under robust estimation. Another option would be to parsimoniously redesign the model by finding an appropriate transformation such that the model assumptions hold. In general, researchers have been using data transformations as a go-to tool to assist scientific work under the

classical and linear mixed regression models instead of developing new theories, applying complex methods or extending software functions. Nevertheless, transformations are often automatically and routinely applied without considering different aspects on their utility. For instance, a standard practice in applied work is to transform the target variable by computing its logarithm. However, this type of transformation does not adjust to the underlying data. Therefore, some research effort has been shifted towards alternative data-driven transformations, which includes a transformation parameter that adjusts to the data. The main contributions of this thesis focus on providing modeling guidelines for practitioners on transformations and on the methodological and practical development of the use of transformations in the context of small area estimation. The proposed approaches are complemented by the development of open source software packages. This aims to close possible gaps between theory and practice. This paper is structured into three parts. In part I (section 2), some modeling guidelines for data analysts in the context of data-driven transformations are presented. This summarizes the papers by Medina and others (2019) and Medina (2017). In part II (section 3), transformations in the context of small area estimation are applied and further developed. This is based on the papers by Rojas-Perilla and others (2020), Kreutzmann and others (2019) and Tzavidis and others (2018). Finally, part III (section 4) presents a discussion of the applicability of transformations in the context of generalized linear models. The publications listed below are the result of this overview.

2

Modeling Guidelines for Practitioners on Transformations

Representing a relationship between a response variable and a set of covariates is an essential part of the statistical analysis. The linear regression model offers a parsimonious solution to this issue, and hence it is extensively used in nearly all science disciplines. In recent years the linear mixed regression model has become common place in the statistical analysis. Standard statistical techniques for linear and linear mixed regression models are commonly associated with

interpretation, estimation, and inference. Numerous assumptions underlying the working model are usually made whenever these models are employed in scientific research. If one or several of these assumptions are not met, the application of transformations can be useful. The work provides an extensive overview of different transformations and estimation methods of transformation parameters in the context of linear and linear mixed regression models. The main contribution is the development of a guideline that leads the practitioner working with data that does not meet model assumptions by using transformations. The referenced work proposes a framework that seeks to help the researcher to decide if and how a transformation should be applied in practice. It combines a set of pertinent steps, tables, and flowcharts that guide the practitioner through the analysis of transformations in a friendly and practical manner. The guideline is structured as follows:

1. Defining relevant assumptions depending on the research goals: Choose the model and be aware of the corresponding assumptions.
2. Choosing a suitable transformation that addresses assumption violations and estimation method according to model assumption violations:
 - › Transformations to achieve normality: The use of transformations is considered as a parsimonious alternative to complex methodologies when dealing with the departure from normality, a feature seldom observed in raw data. A significant part of the effort put into transformations has been focused on achieving approximate normally distributed errors. To ensure normality, it is common to use a proper one-to-one transformation on the target variable (Hoyle, 1973; Thoni, 1969). To find a data-driven transformation, an adjustment is done by including a data-driven transformation parameter, denoted by λ . This parameter should be estimated and this estimate changes according to the data, the assumption violations or to a specific researcher criteria.
 - › Transformations to achieve homoscedasticity: According to and based on Bartlett (1947) and Bartlett (1937), transformations might provide a fair correction for heteroscedasticity. When a functional dependence of the variance of the outcome variable on the mean is present in the data, we may gain

the advantages of using variance-stabilizing transformations. This dependence mostly implies an underlying distributional process and determines the form of the suitable transformation.

- › Transformations to achieve linearity and additivity: In general, transformations to linearize data can be divided into two classes: in one class, the expected response is related to the independent variables by a known non-linear function; in the other, the relationship between the expected response and the explanatory variables is not exactly known (Cook/Weisberg, 1982). For the first class, transformations can be easily selected. Wood and Gorman (1971) show plots for a comprehensive number of non-linear functions that can be transformed into linear ones. In the second class fall transformations such as the Box-Cox transformation, which have the potential to correct non-normality, heteroscedasticity, and non-linearity, so that, after the data is transformed, normal theory methods and linear regression techniques can be employed.
3. Providing a proper inference analysis and interpreting model results more carefully: The inference analysis is a controversial question that arises when a transformation, and especially a transformation with a transformation parameter, is used under the linear and linear mixed regression model. One question is whether we should treat the transformation parameters as fixed in case we are making inferences on the model parameters. If the transformation does not contain a data-driven transformation parameter common model inference can be conducted. In contrast, when using data-driven transformations, one point of discussion concerns if the transformation parameter can be treated as known or not. One of the biggest challenges that researchers face when working with transformations is the interpretation of the results. It implies choosing the scale in which we need to present the results, depending on the research question. O'Hara and Kotze (2010) summarized this issue by pointing out that transformations comes at some cost to the trade-off between accuracy and interpretability.

In order to provide an extensive collection of transformations usable in linear regression models and a wide range of estimation methods for the transformation parameter, the package *trafo* is developed and presented as a part

of this work. This package complements and enlarges the methods that exist in R so far, and offers a simple, user-friendly framework for selecting a suitable transformation depending on the research purpose.

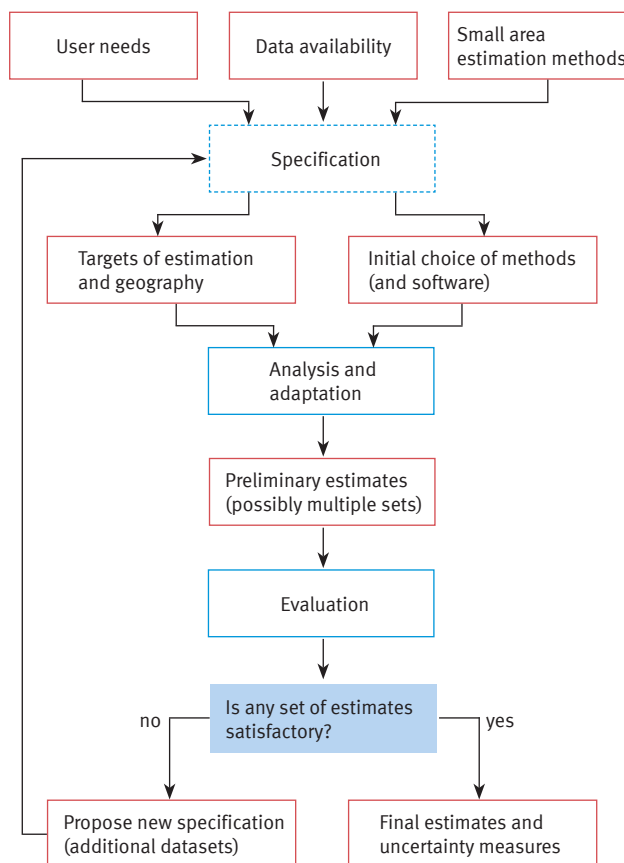
3

Transformations in the Context of Small Area Estimation

In the literature, little attention has been paid to the study of techniques of the linear mixed regression model when particularly working with data-driven transformations. This becomes a special challenge for users of small area estimation (SAE) methods, since most commonly used SAE methods are based on the linear mixed regression model which often relies on Gaussian assumptions. In particular, the empirical best predictor is widely used in practice to produce reliable estimates of general indicators for areas with small sample sizes. The issue of data transformations is addressed in the current SAE literature in a fairly ad-hoc manner. Contrary to standard practice in applied work, recent empirical work indicates that using transformations in SAE is not as simple as transforming the target variable by computing its logarithm. The main contributions of the work are particularly presented in this second part, where transformations in the context of SAE are applied and further developed. The study of SAE methods is a research area in official and survey statistics of great practical relevance for national statistical institutes and related organisations. Despite rapid developments in methodology and software, researchers and users would benefit from having practical guidelines for the process of small area estimation. In the work a general framework for the production of small area statistics that is governed by the principle of parsimony is proposed. This protocol is based on three stages, namely (i) specification, (ii) analysis/adaptation and (iii) evaluation. ↘ Figure 1

Emphasis is given to the interaction between a user of small area statistics and the statistician in specifying the target geography and parameters in light of the available data. Model-free and model-dependent methods are described with focus on model selection and testing, model diagnostics and adaptations such as use of data transformations. In particular, the use of some adap-

Figure 1
Graphical representation of the framework for the production of Small Area statistics¹



¹ Based on Tzavidis and others (2018).

tations of the working model by using transformations is shown as a part of the (ii) stage. Additionally, the use of data-driven transformations under linear mixed model-based SAE methods is extended, in particular, the estimation method of the transformation parameter under maximum likelihood theory. First, an analysis is conducted about how the performance of SAE methods are affected by departures from normality and how such transformations can assist with improving the validity of the model assumptions and the precision of small area prediction. In particular, attention has been paid to the estimation of poverty and inequality indicators, due to its important socio-economical relevance and political impact. Second, an adaptation of the mean squared error estimator is proposed to account for the additional uncertainty due to the estimation of transformation parameters. These methodological developments are

illustrated using real survey and census data for estimating income deprivation parameters for municipalities in Mexico. Finally, in order to improve some features of existing software packages suitable for the estimation of indicators for small areas, the package **emdi** is developed in this thesis. This package offers a methodological and computational framework for the estimation of regionally disaggregated indicators using SAE methods as well as providing tools for assessing, processing, and presenting the results. In particular, package **emdi** offers the following features:

- › It simplifies the estimation of indicators for small areas and its precision estimates by tailored functions.
- › These functions return by default estimates for a set of predefined indicators, including the mean, the quantiles of the distribution of the response variable and poverty and inequality indicators.
- › Self-defined indicators or indicators available within other packages can be included.
- › The users can choose the type of data transformations, including data-driven transformations, for which the parameters are automatically estimated.
- › It includes two bootstrap methodologies: a parametric bootstrap and a semi-parametric wild bootstrap for the for mean squared error (MSE) estimation.
- › Parallel computing is provided in a customized manner for reducing the computational time associated with the use of bootstrap.
- › It provides predefined functions for diagnostic checks of the underlying model, if model-based estimation is chosen. A mapping tool enables the creation of high quality maps. An informative output summarizing the most relevant results can be exported to Microsoft Excel.

4


Discussion on the Applicability of Transformations

We see and interpret the world as a set of discrete individual things that can be grouped: dogs, trees, countries and, thus, the act of counting is, usually, natural to all of

us: two dogs, five trees, ten countries, among others. In statistics, these variables are known as counts and refer to enumerated events or observations often confined within a fixed time-interval or a defined area. Sometimes, one also may like to analyze variables that take only values within the interval $[0, 1]$, such as proportions or percentages: for instance, the proportion of animals inhabiting a specific area. Thus, if the aim is to model these non-continuous variables, linear regression may not be able to be directly used. In fact, it makes different key assumptions about the target variable, the explanatory variables, and their relationship. First, it is based on modeling the expected value of measurements from a continuous quantity (such as weights or income) as a linear function of quantitative and qualitative covariates. This is also called the linearity assumption. Second, the variability is attached by the normal distribution of the error regression terms (normality assumption), which are also assumed to be independent with constant variance (homoscedasticity assumption).

If one aims to explain non-continuous variables using the classical linear regression model, a non-normal distributed error and heterogeneous variance structures arise and the above mentioned assumptions are not fulfilled. Typically, the conditional distribution of these data types can be skewed, their variances can be dependent on the mean, and they often contain many zero values (Blom, 1954). Even counts are easy to interpret: difficulties in the distribution of the observed variable can arise when the target variable is also bounded. Thus, directly using linear regression might yield inaccurate results and, moreover, might yield predictions for the target variable that lie outside the data range. Therefore, possible modifications in the response variable may be needed in order to apply the least squares estimation method and subsequent inference for the classical linear regression model. These modifications are known in the literature as transformations, and are broadly applied in this context in order to improve linearity, normality, and homoscedasticity assumptions (Rocke, 1993). Proper transformations for non-continuous data often depend on the underlying assumed distribution of the target variable or on the variance structure inherent to the data. But even if no evidence of a model-specific process underlying the data is taken into account or cannot be demonstrated, transformations can still be applied. The most prominent ones are the logarithmic function, the Box-Cox transformation, and different powers of

roots, among others. A broad range of models suitable for the analysis of non-continuous data have emerged as an alternative approach. For instance, generalized linear models (GLMs) were proposed by Nelder and Wedderburn (1972) and extended by McCullagh and Nelder (1989). These models allow for directly modeling a target variable coming from the family of exponential distributions that includes in particular the Poisson, binomial, and negative binomial distributions. GLMs are broadly applied in a wide variety of disciplines, such as human biology, ecology, and social sciences. They are specified by a linear predictor; a link function, which describes how the mean of the target variable is related to the linear predictor; and a variance function, which describes the relationship between the variance and the mean. Furthermore, generalized linear mixed models (GLMMs) additionally account for dependency coming from repeated measurements made on the same statistical units. Therefore, the non-continuous variables mentioned above could be modeled by using GLMs and GLMMs.

Choosing which methodology should be preferable, always depends on the research question. Using non-linear transformations for count data sets have different challenges for researches. First, the selection of a suitable transformation should be part of a previous careful analysis of the data to be studied. The distributional form of the underlying distributional process, the data range, and some features of distributional moments are some of the characteristics to be included in this previous analysis. For instance, in case the underlying process of the data is not previously known, data transformations are able to adapt on different count data distributions. In such a scenario, where the exact distribution of the target variable was applied in the context of GLMs, the use of GLMs are usually recommended in practice. Second, selecting only one transformation that improves all distributional assumptions of the linear regression model is not always straightforward. Thus, it is not common to have in practice one transformation, which in parallel corrects the model assumptions in the same way. Therefore, the research should know in which scale is the analysis made or the criteria of selecting one suitable transformation. Third, if a selected transformation is applied on a target variable and the researcher needs to return to the original measurement scale, a bias correction analysis should be proposed. 

BIBLIOGRAPHY

Bartlett, Maurice Stevenson. *Properties of sufficiency and statistical tests*. In: Proceedings of the Royal Society of London, Series A. Volume 160 (1937). Issue 901.

Bartlett, Maurice Stevenson. *The use of transformations*. In: Biometrics. Volume 3 (1947). No. 1, pp. 39

Blom, Gunnar. *Transformations of the binomial, negative binomial, Poisson and X^2 distributions*. In: Biometrika. Volume 41 (1954). No. 3/4, pp. 302

Cook, R. Dennis/Weisberg, Sanford. *Residuals and influence in regression*. New York 1982.

Cuthbert, Daniel/Wood, Fred S./Gorman, John W. *Fitting equations to data: Computer analysis of multifactor data for scientists and engineers*. New York 1971.

Hoyle, M. H. *Transformations: An introduction and a bibliography*. In: International Statistical Review / Revue Internationale de Statistique. Volume 41. No. 2/1973, pp. 203

Kreutzmann, Ann-Kristin/Pannier, Sören/Rojas-Perilla, Natalia/Schmid, Timo/Templ, Matthias/Tzavidis, Nikos. *The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators*. In: Journal of Statistical Software. Volume 91 (2019). Issue 7.

McCullagh, Peter/Nelder, John A. *Generalized linear models (2nd Edition)*. Boca Raton 1989.

Medina, Lily. *Transformations in the linear regression model: An overview*. Master's thesis. Humboldt Universität zu Berlin 2017.

Medina, Lily/Kreutzmann, Ann-Kristin/Rojas-Perilla, Natalia/Castro, Piedad. *The R Package trafo for Transforming Linear Regression Models*. In: The R Journal. Volume 11. Issue 2. December 2019. [retrieved on 4 January 2021]. Available at: <https://journal.r-project.org>

Nelder, John A./Wedderburn, Robert W. M. *Generalized linear models*. In: Journal of the Royal Statistical Society: Series A (Statistics in Society). Volume 135 (1972). Issue 3, pp. 370

O'Hara, Robert B./Kotze, D. Johan. *Do not log-transform count data*. In: Methods in Ecology and Evolution. Volume 1 (2010). Issue 2, pp. 118

Rocke, David M. *On the beta transformation family*. In: Technometrics. Volume 35 (1993). Issue 1, pp. 72

Rojas-Perilla, Natalia. *The Use of Data-driven Transformations and Their Applicability in Small Area Estimation*. Berlin 2018. [retrieved on 4 January 2021]. Available at: <https://refubium.fu-berlin.de>

BIBLIOGRAPHY

Rojas-Perilla, Natalia/Pannier, Sören/Schmid, Timo/Tzavidis, Nikos. *Data-Driven Transformations in Small Area Estimation*. In: Journal of the Royal Statistical Society: Series A (Statistics in Society). Volume 183 (2020). Part 1, pp. 121

Thoni, H. *Transformation of variables used in the analysis of experimental and observational data: A review*. Statistical Laboratory. Iowa State University 1969.

Tzavidis, Nikos/Zhang, Li-Chun/Luna, Angela/Schmid, Timo/Rojas-Perilla, Natalia. *From Start to Finish: A Framework for the Production of Small Area Official Statistics*. In: Journal of the Royal Statistical Society: Series A (Statistics in Society). Volume 181 (2018). Part 4, pp. 927

Herausgeber

Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung

Dr. Daniel Vorgrimler

Redaktionsleitung: Juliane Gude

Redaktion: Ellen Römer

Ihr Kontakt zu uns

www.destatis.de/kontakt

Erscheinungsfolge

zweimonatlich, erschienen im Februar 2021

Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Artikelnummer: 1010200-21001-4, ISSN 1619-2907

© Statistisches Bundesamt (Destatis), 2021

Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.