LCA data conforms to Benford's law

Bogdan Šinik
UP FAMNIT
Koper, Slovenia
bogdan.sinik@famnit.upr.si

Aleksandar Tošić
UP FAMNIT
InnoRenew CoE
Koper, Slovenia
aleksandar.tosic@upr.si

***Abstract*—Life cycle assessment (LCA) has been established as the standard method for evaluating environmental impacts of products, and processes. However, ISO standards depicting the application of these standards are tailored more towards LCA practitioners and less on the data acquisition and quality. The data acquisition process is not very robust, and considerable decision making and quality control is entrusted to the practitioners. Taking the lack of incentives from industry participants to submit quality data, the integrity of LCA databases can be questionable. Moreover, in some cases, participants may be incentivised to protect their data from competition. To address these concerns, data is carefully studied by external experts to verify their credibility. However, it is not entirely clear how these experts are chosen, nor how trust is established. In this paper, we apply a well known method Law of anomalous numbers, commonly refereed to as Benford's law in order to test the conformity of commonly used LCA databases. Our results on testing Ecoinvent, one of the most widely used LCA databases, show that LCA data strongly conforms to Benford's law. Moreover, our analysis includes 5 additional publicly available LCA databases, which also conformed with the exception of Bioenergiedat, which is likely due to the low number of observations. Finally, we tested individual properties given by Ecoinvent and establish that very few columns ($<5\%$), which pass the criteria for Benford's analysis are non-conforming. Although interesting, these results call for a more fine-grained analysis as future work.**

***Index Terms*—LCA, Benford's Law, Anomaly Detection**

I. Introduction

Life Cycle Assessment (LCA) is a systematic methodology for evaluating the environmental impacts of a product, process, or service throughout its entire life cycle. This encompasses all stages from raw material extraction, production, use, and disposal or recycling. The main goal of LCA is to identify opportunities for improving environmental performance and making informed decisions regarding sustainability.

1) Goal and Scope Definition**:** Establishing the purpose, boundaries, and scope of the assessment.
2) Inventory Analysis (LCI): Compiling an inventory of relevant energy and material inputs and environmental releases.
3) Impact Assessment (LCIA): Evaluating the potential environmental impacts associated with the inputs and releases identified in the inventory analysis.
4) Interpretation: Analyzing results to make informed decisions, identify significant issues, and suggest improvements.

Life Cycle Assessment (LCA) is used in various industries and sectors to promote sustainable practices, reduce environmental footprints, and support regulatory compliance. It is recognized as a crucial tool for environmental management and policy-making [1]. Considerable effort has been made over the past decade in an attempt to alleviate early criticism about data quality and trustworthiness [2]. Over the last few years many different tools have been made to asses these problems, but there is still not standard that could be used on all LCA datasets [3]. It was shown that results of LCA for same processes were variable over time which makes it hard to find statistical tool that could easily assess the trustworthiness and reliability of the data. [4] This is the reason why we decided to use Benford's law as a good indicator. It is also known as first-digit law and is very often used for fraud detection [5]. It was perfect choice for our research since it did not require a lot of domain specific knowledge and is not influenced by variability over time. For this purpose we have decided to analyze Ecoinvent database as it has established it self as one of the most used databases due to the amount of data, and tooling provided. More details about databases available in the literature [6, 7, 8].

II. Literature review

We have checked many sources that motivated us to check how good Benford's law would be for detecting inconsistencies in data. It is estimated that less than 20% of papers related to LCA conducted any kind of uncertainty analysis. [9]

Research conducted by Aalto University, Finland [10], compared numbers from five different LCA databases to compare amount of green house gases produced during the creation of the buildings. The findings indicate that the databases exhibit comparable patterns in the evaluation outcomes, with consistent disparities in scale across the reference buildings observed across all databases. Additionally, it was disclosed that there are significant disparities in the numerical values across the databases at some locations, and these disparities come from several data fragments.

Early research offered a comparative analysis of 26 sulfate pulp mills in Sweden, focusing on the quantified emissions released into the air and water between 1986 and 1993. The analysis revealed significant variations in annual emission variables across a group of companies. The emission parameters for most water emissions were not influenced by the annual production rates of pulp. [4]

A more recent study statistical analysis of elemental flows in commonly used LCA and LCIA databases and software was analysed [11]. The main conclusions of the study signaled considerable shortcomings in flow clarity, consistency, and extensibility in elementary flows. However, no common method of identification was proposed.

Moreover, the objectiveness of the evaluators responsible for data integrity has been highlighted as crucial to LCA database quality [12]. While this may be self-evident, a bigger concern would be when objectiveness is not guaranteed.

III. Methodology

The first-digit law is an observation about the frequency distribution of leading digits. It is also known as the Newcomb–Benford law or Benford's law. It has been apparently first discovered by polymath Newcomb and published in [13] and later rediscovered by physicist F. Benford and presented in [14]. The Benford's law [15] defines a fixed probability distribution for leading digits of any kind of numeric data with the following requirements:

- Data with values that are formed through a mathematical combination of numbers from several distributions.
- Data that has a wide variety in the number of figures (e.g., data with plenty of values in the hundreds, thousands, tens of thousands, etc.)
- Data set is fairly large, as a rule of a thumb at least 50 – 100 observations [16].
- Data is right skewed (i.e., the mean is greater than the median), and the distribution has a long right-tail rather than being symmetric.
- Data has no predefined maximum or minimum value (with the exception of a zero minimum).

The distribution of digits is presented in Figure 1; the digit 1 occurs in roughly 30 % of the cases, and the other digits follow in a logarithmic curve. It has been shown that this result applies to a wide variety of data sets [15], including electricity bills, street addresses, stock prices, house prices, population numbers, death rates, and lengths of rivers. The equation for the distribution of the first digits of observed data is given in Equation 1.

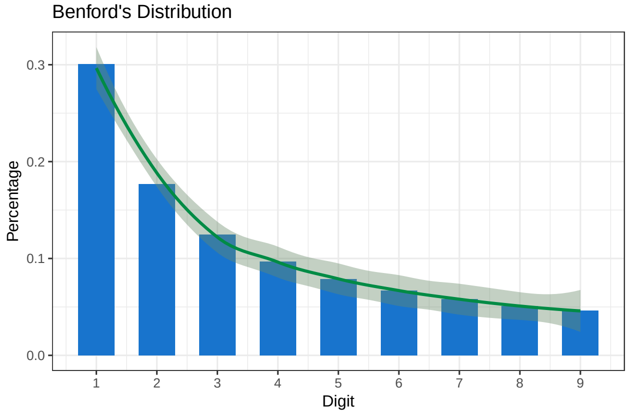

Fig. 1: A graphical representation of the Benford's distribution

$$P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1+\frac{1}{d}) \qquad (1)$$

We utilized the benford.analysis package [17] in R for our research. There were no instances of missing values denoted by "NA" in the form, although there were numerous occurrences of zeros. The Benford function automatically disregarded any zero values, eliminating the necessity for their removal. We needed to extract a subset of the data that includes only columns containing numerical values. The Benford function was modified to accommodate negative numbers as well, as the sign was irrelevant for our specific investigation. The Ecoinvent dataset comprises 2654 columns, of which 2648 are numeric and were utilized in the analysis. Within the 2648 columns, we have identified a total of 1190 distinct chemical substances. These substances have been categorized into five classes according to their place of release: Air, Water, Soil, Natural Resources, and Inventory Indicator.

IV. Results

We have demonstrated that a significant proportion of the columns in the Ecoinvent database conform to Benford's law, specifically 2193 out of 2648. The majority of nonconforming cases had a significant number of missing values, resulting in an insufficient amount of data for analysis. Only 70 columns met the criteria of having sufficient observations and not conforming. Figure 2 shows proportion of columns that conform and those that do not. We can see that our R package also divides conformity into four levels: close conformity, acceptable conformity, and marginally acceptable conformity. For our research we have counted all levels of conformity as the same since we can't expect data to conform perfectly in real life.

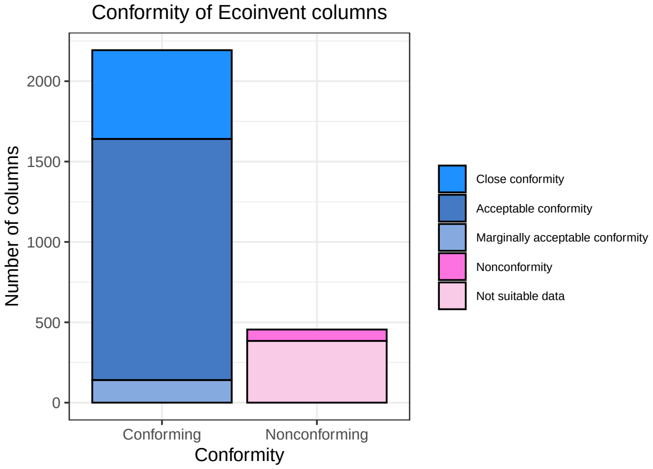

Fig. 2: Conformity of Ecoinvent columns

In order to show that LCA data generally conforms to Benford's law, in addition to Ecoinvent, which underwent individual column testing, other databases were examined as a whole using the Benford method. These databases are

Database	ChiSq	ChiSqP	MantissaArcTest	MantissaArcTestP	MADConformity	MAD	Number of Observations
Ecoinvent	247.684	0	0.033	0	Close conformity	0	49031793
Worldsteel	44.56	0	0.039	0	Close conformity	0.002	73044
OzLCI2019	63.034	0	0.036	0	Close conformity	0.002	178940
Greendelta	29.375	0	0.032	0	Close conformity	0.001	205344
Needs	150.271	0	0.034	0	Close conformity	0.001	808382
Bioenergiedat	29.626	0	0.053	0	Nonconformity	0.018	834

TABLE I: Statistical indicators of Benford's conformity for LCA databases studied.

open source and available online, and even though they are not as big and detailed as Ecoinvent, they also conformed. From the Figure 3 we can see that all of these databases conformed with the testing, except for one that was outdated and lacked sufficient data. By visual inspection, the last database appears acceptable, however, from a statistical standpoint, it does not meet the required criteria. Table I shows all statistical values produced by our Benford function. As aforementioned, the Bioenergiedat [1] stands out as the only non-conforming LCA database due to the low amount of observations (n=834), the credibility of this statistical test is questionable. For other databases, all statistical conformity tests signal strong conformity with the expected distribution, which results into a Close conformity according to commonly used MAD conformity test. All open source databases were found on OpenLCA Nexus website [2]. Worldsteel [3] presents comprehensive worldwide and regional Life Cycle Inventory (LCI) data for 16 different steel products, ranging from hot rolled coil to plate, rebar, sections, and coated steels. This study was conducted using the worldsteel LCI methodology report and ISO standards 14040 and 14044. It is considered the most extensive and precise LCI dataset for steel products worldwide. OzLCI2019 [4] is a free LCA database created by The Evah Institute in Australia. The database covers the supply of goods from the Australasian area, including imports, and was created using openLCA. Greendelta [5]is focused on providing secondary data for Product Environmental Footprints (PEFs) within the openLCA software. The objective is to address the environmental effects of products, such as carbon emissions, by creating a standardized European approach for evaluating and categorizing items. Needs [6] database was established by the NEEDS (New Energy Externalities Developments for Sustainability) project and contains life cycle inventories of future energy supplies in Europe. The dataset includes life cycle inventory (LCI) information related to upcoming transportation services, power, and material supply in the industrial sector. Bioenergiedat was developed as part of the German BioEnergieDat project and was finalized in February 2013. The objective of the project was to establish supply chains for bioenergy alternatives, with a particular focus on the German context. The primary focus was on bioenergy derived from wood and wastewood, wheat, and biowaste.

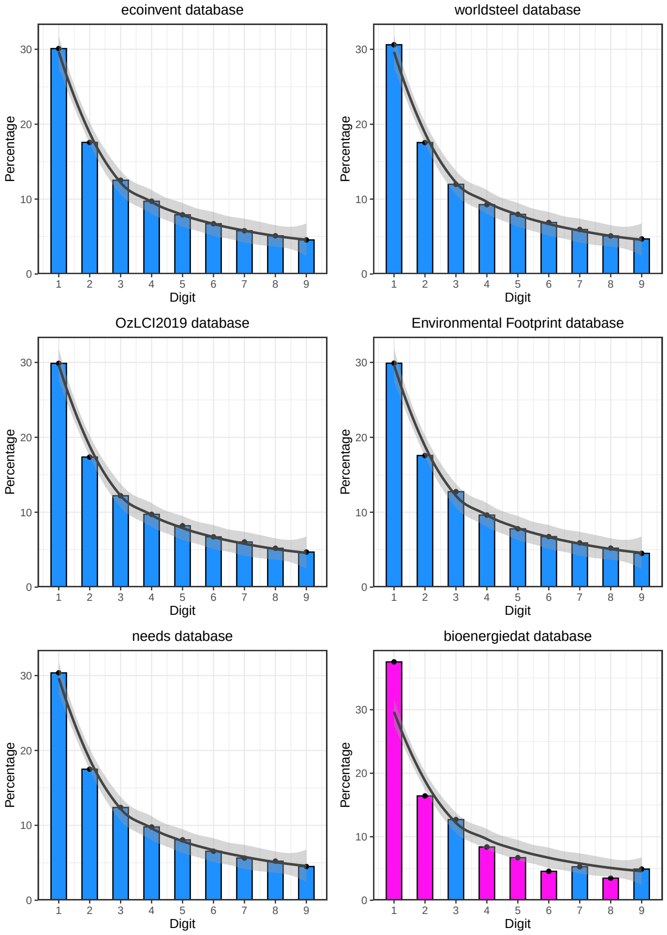

Fig. 3: Comparison of first digit distribution between different databases

Conclusion and future work

Our analysis has determined that the generally, LCA data should conform to Benford's law. The results were consistent with expectations and demonstrated that this straightforward statistical approach can be employed to evaluate first assessments of data reliability. We anticipated that the entire LCA dataset would conform, while we acknowledged the possibility that certain minor components would not. Non-conformity in the results does not imply the inaccuracy or fabrication of the data; rather, it indicates the necessity for expert verification.

[1] https://bioenergiedat.de/

[2] https://nexus.openlca.org/

[3] https://worldsteel.org/

[4] http://www.evahnaturepositive.com/

[5] https://www.greendelta.com/

[6] https://nexus.openlca.org/database/NEEDS

Establishing general conformity is a stepping stone as the method can be used to reliably test for database anomalies quickly and efficiently. However, nonconformity would still require expert assessment to determine the reasons. Typically, anomalies might be statistically undetectable in large number of observations. This is commonly addressed by sub-sampling the data. By testing individual parameters we have shown that for Ecoinvent, almost all are conforming (>95%). Future work should include other sub-sampling techniques such as geographic location, compartment (Air, Water, Soil, Natural Resources, and Inventory Indicator), and others. These approaches might provide a more granular view and possibly be able to detect anomalies in individual sub-samples thereby making consequent analysis easier and more accurate.

References

[1] M. A. Curran, "Life cycle assessment: a review of the methodology and its application to sustainability," *Current Opinion in Chemical Engineering*, vol. 2, no. 3, pp. 273–277, 2013, energy and environmental engineering / Reaction engineering and catalysis.

[2] R. U. Ayres, "Life cycle analysis: A critique," *Resources, Conservation and Recycling*, vol. 14, no. 3, pp. 199–223, 1995, life Cycle Management.

[3] A. E. Björklund, "Survey of approaches to improve reliability in lca," *The International Journal of Life Cycle Assessment*, vol. 7, no. 2, pp. 64–72, 2002.

[4] O. J. Hanssen and O. A. Asbjørnsen, "Statistical properties of emission data in life cycle assessments," *Journal of Cleaner Production*, vol. 4, no. 3, pp. 149–157, 1996.

[5] R. M. Fewster, "A simple explanation of benford's law," *The American Statistician*, vol. 63, no. 1, pp. 26–32, 2009.

[6] R. Frischknecht, N. Jungbluth, H.-J. Althaus, G. Doka, R. Dones, T. Heck, S. Hellweg, R. Hischier, T. Nemecek, G. Rebitzer *et al.*, "The ecoinvent database: overview and methodological framework (7 pp)," *The international journal of life cycle assessment*, vol. 10, pp. 3–9, 2005.

[7] G. Wernet, C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz, and B. Weidema, "The ecoinvent database version 3 (part i): overview and methodology," *The International Journal of Life Cycle Assessment*, vol. 21, pp. 1218–1230, 2016.

[8] R. Frischknecht and G. Rebitzer, "The ecoinvent database system: a comprehensive web-based lca database," *Journal of Cleaner Production*, vol. 13, no. 13-14, pp. 1337–1343, 2005.

[9] N. Bamber, I. Turner, V. Arulnathan, Y. Li, S. Z. Ershadi, A. Smart, and N. Pelletier, "Comparing sources and analysis of uncertainty in consequential and attributional life cycle assessment: review of current practice and recommendations," *The International Journal of Life Cycle Assessment*, vol. 25, pp. 168–180, 2020.

[10] A. Takano, S. Winter, M. Hughes, and L. Linkosalmi, "Comparison of life cycle assessment databases: A case study on building assessment," *Building and Environment*, vol. 79, pp. 20–30, 2014.

[11] A. Edelen, W. W. Ingwersen, C. Rodríguez, R. A. Alvarenga, A. R. de Almeida, and G. Wernet, "Critical review of elementary flows in lca data," *The international journal of life cycle assessment*, vol. 23, pp. 1261–1273, 2018.

[12] A. Martínez-Rocamora, J. Solís-Guzmán, and M. Marrero, "Lca databases focused on construction materials: A review," *Renewable and Sustainable Energy Reviews*, vol. 58, pp. 565–573, 2016.

[13] S. Newcomb, "Note on the Frequency of Use of the Different Digits in Natural Numbers," *Amer. J. Math.*, vol. 4, no. 1-4, pp. 39–40, 1881.

[14] F. Benford, "The law of anomalous numbers," *Proceedings of the American philosophical society*, pp. 551–572, 1938.

[15] T. W. Singleton, "IT Audit Basics: Understanding and Applying Benford's Law," *Isaca Journal*, vol. 3, p. 6, 2011.

[16] D. A. Kenny, "Measuring model fit," 2015.

[17] C. Cinelli and M. C. Cinelli, "Package 'benford. analysis'," *Benford analysis for data validation and forensic analytics. Version 0.1*, vol. 5, 2022.