Quantification of Biodiversity from Historical Survey Text with LLM-based Best-Worst-Scaling

Anonymous

Abstract

In this paper, we evaluate methods to determine biodiversity via quantity estimation from historical survey text. To that end, we formulate classification tasks and finally show that this problem can be successfully framed as regression based on best-worst-scaling with LLMs. We find that this approach is more cost effective and similarly robust to a fine-grained multi-class approach, allowing automated quantity estimation across species.

1 Introduction

Long-term observation data plays a vital role in shaping policies aimed at preventing biodiversity loss driven by habitat destruction, climate change, pollution, or resource over-exploitation (Dornelas et al., 2013; Hoque and Sultana, 2024). Accurate measurement and monitoring over extended periods provide invaluable insights into ecological trends, yet these efforts hinge on the availability of reliable, relevant data and robust methods.

We are faced with both, scarcity and heterogeneity of historical records that inform us about biodiversity data from the past. Available data varies in resolution, from detailed records (e.g., point occurrences, trait measurements) to aggregated compilations (e.g., Floras, taxonomic monographs) (König et al., 2019), and projects often focus on the disaggregated end of the data spectrum (e.g., GBIF, cf. Telenius (2011)), particularly with presence/absence data (Dorazio et al., 2011; Iknayan et al., 2014).

Despite its significance, longitudinal biodiversity data is typically confined to post-1970s sources (van Goethem and van Zanden, 2021), leaving significant historical gaps. Historical sources such as natural history collections and records from the archives of societies offer valuable opportunities for extending datasets further back in time (Johnson et al., 2011; Brönnimann et al., 2018). Such (text or image-based) sources are rich in data but typically unstructured. This requires sophisticated extraction tools to produce meaningful insights from quantification. Initial research in this area leverages recent advances in NLP methods, enabling information retrieval based biodiversity detection in (scientific) literature (Kommineni et al., 2024; Langer et al., 2024; Lücking et al., 2022).

This paper focuses on evaluating methods for biodiversity quantification from semi-structured historical survey texts. To achieve this, we have tested tasks to distill meaningful metrics from textual information found in survey records. A particular interest is in testing the feasibility of best-worst-scaling (BWS) with a large language model as an annotator, which promises greater efficiency and cost-effectiveness compared to manual annotation (Bagdon et al., 2024).

In the following, we describe the data (section 2), outline the tasks and machine learning methods (section 3), and finally a case study (section 4).

2 Data

In 1845, the Bavarian Ministry of Finance issued a survey to evaluate biodiversity in the Bavarian Kingdom, a region that encompasses a variety of different ecosystems and landscapes. To that end, 119 forestry offices were contacted to complete a standardized questionnaire. Namely, trained local foresters should record in free text how frequently 44 selected vertebrate species occurred in the respective administrative territory, and in which habitats and regions they could be found.

Figure 1 shows the facsimile of a digitized survey page. It features a header containing instructions and a number of records describing animal species with their respective responses. These historical survey documents are preserved by the Bavarian State Archives (cf. Rehbein et al., 2024).

Animal	Text	Binary	BWS	Multi	Multi descriptor
Ducks	Bedecken Isar-Strom, wie Amper und Moosach in ganzen Schwärmen. Cover Isar-torrent, likewise Amper and Moosach in whole swarms.	1	1.00	5	Abundant
Roe Deer	Ist hier zu Hause, und beinahe in allen Waldtheilen zu finden. Is at home here and can be found in almost all parts of the forest.	1	0.88	4	Common
European Adder	Kommt wohl aber eben nicht häufig vor. Does indeed appear but just not that often.	1	0.44	3	Common to Rare
Lynx	Höchst selten wechseln derlei Thiere von Tyrol herüber. Very rarely do such animals cross over from Tyrol.	1	0.12	2	Rare
Wild Goose	Kommt nur äußerst selten zur Winterszeit vor. Occurs only very rarely at winter time.	1	0.06	1	Very Rare
Owl	Horstet dahier nicht und verstreicht sich auch nicht in diese Gegend. Does not nest here and does not stray into this area.	0	0.00	0	Absent
Wolf	Kommt nicht mehr vor. No longer occurs.	0	0.00	-1	Extinct

Table 1: Data Examples with Annotation (our own translations)

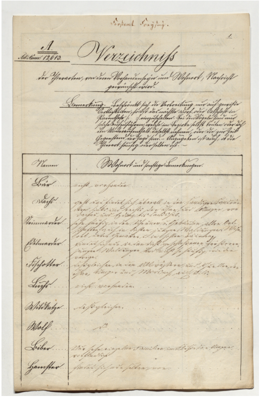

Figure 1: Facsimile of a survey page, Freysing forestry office in the Upper Bavaria district.

The archival sources were digitized and enriched with metadata, including, among others, taxonomic norm data according to the GBIF[1] database (Telenius, 2011) and geographical references to forestry offices. The transcription of the digitized documents written in Kurrent was carried out by student researchers with the aid of Transkribus[2] handwritten text recognition. The dataset is freely available on zenodo (Rehbein et al., 2024): `https://doi.org/10.5281/zenodo.14008158`

In total, the dataset contains 5,450 entries[3] among which are also a number of empty (striked out) or 'see above' responses. The unique set we used for our experiments contains 2,555 texts. We find that the foresters' replies vary considerably in length where most texts contain 3 to 10 token and only a few texts more than 20 tokens. See Table 1 for examples with annotations according to the tasks detailed in the next section.

3 Tasks & Experiments

The main task in this paper is to assign a quantity label to a text, indicating the frequency with which an animal species occurs in a specific area. This can be operationalized in various ways, either through a classification task or through regression. In both it can be difficult to obtain consistent labels from humans by asking them to assign a value from a rating scale (Schuman and Presser, 1996; Likert, 1932). Not only is it difficult for annotators to rate texts consistently, it is also difficult for researchers to design rating scales, considering design decisions such as scale point descriptions or granularity may bias the annotator. We evaluated three different task setups, as detailed in Table 1: Binary 'Presence vs. Absence' Classification, a 7-ary Multi-Class setup (Abundant to Extinct), and Continuous values scaled to $[0, 1]$. The former two tasks were annotated via manual annotation, while continuous values are derived through best-worst-scaling (BWS) with GPT-4 (Bagdon et al., 2024).

3.1 Binary 'Presence vs. Absence'

The simplest form of animal occurrence quantification is a binary distinction between the absence (0) or presence (1) of a given species, an annotation scheme as popular as it is problematic in biodiversity estimation.[4] In our annotation, the PRESENT label is given if the species was described as present at the time of the survey (thus excluding explicit mentions of extinction).

The annotation workflow was done in iterative steps with discussions. Agreement was nearly perfect. Overall, from the set of 2,555 unique texts, 1,992 (78%) are in class PRESENT, while 563 (22%) are in class ABSENT.[5]

[1] `gbif.org`

[2] `transkribus.org`

[3] One text entry per species and office, but also entries for species that were not explicity prompted.

[4] ABSENCE may just stem from non-detection, rather than real absence (Dorazio et al., 2011; Iknayan et al., 2014).

[5] In the complete dataset, absence texts make up more than

To test the feasibility of the binary task, we created training curves with different models, namely BERT against Logistic Regr., SVM, and Random Forest on Unigrams. We use 20% of the data for testing, and take another 20% from the training set for hyperparameter search at each cumulative 100 text increment. Despite the 78% majority baseline, we find that the models perform well, and they need only a few hundred texts for training to reach an F1-macro score in the high 90s.

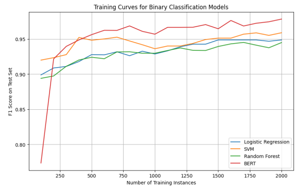

Figure 2: Training Curves of different models on incremental training data (binary classification)

Upon feature weight interpretation of the Logistic Regr. and LIME on BERT (Ribeiro et al., 2016) we find that there is some bias in the data, such that classification decisions occur on tokens that are not explicit quantifiers ('Danube') that are substitutable w/o changing the classification result. This presents a promising future research direction, but a matching (Wang and Culotta, 2020) or counterfactual approach (Qian et al., 2021) appears challenging on our heterogeneous data. Yet, we annotated the best features w.r.t. their 'spuriousness' and find that classifiers are still robust without spurious features. This annotation also gave us a list of quantifiers which we utilized for transfer learning of a regression model (section 3.3).

3.2 Multi-Class

Since the quantification of frequency in practice exceeds the differentiation between presence and absence of animals, a multi-class approach can provide more details. We used a 7-class system, categorizing texts based on the schema as shown by the descriptors in Table 1, ranging from Abundant (5) to Extinct (-1). We decided to annotate four species for our case study (section 4): Roe deer, Eurasian Otter, Eurasian Beaver, Western Capercaille, each within the 119 forestry offices. A sample of 100 texts was annotated by a second person, which resulted in a Cohen's κ of 0.78, indicating high agreement.

We trained a few models with a 5-fold cross validation, and find that the language agnostic LaBSE model (Feng et al., 2022) performs better than monolingual BERT models and a Logistic Regression. We also test a zero shot classification with GPT-4. See Appendix for the prompt.

Model	F1 Micro	F1 Macro
Logistic Regression	0.69	0.61
gbert-base	0.63	0.51
bert-base-german	0.73	0.63
LaBSE	**0.77**	**0.68**
GPT4 Zero Shot	0.70	0.56

Table 2: Multi-class model performance.

As seen in Table 2, the task is overall fairly challenging. We find that the main problem is posed by the underrepresented classes, as shown by the discrepancy between micro and macro score, indicating that more data would help, which is however expensive to obtain. Zero shot classification is biased towards the 'rare' frequencies.

3.3 Continuous

Finally, we experiment with operationalizing our task as a regression problem, with the aim of generalizing the quantification problem to less arbitrary categories, and a possibly imbalanced dataset (Johan Berggren et al., 2019). While a naive labeling of quantifiers showed promising results, it is a challenge to create a comprehensive test set based on heuristic annotation. Thus, we experiment with Best-Worst-Scaling, aided by an LLM.

3.3.1 Best-Worst-Scaling with GPT-4

Best-worst scaling is a comparative judgment technique that helps in ranking items by identifying the best and worst elements within a set. This approach is easier to accomplish than manual labeling and there are fewer design decisions to make. In a best-worst-scaling setting, the amount of annotations needed to rank a given number of text instances depends on three variables, namely 1) the total number of texts used (corpus size), 2) the number of texts in each comparison set (set size), and 3) the number of comparison sets each text appears in (comparisons).

The number of comparisons divided by corpus size is regarded as the variable N, where $N = 2$ generally yields good results in the literature (Kiritchenko and Mohammad, 2017). A reliable set size is 4, since choosing the best and worst text instance from a 4-tuple set essentially provides the same results as five out of six possible pairwise comparisons (Bagdon et al., 2024).

half of all text descriptions, but often amount to empty or 'strike-out' responses.

We took a random sample of 1000 texts from our corpus (excluding texts with ABSENCE annotation, thus making the task harder, but giving us a more realistic distribution). With a tuple size of 4, $N = 2$, thus every text occurring in exactly 8 different tuples, we get 2000 comparison sets that were then prompted to GPT-4 individually. Prompts are in the Appendix. We asked a few native German post-graduate students to annotate a subset of 50 tuples each. Table 3 shows the Cohen's κ agreement between humans and LLM. We find that human agreement is similar to the agreement between humans and GPT-4, indicating i) that the task is overall feasable but by far not trivial, and ii) that GPT-4 is largely on par with humans. Furthermore, it appears that it easier to identify the worst instance, rather than the best.

Annotator1	Annotator2	B	W	B + W
AO	GPT4	0.68	0.55	0.45
AR	GPT4	0.49	0.57	0.34
TP	GPT4	0.63	0.57	0.43
DS	GPT4	0.44	0.71	0.43
KB	GPT4	0.47	0.68	0.41
MR	GPT4	0.49	0.63	0.41
Average		0.53	0.62	0.41
AR	DS	0.56	0.65	0.45
DS	KB	0.56	0.62	0.40
MR	AR	0.51	0.65	0.39
TP	AO	0.73	0.55	0.48
Average		0.59	0.62	0.43

Table 3: Cohen's κ Agreement between humans and LLM in Best-Worst-Annotation (B: Best, W: Worst, B+W: Best + Worst)

By counting how often each text got chosen as best, worst, or as one of two other texts, we calculated a score $s(i)$ as detailed in equation (1), resulting in an interval scale $[-1, 1]$, which we normalized to a scale $[0, 1]$. This scales (and ranks) the entire dataset, so it can be used for regression. We find a unimodal flat inverted U-shape in the score distribution without notable outliers.

$$s(i) = \frac{best(i) - worst(i)}{overall(i)} \tag{1}$$

3.3.2 Regression Models

We trained a variety of different models with 5-fold cross validation to optimize the regression of the text-label pairs generated through best-worst-scaling, as detailed in Table 4. We compare a Kernel Ridge Regression (KRR) baseline against Sentence BERT models with regression head, and test a transfer learning setup, for which we scale the 114 n-gram quantifiers as extracted from the binary Logistic Regression with another GPT-4 BWS. We then match these scores to the texts in the training set and tune a LaBSE model before using that tuned model for the final task.

Features/Training Strategy	Model	MAE	R^2
Unigrams	KRR	0.170	0.415
Frozen LaBSE Embeddings	KRR	0.118	0.678
Regression Head	bert-base-german	0.149	0.516
Regression Head	LaBSE	0.133	0.607
Reg. Head + Transfer	LaBSE	0.107	0.730

Table 4: Comparison of different training strategies for regression based on BWS-Scaling

It is curious that the KRR with LaBSE embedding features benefits substantially from hyperparameter tuning, reaching superior results over LaBSE with regression head. The Transfer Model offers the best performance, with acceptably high explained variance (R^2 = .72) and only .11 Mean Absolute Error (MAE), which makes this model useful for downstream prediction in our case study. However, more data would likely also help (training curves show continuous improvement).

4 Case Study

For a proof of concept, we map the regression model results to the multi-class gold annotation. See Figure 3 for a mapping of the entire dataset, and Figure 4 for species specific distributions across all 119 offices (Roe deer, Eurasian otter). We find that in the former that there is a strong correlation, but also that extinction is not encoded in the regression and furthermore that higher values are challenging to predict. Regarding Figure 4, we can see a fairly good match between the regression result (left) and the multi-class annotation (right). We conclude that the BWS approach is robust and cost effective to estimate animal quantity.

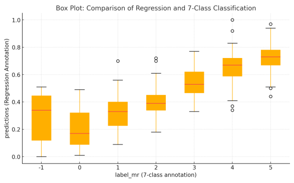

Figure 3: Multi-Class vs. Regression Distribution

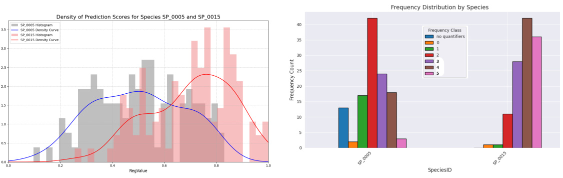

Figure 4: Density and histogram for Roe deer (SP_0015, red) and Eurasian otter (SP_0005, blue).

References

Christopher Bagdon, Prathamesh Karmalker, Harsha Gurulingappa, and Roman Klinger. 2024. " you are an expert annotator": Automatic best-worst-scaling annotations for emotion intensity modeling. *arXiv preprint arXiv:2403.17612*.

Stefan Brönnimann, Christian Pfister, and Sam White. 2018. Archives of nature and archives of societies. In *The Palgrave Handbook of Climate History*, pages 27–36. Palgrave Macmillan UK, London.

Robert M Dorazio, Nicholas J Gotelli, and Aaron M Ellison. 2011. Modern methods of estimating biodiversity from presence-absence surveys. *Biodiversity loss in a changing planet*, pages 277–302.

Maria Dornelas, Anne E. Magurran, Stephen T. Buckland, Anne Chao, Robin L. Chazdon, Robert K. Colwell, Tom Curtis, Kevin J. Gaston, Nicholas J. Gotelli, Matthew A. Kosnik, Brian McGill, Jenny L. McCune, Hélène Morlon, Peter J. Mumby, Lise Øvreås, Angelika Studeny, and Mark Vellend. 2013. Quantifying temporal change in biodiversity: challenges and opportunities. *Proceedings of the Royal Society*, 280.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. http://arxiv.org/abs/2007.01852 Language-agnostic bert sentence embedding.

Thomas van Goethem and Jan Luiten van Zanden. 2021. Biodiversity trends in a historical perspective.

Sk Rezaul Hoque and Sk Rima Sultana. 2024. Addressing global environmental problems: Challenges, solutions, and opportunities. *The Social Science Review: A Multidisciplinary Journal*, 2(2):124–130.

Kelly J Iknayan, Morgan W Tingley, Brett J Furnas, and Steven R Beissinger. 2014. Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, 29(2):97–106.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. https://doi.org/10.18653/v1/W19-4409 Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy. Association for Computational Linguistics.

Kenneth G Johnson, Stephen J Brooks, Phillip B Fenberg, Adrian G Glover, Karen E James, Adrian M Lister, Ellinor Michel, Mark Spencer, Jonathan A Todd, Eugenia Valsami-Jones, Jeremy R Young, and John R Stewart. 2011. Climate change and biosphere response: Unlocking the collections vault. *Bioscience*, 61(2):147–153.

Svetlana Kiritchenko and Saif Mohammad. 2017. https://doi.org/10.18653/v1/P17-2074 Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Vamsi Krishna Kommineni, Waqas Ahmed, Birgitta Koenig-Ries, and Sheeba Samuel. 2024. Automating information retrieval from biodiversity literature using large language models: A case study. *Biodivers. Inf. Sci. Stand.*, 8.

Christian König, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, and Holger Kreft. 2019. Biodiversity data integration—the significance of data resolution and domain. *PLoS biology*, 17(3):e3000183.

Lars Langer, Manuel Burghardt, Roland Borgards, Ronny Richter, and Christian Wirth. 2024. The relation between biodiversity in literature and social and spatial situation of authors: Reflections on the nature–culture entanglement. *People and Nature*, 6(1):54–74.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Andy Lücking, Christine Driller, Manuel Stoeckel, Giuseppe Abrami, Adrian Pachzelt, and Alexander Mehler. 2022. Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Language resources and evaluation*, 56(3):807–855.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. https://doi.org/10.18653/v1/2021.acl-long.422 Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online. Association for Computational Linguistics.

Malte Rehbein, Andrea Belen Escobari Vargas, Sarah Fischer, Anton Güntsch, Bettina Haas, Giada Matheisen, Tobias Perschl, Alois Wieshuber, and Thore Engel. 2024. https://doi.org/10.5281/zenodo.14008158 Historical animal observation records by bavarian forestry offices (1845): Description of the data sets. Version 1.3 as of 2024-10-29.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Anders Telenius. 2011. Biodiversity information goes public: Gbif at your service. *Nordic Journal of Botany*, 29(3):378–381.

Zhao Wang and Aron Culotta. 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.308 Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

APPENDIX

GPT-4 Multi-Classification Prompt

System-prompt: You are a German native expert in text classification. Use the provided classification scheme to classify German texts based on species frequency descriptions.

User-prompt: You are a classification model. Classify the given German text into one of the following categories:
- Abundant (5): Species is very frequently observed or present.
- Common (4): Species is commonly found in the area.
- Common to Rare (3): Species is observed, but not very frequently.
- Rare (2): Species is rarely seen in the area.
- Very Rare (1): Species is seen only in exceptional circumstances.
- Absent (0): Species is not observed in the area.
- Extinct (-1): Species no longer exists in the area.

Read the provided text and classify it according to this scheme. Here is the text to classify:
Text

GPT-4 Best-Worst-Scaling Prompt

System-prompt: You are an expert annotator specializing in Best-Worst Scaling of German texts based on quantity information about animal occurrences.

User-prompt: (Texts 1 to 4 were substituted with the actual texts of a tuple): Task: From the following german texts about animal occurrence, identify:
Best: The text conveying the highest quantity (e.g., presence, frequency, population size)
Worst: The text conveying the lowest quantity.
1. Text 1
2. Text 2
3. Text 3
4. Text 4

JSON format for your answer:
{ ”Best”: [Text Number],
”Worst”: [Text Number]}