More Efficient Manual Review of Automatically Transcribed Tabular Data

By Bjørn-Richard Pedersen, Rigmor Katrine Johansen, Einar Holsbø, Hilde Sommerseth and Lars Ailo Bongo

To cite this article: Pedersen, B.-R., Johansen, R. K., Holsbø, E., Sommerseth, H., & Bongo, L. A. (2024). More Efficient Manual Review of Automatically Transcribed Tabular Data. *Historical Life Course Studies, 14*, 3–15. https://doi.org/10.51964/hlcs15456

HISTORICAL LIFE COURSE STUDIES

VOLUME 14

2024

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies was established within *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation, the International Institute of Social History, the European Society of Historical Demography, Radboud University Press, Lund University and HiDO Scientific Research Network Historical Demography. Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

HISTORICAL LIFE COURSE STUDIES
VOLUME 14 (2024), 3–15, published 04-04-2024

More Efficient Manual Review of Automatically Transcribed Tabular Data

Bjørn-Richard Pedersen — Norwegian Historical Data Centre, UiT The Arctic University of Norway
Rigmor Katrine Johansen — Department of Health and Care Sciences, UiT The Arctic University of Norway
Einar Holsbø — Department of Computer Science, UiT The Arctic University of Norway
Hilde Sommerseth — Norwegian Historical Data Centre, UiT The Arctic University of Norway
Lars Ailo Bongo — Department of Computer Science, UiT The Arctic University of Norway

ABSTRACT

Any machine learning method for transcribing historical text requires manual verification and correction, which is often time-consuming and expensive. Our aim is to make it more efficient. Previously, we developed a machine learning model to transcribe 2.3 million handwritten occupation codes from the Norwegian 1950 census. Here, we manually review the 90,000 codes (3%) for which our model had the lowest confidence scores. We allocated these codes to human reviewers, who used our custom annotation tool to review them. The reviewers agreed with the model's labels 31.9% of the time. They corrected 62.8% of the labels, and 5.1% of the images were uncertain or assigned invalid labels. 9,000 images were reviewed by multiple reviewers, resulting in an agreement of 86.4% and a disagreement of 9%. The results suggest that one reviewer per image is sufficient. We recommend that reviewers indicate any uncertainty about the label they assign to an image by adding a flag to their label. Our interviews show that the reviewers performed internal quality control and found our custom tool to be useful and easy to operate. We provide guidelines for efficient and accurate transcription of historical text by combining machine learning and manual review. We have open-sourced our custom annotation tool and made the reviewed images open access.

Keywords: Population census data, Machine Learning, Historical data, Manual review, Occupation codes, Norway 1950, Norwegian population data, Efficient manual review, Automatically transcribed, Tabular data, Manual review and correction, Interviews, Norwegian occupation data, Historical occupation data, Norwegian Historical Data Centre, UiT The Arctic University of Norway

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs15456

1 INTRODUCTION

The Norwegian Historical Data Centre (NHDC) at UiT The Arctic University of Norway is one of the partners constructing the Norwegian Historical Population Register (HPR) (http://www.rhd.uit.no/nhdc/hpr.html). HPR will include the records of 9.7 million people who lived in Norway from 1801 to 1964, with the main building blocks being population censuses and church books. One important goal is to transcribe historical population sources with a high degree of accuracy, and machine learning can be a key tool in achieving this while keeping costs and time usage low. The building blocks in the HPR have been transcribed using three approaches. First, professional transcribers and volunteers have manually transcribed about 30 million person entities from the 19th- and early 20th century population censuses and church books dating back to the early 18th century. Second, selected columns in the church books have been manually transcribed through an agreement between the National Archives of Norway and three commercial genealogy companies. Third, we have developed a machine learning model to automate the transcription of 2.3 million handwritten occupation codes from the Norwegian 1950 population census (Pedersen et al., 2022).

In the third approach, we had three main requirements for the automated occupation code transcription: (i) the error rate of the automatic transcription solution should fall within the same range that is accepted for human transcribers at the NHDC (<=3%), (ii) given the 2.3 million code images we extracted from the 1950 census[1], fewer than 100,000 images should be sent for manual validation and correction, and (iii) no systematic errors should be introduced to the automatically transcribed data (Pedersen et al., 2022). There is a trade-off between the first and second goals since a higher accuracy requires more manual transcription. We can tune this trade-off by adjusting the prediction confidence required for a code to be automatically transcribed. A higher confidence threshold improves accuracy but requires more manual validation and correction. Thus, more efficient manual work can increase machine learning accuracy. An important part of tuning a machine learning model for transcription is, therefore, finding the best trade-off between these two conflicting goals.

Our resulting model satisfied all three requirements. We achieved 97% accuracy, but for 82,177 images the classification confidence score was below the 65% threshold we had set as our criterion (Pedersen et al., 2022). In addition, during post-processing of the results, we found that 16,156 images had been given invalid or non-numerical codes. Since all images must be transcribed these 98,333 images require manual validation and possible correction. This manual review necessitates a significant amount of human effort and time. For other data, with non-numeric columns that are more challenging for machine learning, we expect lower model accuracy and, therefore, even more manual work to maintain the required accuracy for the automatically transcribed images. It is therefore important for transcription projects to make this time-consuming and costly manual work as efficient as possible.

To our knowledge, this is the first study on the evaluation of machine learning classification results for historical population sources. However, the need for human labeling in machine learning projects has been a motivation for the development of data labeling tools such as Amazon Mechanical Turk (https://www.mturk.com/), Label Studio (https://labelstud.io/), Prodigy (https://prodi.gy/), Snorkel Flow (https://snorkel.ai/snorkel-flow-platform/), and many more (https://github.com/doccano/awesome-annotation-tools provides a list). These tools typically aim to reduce the effort of labeling the data required to train machine learning models. Consequently, these systems are designed for efficient labeling of training data by a team of labelers working for large companies or organizations, which may result in expensive licensing fees. Many of these tools enable the use of humans-in-the-loop when training a machine learning model (Bernard et al., 2018a; Cohen-Wang et al., 2019). Additionally, there is a growing focus in the machine learning community on improving the quality of labels, not just on model performance. For example, by analyzing and flagging labels that differ between human annotators. As a result, most commercial tools support such quality control workflows. Furthermore, there are specialized tools for data cleaning (such as Trifacta, https://www.trifacta.com/), and for correcting labels (Xiang et al., 2019). Moreover, recent research has focused on explaining why machine learning models make the classifications that they do (Dwivedi et al., 2023; Kim et al., 2023), resulting in the development of numerous libraries and tools (https://github.com/wangyongjie-ntu/Awesome-explainable-AI provides lists of such tools).

1 The 1950 population census consisted of 7.3 million occupation code images in total, but approximately 5 million of these were programmatically found to be blank, and not containing any actual occupation code. However, this process did not guarantee that a blank image could not still be found among the 2.3 million remaining images.

Although we tested some of the tools, we found that we could not utilize the advanced functionality they provided well enough because our approach differs from typical labeling approaches. Initially, we identify the images that are likely mislabeled using our model's confidence scores. Subsequently, the selected images are manually verified and corrected. Our custom annotation tool groups the images according to the predicted label. More advanced algorithms for autocorrecting large sets of images are described in (Hung et al., 2015; Liu et al., 2019; Xiang et al., 2019). We do not use the corrected labels to find additional labels for correction, since our error analysis results show that the reviewed images are already selected based on specific features.

Understanding how to structure the workflow for human reviewers is also important, as the task of high-quality manual review can become tedious and monotonous. We believe the workflow should take into account that professional transcribers have different training and motivation for the tasks than temporary workers who typically label datasets for training machine learning models.

To address the challenge of making manual review efficient, accurate, and motivating for historical data reviewers, we used a custom annotation tool to verify and correct the Norwegian 1950 census occupation codes that our model could not automatically transcribe with confidence, to get insight into how accurate and time-consuming this manual work is. Subsequently, we interviewed the reviewers to understand how to improve the manual workflow.

To the best of our knowledge, this is the first paper to focus on manual quality control of machine learning transcribed historical text, with emphasis on efficiency and accuracy. We make three main contributions:

1. To understand the quality control requirements for manual review, we provide an error analysis for machine learning and human transcriptions of the 3% of the texts that our machine learning model was unable to transcribe.
2. To improve efficiency while maintaining high quality for the transcriptions and keeping the reviewers motivated for the job, we assessed the manual transcription workflow to offer guidelines on organizing the manual review of texts transcribed by machine learning for use by professional transcribers.
3. To enable others to use our tools in their projects, we have open-sourced our custom annotation tool, and made the reviewed images open access.

2 METHODS

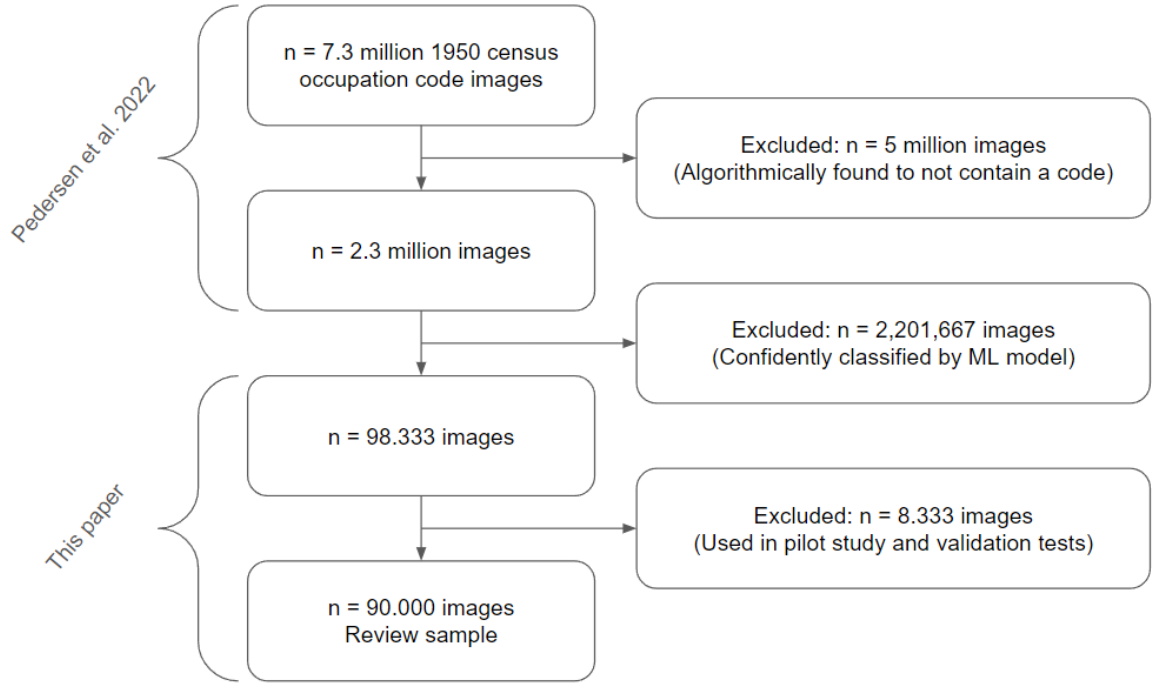
2.1 SELECTING THE RECORDS FOR REVIEW

Previously, we developed a machine learning (ML) model for automatic transcription of occupation codes in the Norwegian 1950 census which comprises a total of 7.3 million images, among which 2.3 million contain occupation codes (Pedersen et al., 2022). After classifying all 2.3 million images, 98,333 images were selected for manual review based on various factors; 82,177 images failed to meet the confidence threshold set for automatic transcription, 2,273 images had a non-numerical label (e.g., 't4b'), and 10,376 images had labels that were not in Statistics Norway's official list of occupation codes for the 1950 census. Furthermore, 3,507 images received labels that, while present in the official code list, were so rare that they did not appear in our training set. Although our machine learning model can correctly transcribe unseen labels, we decided to include these since they represent rare codes which we know our model may not accurately classify.

We conducted a pilot study for the review process to (i) identify useful features of state-of-the-art data labeling tools for this task, (ii) recognize potential problems to be aware of during the manual review, and (iii) determine how to analyze the results after the end of the project period. Our experience from the pilot was that the additional functionality of the commercial labeling tools did not benefit this project, since we do not need to administer large teams of reviewers and since we can as easily implement a custom graphical user interface (GUI) for this specific review task. Additionally, we learned that we needed to create a set of clear and common instructions that reviewers could refer to throughout the labeling process.

Of the 98,333 images mentioned above, we used 8,333 for validation tests and the pilot study, leaving a dataset of 90,000 images that underwent manual review in this study (Figure 1).

Figure 1 *Flow chart of the selection process for the images that needed to be manually reviewed*



2.2 REVIEWING THE REVIEWS

In our previous paper (Pedersen et al., 2022), we manually labeled a training set using an annotation tool that we created specifically for manual verification and correction of labeled images (https://github.com/HistLab/More-efficient-manual-review-of-automatically-transcribed-tabular-data). We used the PyQt GUI toolkit (https://riverbankcomputing.com/software/pyqt/intro) and implemented the back-end in Python. Approximately 3–4 workdays were spent on the tool's implementation. This tool was also employed in the current project. It is installed locally, and the tool's graphical user interface (GUI) can display up to 60 images at once, grouped according to the machine learning model's predicted code for each image. This predicted code is displayed in the top of the GUI (Figure 2). Reviewers manually reclassify incorrect labels by entering the new label into the textbox corresponding to the corrected image. The correctly labeled images are ignored in this process.

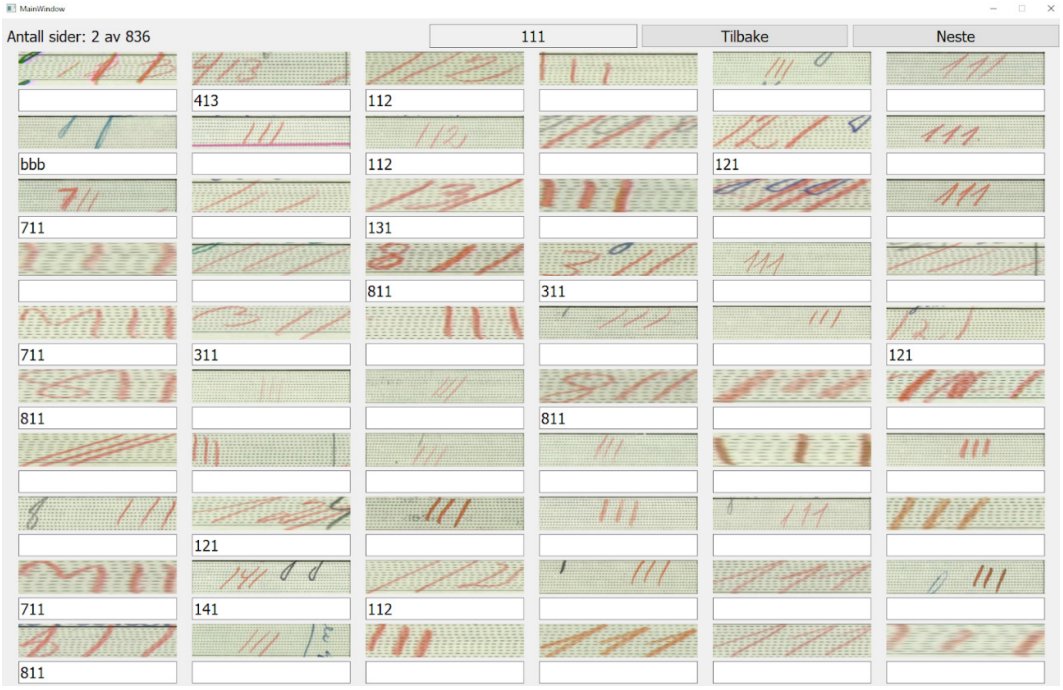
Manual reclassification should adhere to the Histform standard (https://rhd.uit.no/histform/histform.html). This standard, consisting of guidelines originally developed for transcribing Norwegian historical population censuses, is also well-suited for the review process in this project. Included in this standard is a set of specific symbols, such as '??' for instances when the reviewer is uncertain about the code in an image, and ('<new text> %<old text>%') to indicate when the original text has been crossed out and replaced, among others. A set of instructions summarizing the relevant aspects of the Histform standard was created. These instructions were handed out to each reviewer before they started.

To understand the manual effort required for this project, the tool was modified to record the time each reviewer spent on each group of images. Records indicating that a reviewer spent more than one hour on a page were removed under the assumption that a break had been taken, due to the absence of a pause function in the tool. The reviewers were aware that their time was being recorded, and this was also discussed during the interviews. We anonymized all data and stored the images, corrected labels, and per-group times in a single database.

We recruited six employees from the NHDC to perform the manual review of our images. All six are women and three of them are professional transcribers, holding several years' worth of experience with transcription of historical population sources. The other three were research assistants or students who have been working with historical data, but never with this type of handwritten source material. In addition, one of the authors, Bjørn-Richard Pedersen (B.R.P.), contributed as a reviewer. We believe these seven reviewers represent levels of experience that are relevant for this type of project.

We divided the 90,000 unclassified images evenly among the seven reviewers, yielding roughly 13,000 images each. They were given one month to review these images.

Figure 2 *A screenshot of our custom annotation tool used for the manual review*



Note: The text on the upper left translates to "number of pages: 2 of 836". The buttons on the right translate to "back" and "next".

2.2.1 INTRACODER AGREEMENT

We measured the consistency of the reviewer's own classifications by reassigning approximately 14% of their images back to them. These images were not used when calculating inter-reviewer performance, but rather as a test of the individual reviewer's attention to their task. We wanted to observe if they consistently labeled the images.

2.2.2 INTERCODER AGREEMENT

A random 10% of the 90,000 automatically labeled images were assigned for review by pairs of reviewers. The results from these overlapping reviews were then used to evaluate the inter-reviewer agreement. Note that we are ignoring the machine learning model's labels in the analysis. Finally, we ended up with 4 categories (Table 1):

1. *Certain and agreed.*
2. *Certain and disagreed.*
3. *Uncertain and agreed.*
4. *Uncertain and disagreed.*

Table 1 *An overview of the categories we use in the project, and examples of how an occupation code image may be categorized based on the labels from the human reviewers*

Reviewer A	Reviewer B	Model label	Category
531	531	531	Certain and agreed
531	531	999	Certain and agreed
531	999	888	Certain and disagreed
531	888	888	Certain and disagreed
182??	182??	141	Uncertain and agreed
531@537	531	181	Uncertain and disagreed

2.2.3 MACHINE MISCLASSIFICATION

The distribution of occupation codes may affect the trained model. Specifically, the 11 most common occupations comprise 70% of the images in our data, while the remaining 30% of images belong to one of 329 other occupations. Some codes are seen only once. Hence, the distribution of codes in this dataset is highly skewed, and the model is trained on thousands of examples of certain codes and may see other codes only once or not at all. We evaluate if small occupations get misclassified more often and, if so, whether they get misclassified as more common occupations.

Performing a manual review after the classification allows us to calculate the misclassification-rate of our model per class, and to identify the classes to which codes are most frequently mislabeled. However, this manual-review misclassification rate is calculated based primarily on the images for where the model's confidence was low, and thus may not reflect the overall misclassification rate. For the analysis, labels containing uncertainty symbols, as defined by the Histform standard, were removed, indicating that the human reviewer could not confidently assign a label to these images. We therefore assume that the remaining human-labeled codes are correct. We then separated the remaining labels into two categories: where human and machine agreed, and where they disagreed. We can then count the number of instances for each label and order our classes by size. To find the classes the model is biased towards, we order the incorrect labels by frequency.

2.3 EVALUATION OF THE ANNOTATION WORKFLOW

Before commencing the research project, B.R.P. and Hilde Sommerseth (H.S.) arranged a Zoom meeting with the six reviewers who had given oral consent to participate. The reviewers were informed about the project goals, and that any data gathered would only be used for this research project and then deleted. Additionally, they were informed that Lars Ailo Bongo (L.A.B.) and B.R.P. would conduct interviews after the completion of the manual work. All reviewers consented and were given the opportunity to ask clarifying questions. As the participants were aware of each other's involvement, there was no anonymity among the reviewers. To ensure the interviewed reviewers' confidentiality, access to the interviews was restricted to the interviewers and the paper's authors who analyzed them.

We used qualitative interviews to gain insight into how each reviewer experienced the work, workflow, and the use of our custom tool. When the seven reviewers (including B.R.P.) had completed their manual verification and correction work, L.A.B. and B.R.P. interviewed the other six reviewers. We utilized a semi-structured interview guide, prepared with four main questions, to ensure the acquisition of data believed to be the most interesting for this project. The interviews were conducted in Norwegian, with the interview guide questions translated into English as follows:

1. How did you experience the task?
2. Did you find some part(s) more challenging than the rest?
3. What are your impressions of using the custom annotation tool?
4. You were informed in advance that you would be timed when performing this task, did that impact your work in any way?

We also asked follow-up questions when relevant. During the interviews we took notes, and after each interview the notes were reread, and additional information was added. We also discussed the interviews and shared our first impressions of them, particularly when it came to the topics and themes that seemed most important to the individual reviewers. After all interviews were conducted, the compiled notes were analyzed to identify key findings.

3 RESULTS

3.1 CORRECTED MACHINE LEARNING LABELS

In the dataset of 90,000 images, reviewers corrected the model's labels in 62.8% of cases. In 31.9% of the cases, they agreed with the model's labels. Reviewers could not confidently label 0.2% of the images, while 4.8% received some form of uncertainty symbol. Most of these cases are believed to be resolvable programmatically. For example, if reviewer A labeled the image '531' and reviewer B labeled

it '531@537', we are confident that the label should be '531'. For the remaining 0.3% of the images, reviewers provided invalid input, such as entering two codes in the same text box, or they agreed with the model on implausible labels, like '5bb'.

3.2 INTRACODER AGREEMENT

Each reviewer got, on average, 179 duplicate images from their own set of images. Three reviewers assigned the exact same label to all their duplicate images, while the remaining four reviewers labeled two to three images differently. Some of these differences are the results of reviewers not fully entering the second label — for instance, labeling an image as '555' initially and '55' for the duplicate — or when they reordered the digits, such as writing '861' for the first label and '168' for the second. These results indicate a high level of consistency among the reviewers.

3.3 INTERCODER AGREEMENT

We found that 86.4% of the images reviewed by two reviewers were categorized as *Certain and agreed*. Both reviewers agreed with the machine learning label for 33.8% of these *Certain and agreed* images, which results in both reviewers agreeing with the model for 29.2% of all images that were assigned to two reviewers. 8.9% are categorized as *Certain and disagreed*; for 44.9% of these, one reviewer agreed with the machine learning label. 4.5% belonged to the *Uncertain and agreed* category and 0.2% of the images were in the *Uncertain and disagreed category*.

The results show that in 8.9% of cases the reviewers disagreed with each other, and at least one of them was therefore wrong. Since the images in this manual verification dataset are some of the hardest images to label, we conclude that an error rate of 8.9% is good enough, and so a second reviewer is not needed.

We also found that in these overlapping reviews, only 34% of images had at least one reviewer agreeing with the model's labels. We believe these results combined show that these particular images are challenging to label, both for humans and the machine learning model.

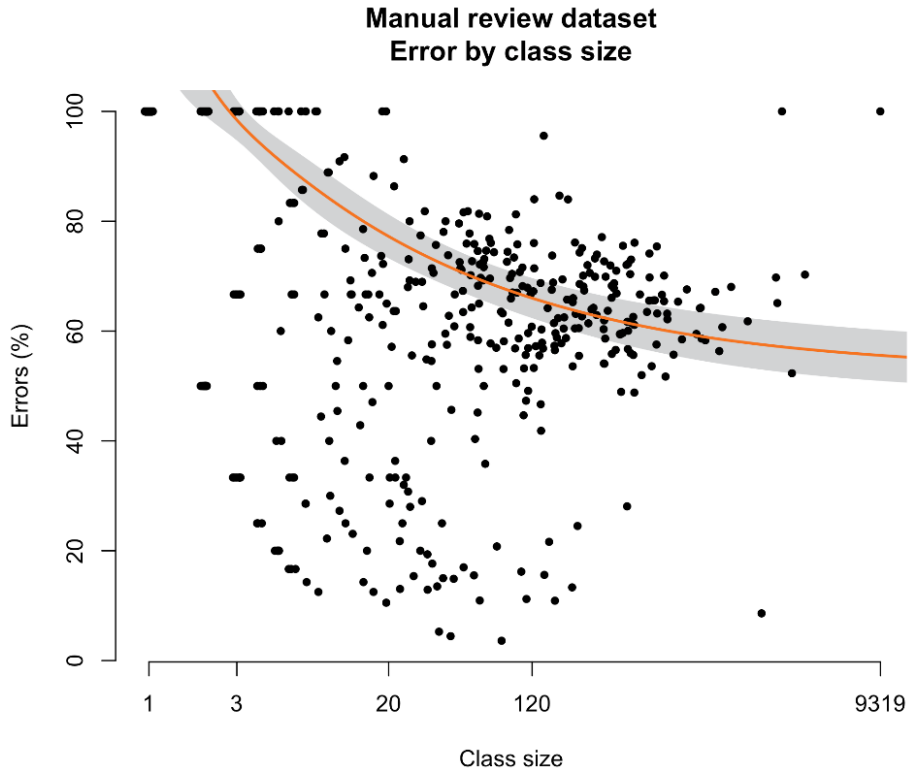
3.4 TIME USAGE

We examined the time spent reviewing images throughout the task and found no clear trend. As expected, some reviewers spent more time at the start of the project period, most likely to get accustomed to the annotation tool. The results suggest that reviewers' time spent did not decrease towards the end of the tasks, indicating consistent label quality over time.

3.5 MACHINE MISCLASSIFICATION

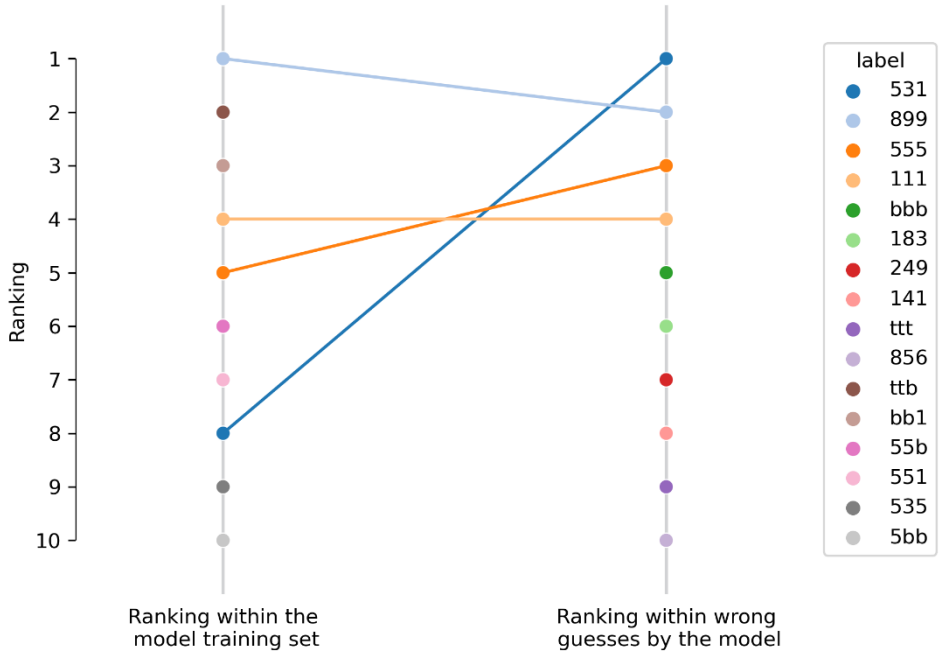
Our model's misclassification rate is higher for the smaller classes (Figure 3). For some of the smallest classes, the model has a misclassification rate of 100%. The same is true for two of the largest classes as well ('blank' and 'text'), these were included in the model to handle cases where the process of extracting images from the census pages did not manage to cut only images containing a code (Pedersen et al., 2022). In this dataset, the model always predicted some combination of 'blank'/'text' and digits for these images, most likely because of the noisy nature of an incomplete image extraction, and never just 'blank' or 'text'.

Figure 3 *Scatter plot for the classification error rate of the model*



Note: The x-axis represents all classes ordered by size, with some class size values shown. The regression is a natural spline with six knots that shows the general trend. We assume that all human labels are correct.

Figure 4 *Rankings by frequency of misclassification and frequency in training set*

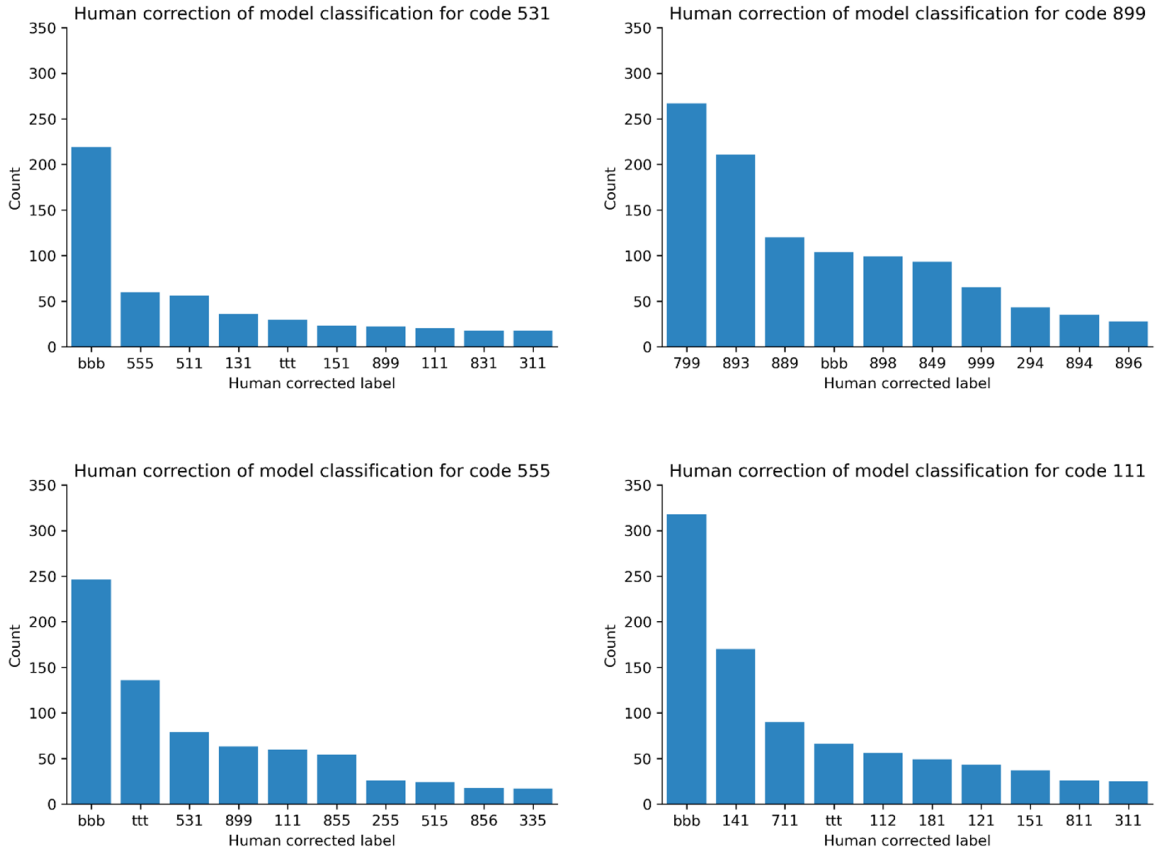


Note: The rankings of 10 most frequently misclassified labels (left) and the 10 most frequent classes in the model training set (right). They are connected by lines to help make it easier to see the labels' placements within the two datasets. Labels without a line connecting them were only found in one of the datasets.

We analyzed the labels found both in the model's misclassifications and the model's training set. The reviewers wrote down 549 unique codes. The training set has 286 codes, and there are 339 official occupation codes. There is some overlap between the training set and the model's misclassifications (Figure 4). For example, label '531' ranks as the most frequent in the model's training set (right column) and the 8th most frequently misclassified label (left column). In total, four of the largest classes in the training set — '531', '899', '555', and '111' — are all among the most frequently misclassified labels of the model. This is as expected since we assume the model is more likely to predict a large class for a prediction where it is not confident. However, the smallest classes exhibit a different pattern. After examining two of the smallest classes, '324' and '333' where the model had a 100% misclassification rate, we found that the model would most often predict a label that resembled the codes in appearance or were made up of the same individual digits but in a different order. To illustrate, for the code '324' the model predicted, from most often to less often: '224', '334', '384', or '534'. For the code '333', the model would predict: '533', '535', '353', or '323'. The human reviewers can also make this type of mistake, as the differences between certain similar digits can disappear depending on the handwriting of the original annotator or through images that were not cropped properly.

We investigated which codes the model incorrectly classified as one of the four largest classes in the training set (Figure 5). Digit similarity is one of the main indicators for the model's labels, but we also find that several of the images that were misclassified as some of the biggest classes were in fact images that should have been labeled 'blank' or 'text' ('bbb' and 'ttt' respectively in the figures). This could indicate that, if the model was unsure about a label it would try to find the closest match, but if there were no actual digits in the image to base this comparison on the model defaulted to a big class. Similarly, when faced with multiple similar labels, the model is likely to choose the label from largest class.

Figure 5 *Re-classifications done by the reviewers for the four biggest classes in the training set for the model.*



3.6 EVALUATION OF THE ANNOTATION WORKFLOW

The reviewers experienced the task differently. Several found the task tedious and not something they would like to do often. However, they did not mind doing it occasionally or in shorter intervals as a break from their regular tasks of transcribing handwritten text in church books or other census information. Some felt that they contributed to an important and relevant project which in the future may impact their work. Some also found the task very enjoyable and fun; one reviewer said it suited their personality better than their other tasks and that they could happily do this type of work more often. We believe that the reviewers' positive attitudes reflect that we succeeded in informing and involving the reviewers in this research project.

The reviewers reported that the main challenge in the labeling was caused by inaccuracies in the image extraction algorithm, leading some images to not be correctly cut from the census pages, making the missing parts harder and more time-consuming to interpret. Other issues consisted of codes that had been rendered unreadable due to excessive correction or overlapping writing, and codes that were too similar to accurately be told apart. The professional transcribers are used to reverting to the uncut source when they felt it necessary, and therefore missed this functionality in the annotation tool.

We found that some reviewers had an internal quality control when they encountered images that were hard to discern. The professional transcribers discussed these challenging images with their colleagues, as they do when transcribing other sources. We did not find the same behavior among the less experienced reviewers. However, some thought that they were not allowed to do so since this was a research project. None of the reviewers expressed concern about being timed during the manual review. Instead, they focused on making sure their labels were correct, thereby prioritizing quality over quantity. This is an important finding, since it strongly suggests that if quality control is a part of the work culture there is no need to implement a process for quality control, and therefore also not a need for a tool that supports such a process.

All reviewers had positive experiences using the custom annotation tool, finding it fast and easy to use. Two reviewers encountered and reported a bug that was fixed within 24 hours. The only feature they requested was a clearer indicator of overall progress, such as a line of text in the tool specifying the number of images completed and remaining. We believe that this shows that it is possible to use a special purpose tool that can be implemented in a short amount of time.

Almost all reviewers said it was easy to determine if the code in the image differed from the model's label, at least in cases where the code was not obstructed from view. They were confident in their own relabeling, typically performing a final overview of all images on the current page before proceeding to the next. However, one reviewer was concerned that she may be influenced by the suggested labels by the machine learning model.

During the interviews most reviewers talked about images they had found challenging to code and expressed concerns about their interpretation of the Histform coding rules. They also commented which numbers and codes the automatic transcription seemed to make mistakes with. We believe this first point shows a need for better training for the reviewers before starting the task, for example by doing a tutorial. The second point shows that the reviewers can provide input to an error analysis or help tune the model.

When asked if they thought they had worked more efficiently in the beginning, middle, or end of the project period, several of the reviewers felt that they worked more efficiently towards the end of the project. They said that it took a while to understand how the Histform standard's rules would apply to the different images. Once they learned the rules, they felt their pace increasing. The reviewers believe that they could have worked faster if they were told that speed was important, but most felt that they wanted to put the emphasis on the quality of the work. We therefore believe that the quality of the reviews is consistently high.

4 DISCUSSION AND CONCLUSION

4.1 MAIN FINDINGS

Seven reviewers examined a total of 90,000 occupation code images using our custom annotation tool. 62.8% of the images were corrected. Additionally, we randomly extracted 9,000 occupation code images, which were reviewed by two people, and compared the inter-reviewer results. For images with two reviewers, 68.1% were corrected. They had an inter-review agreement of 86.4%. This high degree of agreement shows that it will be enough to have one reviewer per image. It also means that a better use of human resources is to have fewer reviewers conducting the manual review with no overlapping images, and any extra resources could be spent on annotating a larger training set during the beginning of a project.

We see that the humans corrected the model in approximately 66% of cases. This indicates the confidence score threshold we set in our previous paper (Pedersen et al., 2022) was appropriate. Of course, there will always be some uncertainty as to what extent the accepted 2.2 million codes are in fact correct within the accepted error rate of <3%, despite the model giving a high confidence score. However, our results lend credence to our selected threshold value.

We found that our model is biased towards classes with a higher frequency and has a misclassification rate of up to 100% for the smallest classes. This proves that it is important to conduct this kind of manual validation and correction check, especially for imbalanced datasets in which the smallest classes have few samples. Without a manual review, several occupations would have vanished from the official records. However, we have no way of finding the codes from small classes that are misclassified as one of the frequent classes.

We explored the users' experiences of the work and the custom annotation tool using semi-structured qualitative interviews. We learned from the interviews that the reviewers prioritize quality over speed, and that some of the reviewers perform an internal quality control as part of their normal tasks. We believe that this behavior should be encouraged among all reviewers, especially if those with less experience are working together with the more experienced ones. This approach further reduces the need for multiple persons per image. In addition, we found that the reviewers can record uncertainty as part of the labeling task (as also suggested in Lu et al. (2022)). This recorded uncertainty can be used for additional control of the most difficult analysis or assist researchers in analyzing the transcribed data.

Several of the reviewers told us that the provided set of instructions for how to operate the annotation tool was very useful, especially in the initial phase, but suggested that including more concrete examples for labeling various problem cases would have been beneficial. So, there should be detailed instructions that take edge cases into consideration. The reviewers also pointed out that some of the most common problems when reviewing the images were bad cropping of the image from the original census page and excessive correction, both of which made it difficult to read the code. We plan to improve the automated cropping in future projects and introduce a new class for overly corrected images to flag them for manual review straight away. We also found that even if the task itself is perceived as boring, doing it as part of a short-term project is meaningful and interesting. We found that our custom tool had the functionality and user-friendliness needed for the task. We therefore do not see a need to use commercial labeling tools. Finally, we believe that the interviews show the importance of close collaboration between the machine learning developers and the annotators.

4.2 LIMITATIONS

Our review results show that there are many errors for the smallest classes. However, we only reviewed the images for which the machine learning model had lowest confidence or those assigned an invalid label. It may therefore also be interesting to do a similar evaluation of the errors and human effort for additional images. For example, the images with 3–6% lowest confidence and a randomly selected 3% of images with high confidence.

The occupation codes were added to the original census sheets by Statistics Norway after the census had been concluded. They also summarized all occupation codes used in a list. We did not constrain

our reviewers to follow this list, since we did not want to introduce bias and because we have found that the list is incomplete.[2]

We have not used the corrected labels to retrain the machine learning model. We have also not evaluated an iterative review process in which the reviewers use an active learning approach (Bernard, Zeppelzauer et al., 2018b). However, it may be useful to use information from a first review round to find the codes that are often mistranslated by the model, or to review the images assigned the least frequent codes.

The findings from the interviews are limited to reflect the views of the six reviewers that participated in this study. However, even themes expressed by a single individual can be relevant to the overall interview findings.

4.3 CONCLUSION

In this article, we conducted manual validation and correction of 90,000 images that had been automatically transcribed by a machine learning model. The aim was to find a simple and efficient method for this manual work, while maintaining high accuracy for the image labels, as they will be added to the Norwegian Historical Population Register. We learned the following lessons which we believe can be used as guidelines for efficient manual verification and correction in machine learning-based transcription projects:

1. The model's classifications are biased towards the largest classes, with the misclassification rate increasing for smaller classes, so manual verification and correction is necessary to improve the transcription accuracy for the smallest classes.
2. The reviewers can be trained to use an inherent quality control during the verification process, and they should encode uncertainty into the labels by using specific uncertainty flags. For this type of image classification, one reviewer per image is enough, as we demonstrated high inter-reviewer agreement. Therefore, the additional multi-annotator features of commercial state-of-the-art data labeling tools are not needed, and custom annotation tools can be quickly implemented and used instead.
3. The reviewers should be involved in the planning of the work so that the tasks are experienced as an interesting break from their usual tasks. However, the amount of work should not be overwhelming, and it should be easy to see progress in the form of percentage of images that are done within the graphical user interface.

Combined with our previous work (Pedersen et al., 2022), this paper contributes an end-to-end pipeline for highly accurate automatic transcription of tabular text data including an efficient manual review process for text that cannot be automatically transcribed. We have shared both the dataset and the code required to create the custom annotation tool, and we believe our solution can be applied by other groups working on transcribing tabular data.

DATA AND CODE AVAILABILITY

The 90,000 reviewed images and labels are open access (CC0 license). These are available at https://doi.org/10.18710/LYXKN1.

The manually annotated occupation code training dataset is open access (CC0 license) and available at https://doi.org/10.18710/7JWAZX.

The code for our custom annotation tool is open sourced using the MIT license. It is available at https://github.com/HistLab/More-efficient-manual-review-of-automatically-transcribed-tabular-data.

2 We have not been able to find the reason for why the extra occupation codes were added to the census.

ACKNOWLEDGEMENTS

Thanks to the Målselv transcription team and the three research assistants/students at the Norwegian Historical Data Centre for their participation in the project.

FUNDING

This work was funded by UiT the Arctic University of Norway through the interdisciplinary strategic project High North Population Studies and funding provided by the Norwegian Research Council (project number 322231).

REFERENCES

Bernard, J., Hutter, M., Zeppelzauer, M., Fellner, D., & Sedlmair, M. (2018a). Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 298–308. https://doi.org/10.1109/TVCG.2017.2744818

Bernard, J., Zeppelzauer, M., Lehmann, M., Müller, M., & Sedlmair, M. (2018b). Towards user-centered active learning algorithms. *Computer Graphics Forum, 37*(3), 121–132. https://doi.org/10.1111/cgf.13406

Cohen-Wang, B., Mussmann, S., Ratner, A., & Ré, C. (2019). Interactive programmatic labeling for weak supervision. *Proceedings of the KDD DCCL Workshop, Anchorage, AK, USA*, 4–8.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), 1–33. https://doi.org/10.1145/3561048

Hung, N. Q. V., Thang, D. C., Weidlich, M., & Aberer, K. (2015). Minimizing efforts in validating crowd answers. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 999–1014. https://doi.org/10.1145/2723372.2723731

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help me help the AI": Understanding how explainability can support human-AI interaction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581001

Liu, S., Chen, C., Lu, Y., Ouyang, F., & Wang, B. (2019). An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 235–245. https://doi.org/10.1109/TVCG.2018.2864843

Lu, Y., Chang, C.-M., & Igarashi, T. (2022). ConfLabeling: Assisting image labeling with user and system confidence. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late breaking papers: Interacting with eXtended Reality and Artificial Intelligence* (pp. 357–376). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21707-4_26

Pedersen, B.-R., Holsbø, E., Andersen, T., Shvetsov, N., Ravn, J., Sommerseth, H. L., & Bongo, L. A. (2022). Lessons learned developing and using a machine learning model to automatically transcribe 2.3 million handwritten occupation codes. *Historical Life Course Studies, 12*, 1–17. https://doi.org/10.51964/hlcs11331

Xiang, S., Ye, X., Xia, J., Wu, J., Chen, Y., & Liu, S. (2019). Interactive correction of mislabeled training data. *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 57–68. https://doi.org/10.1109/VAST47406.2019.8986943