Master's Thesis – Master Sustainable Development – Utrecht University

Reconstructing land use change from archival records

Using text mining, large language models and GIS to reconstruct land use change processes in former Dutch colonies

---

Quantifying historical land use change (LUC) can help estimate historical contributions to greenhouse gas emissions and biodiversity loss. Additionally, understanding LUC in former colonies can help fairly distribute the remaining carbon budget. Historical LUC is often quantified using a modelling approach and historical data is only sporadically used. To improve these models, in this thesis the archives of the Dutch East India Company (Vereenigde Oost-Indische Compagnie, VOC) are assessed to find information on land use. Additionally, a text mining methodology is developed for the identification of such information from transcribed archives of the VOC. A total of 14 lists concerning land use were identified with the text mining methodology. For Ceylon (modern-day Sri Lanka) a land use reconstruction is made using one of these lists. The list is georeferenced onto a contemporary map allowing 34% of the locations in the list to be connected to a location on a map. Additionally, two historical maps on land use were georeferenced, offering another data source for the estimation of land use. By identifying more lists and improving the georeferencing procedure, future research can disentangle the drivers for LUC and make land use estimates for all the colonies of the VOC. This thesis offers the foundations for such an approach and is a starting point for improving our understanding of historical LUC.

Name: Yannick Egberink

Student number: 6428231

Supervisor: Dr. Brian Dermody

Second Assessor: Dr. Janpieter van der Pol

1 Introduction

Land use change (LUC) such as deforestation and urbanisation attracts a lot of attention in both society and scientific literature (Gilg, 2009; van Vliet et al., 2016; Winkler et al., 2021) as it is a key driver of greenhouse gas emissions and biodiversity loss (Luyssaert et al., 2014). Most literature and news items focus on contemporary deforestation in countries such as Brazil and Indonesia (Arima et al., 2014; Austin et al., 2019; Den Elzen & Schaeffer, 2002). However, LUC is not new and has occurred throughout human history. Studying historical LUC provides insights into the environmental transitions that have happened throughout the world (Foley et al., 2005), and for the colonial and post-colonial periods are a crucial data source to estimate historical contributions to greenhouse gas emissions and biodiversity loss (IPCC, 2023). In addition, land use (LU) shapes society's views, norms, and values, and historical reconstructions of LUC can therefore help understand cultural heritage (Plieninger et al., 2006), and aid in the resolution of conflicts (Sterk et al., 2011). Previous studies have quantified historical LUC using modelling approaches such as HYDE3.2 (Goldewijk et al., 2017), LUH2 (Hurtt et al., 2020), and SAGE cropland (Ramankutty & Foley, 1999). Historical documents are only sporadically used to inform the parameters of these models (Winkler et al., 2021). Therefore, analyses using solely historical documents could offer additional insights into these models and might be used to validate the outcomes of these approaches and improve the mechanisms for modelling land use change.

One of the earliest colonial powers was the Dutch Republic whose main instrument for trade and colonisation was the Dutch East India Company (Vereenigde Oost-Indische Compagnie, VOC). The Dutch East-India Company was established in 1602 and until the 19th century controlled most of the colonial trade in Asia (see Figure 1 for a map of the trade outposts of the VOC). By establishing colonies, the VOC controlled the production of commodities, often agricultural products (Weststeijn, 2014). It is yet unclear what the impact of the VOC on land use was, however, the documentation on the operations of the VOC could offer insights into the process through which the VOC impacted LU in its former colonies.

The archives of the VOC contain a variety of documents, ranging from descriptions of indigenous medicine (Scott & Hewett, 2008), to comic books (La Nasa et al., 2019). The archives also contain information that concerns the general operations of the VOC, such as shipped goods, population, and maps of the controlled territories (Meilink-Roelofsz, 2023).

However, the size of the archives and the absence of ex-ante knowledge on where relevant information regarding LUC can be found and what this information might look like, makes it infeasible to analyse the documents by hand. Therefore, a text-mining approach is used to find information relevant to land use change.

Advancements in Artificial Intelligence (AI) and increased digitisation of historical documents make it possible to apply text mining approaches to analyse large amounts of historical documents (Colavizza et al., 2021). This allowed studies on for example historical trade with the British Empire (Hinrichs et al., 2015), LUC in the Mediterranean (Cervera et al., 2019), and environmental change in the English Lake District (Smail et al., 2021). These approaches often require a deep understanding of AI and text-mining to adequately apply the methods (Greene, 2024). In addition, many language models and text mining methodologies have been developed for the English language, making studies on documents in other languages challenging (Medagoda et al., 2013). The increased availability of pre-trained Large Language Models (LLMs) creates new possibilities for researchers without a strong background in computer science and helps research documents written in languages other than English.

This study aims to use archival records from the VOC to make reconstructions of LUC in former colonies of the VOC. Accordingly, three types of data are analysed: trade data, maps, and textual data. All three can be found in the archives of the VOC, however, to make a clear distinction between the data sources, when the VOC archives are mentioned in this study, it refers to the textual data only. The focus of the research on LUC is on Ceylon, since LU maps are available from the time of the VOC (l'Honoré Naber & Boink, 2023), and there is documentation of changes in LU during the rule of the VOC (Kanapathypillai, 1969).

Trade of agricultural commodities is a driver of LUC (Jadin et al., 2016; Meyfroidt et al., 2010). Therefore, analysing trade patterns could be a proxy for LU in the times of the VOC. For this study, the trade data from Huygens (2022) is used as this includes many regions and therefore would allow for applying the methods to other regions in future research. The analysis of the trade data is not confined to Ceylon and rather is limited to the nine commodities which are specified in the methods (3.1). In this thesis, the quality of this trade data is assessed and the usefulness for future research is discussed, but it is not used to inform the LU reconstruction, due to time constraints.

The second aim of the research is to develop a methodology to identify information on LUC in the VOC archives using text mining in which no ex-ante knowledge of the content of the archives is needed. The research describes different approaches and assesses them based on the speed of the process, and the quality of the outcomes. The approach that was deemed the most useful is then described in more detail. To test the methodology, three geographic regions are used: Ceylon (now Sri Lanka), Ambon (part of the Mollucas), and the Banda islands (also part of the Mollucas). Applying the methodology to different regions and not only the Ceylon region ensures more general applicability of the methods.

The two aims of this study give rise to the following central research questions and sub-questions:

RQ1 What is a reliable and generalisable method for identifying relevant documents on land use change in the archives of the VOC without ex-ante knowledge?

1. Do automated approaches such as large language models (LLMs) allow the identification of places and commodities in historical documents to reconstruct land change and trade?
2. In what ways do automated approaches need to be complemented by other methods to answer RQ2?

RQ2 Is it possible to reconstruct land use change in former colonies of the Dutch VOC using historical archives?

1. Can quantitative reconstructions of traded agricultural commodities be made using historical sources?
2. Can land use change be reconstructed from historical sources?
3. Can the processes of colonial land change be understood by analysing and integrating historical sources of land use change and trade?

The goals of this thesis and the research question are inductive, meaning no specific method was established beforehand. Since no previous research has addressed a similar task, the methodology was developed uniquely for this study. This also means that sometimes pragmatic choices are made, i.e. if something works, it works. Therefore, the approach taken in this study can be refined in the future, and potential improvements in the methodology are an important outcome of the study.

In the following section, a short historical background on the VOC, their archives, and Ceylon is given, after which the method is discussed, followed by the results and the discussion. In each section, first the text mining is discussed, then the trade reconstruction and lastly the LU reconstruction.

2 Historical Background

The VOC was established in 1602 from a fusion of six smaller companies (Karnadipa et al., 2023). Previously, the six companies had competed for the profits of trade voyages to Asia. However, because the Dutch Republic was at war with Spain and Portugal, the government forced the companies to unite to play a part in this war (Meilink-Roelofsz, 2023). The ships of the VOC were used to fight against the Spanish and Portuguese, and the profits of the operation were used to further increase the company's fighting power: in some ways the VOC operated more as a country than as a company (Weststeijn, 2014). The six former companies formed six "chambers" led by the heads of the former companies, with a general board (Heren XVII) that had representatives of all the chambers. The archives are also structured according to these chambers named after the city in which they were founded, as they all kept their own records (Meilink-Roelofsz, 2023).

After the institution of the VOC, it quickly became the largest trading company in Asia. It was the first company to leverage the capital market to secure funding for its operations, which enabled the VOC to establish itself as the largest and richest company of the time (Nijman, 1994). The VOC even had the right to engage in military operations against countries and other entities to also establish a monopoly on trade outside of the Netherlands (Meilink-Roelofsz, 2023). In the Netherlands, the VOC got the monopoly on Dutch trade with Asia by right of Octrooi (a sort of patent), which made the VOC the only Dutch company legally allowed to trade with Asia (Weststeijn, 2014). The VOC, however, established monopolies that went beyond this Octrooi, making the VOC the only company that traded monopolised commodities with the whole of Europe, not only the Netherlands. To enforce these monopolies, the VOC formed colonies in many countries including in what is now known as Indonesia and India (Figure 1).

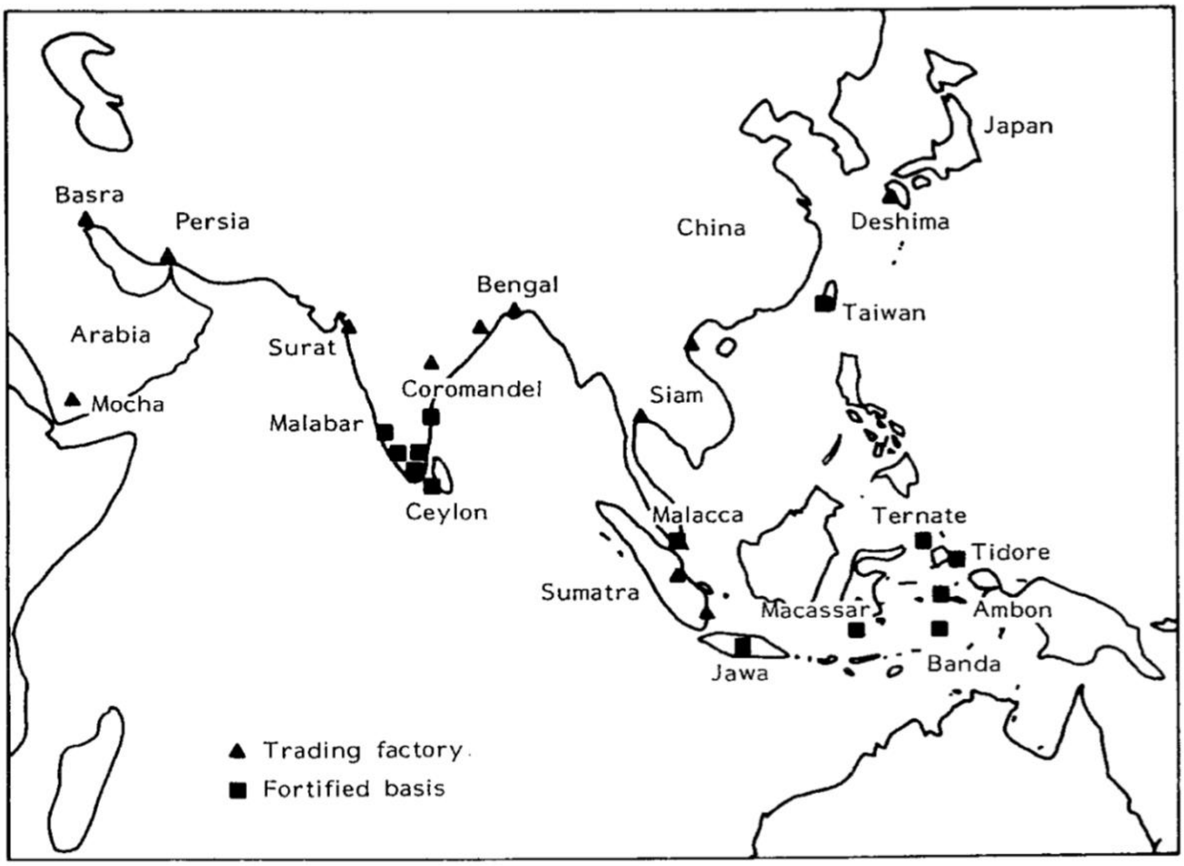


Figure 1: Largest outposts of the VOC in Asia during the 17th century. Some places were gained and lost during this period. The VOC also had smaller settlements inland of these colonies, however the map gives an overview of where in Asia the VOC was active. Source: Nijman (1994)

Initially, the VOC established monopolies for trade in fine spices. The capture of the Banda islands in 1622 gave the VOC the monopoly on the trade of mace and nutmeg, and by destroying plantations in the Moluccas and the capture of Makassar it achieved a monopoly on the trade of cloves (Meilink-Roelofsz, 2023). Lastly, a monopoly on the trade of cinnamon was established by fighting off the Portuguese in Ceylon (Meilink-Roelofsz, 2023). At the end of the 17th century, European trade with Asia changed to other, larger commodities (Nijman, 1994). Textiles from India, coffee from Arabia, and tea from China became more important than fine spices. The increase in the size of the products also meant that more ships were needed for transportation, and they required larger growing areas, making it much harder to maintain monopolies (Meilink-Roelofsz, 2023). Other companies and countries were also able to establish trade routes to Asia, and the relative importance of the company decreased (Rossum, 2011).

The VOC did not only establish trade routes between Asia and countries in Europe. It had a large trade network within Asia as well, transporting four to five times more goods within Asia compared to the trade with Europe in the 17th century (Rossum, 2011). This intra-Asiatic trade was an important source of income in the early years of the VOC and was the most profitable part of the operation during this period (Meilink-Roelofsz, 2023). The intra-Asiatic trade, however, steadily diminished in importance to the VOC as it became less profitable due to difficulties with the availability of ships and crew (Rossum, 2011).

The increase in importance of other companies, most notably the British East-India company, and the decreased profitability of the intra-Asiatic trade, diminished the financial reserves of the VOC (Nijman, 1994). A war between the Netherlands and England in 1780 led to a big loss of returning ships (Rossum, 2011), and because the travels of the ships were paid for with the prospect of income when the ship returned, the VOC acquired high debts (Meilink-Roelofsz, 2023). On January the 1st 1800, the right of Octrooi on the trade monopoly with Asia was not extended, and the VOC was nationalised. This marked the end of the VOC.

2.1 Archives

As mentioned in section 2, the different chambers and the Heren XVII held their own archives in the Netherlands (Meilink-Roelofsz, 2023). In addition to the archives in the Netherlands, parts of VOC archives are held in many other countries, for example: Indonesia (Karnadipa et al., 2023), Ceylon (Paranavitana, 1999), and the Cape (Twidle, 2013). The archives in Indonesia are an especially wealthy source of information as they contain a vast collection of documents from the general administration in Batavia (now Jakarta), which was the seat of the central government of the VOC in Asia (Meilink-Roelofsz, 2023). These archives are currently being digitised but are not yet available for text-mining purposes and therefore not used in this study (Karnadipa et al., 2023).

In 2018, the project Ijsberg was initiated to transcribe a part of the Dutch VOC archives along with other historical documents, using hand-written text recognition (HTR) (Keijser, 2018). In 2021 this project was finished, with the result of 1.5 million documents transcribed and digitised (Nationaal Archief, 2021). After 2021, the transcription of these documents continued in the GLOBALISE projects, making many more transcriptions (> 4 million) available for the public (GLOBALISE, 2023). The documents that are made available through the GLOBALISE project contain the inventory numbers 1053-4454 and 7527-11024 from the VOC archives of the Nationaal Archief (index number 1.04.02), and an index of the content of the whole archives can be found in Meilink-Roelofsz (2023). These transcriptions are used in this study.

2.2 Ceylon

Ceylon was considered one of the most desirable islands for colonial powers. The cinnamon grown on the island was of the highest quality and occurred naturally in forests here (Barceloux, 2008). Even before the arrival of colonial powers, Ceylon's cinnamon was traded with countries in Asia and Europe. Between the 10th and 15th centuries, Arabs controlled most of the cinnamon trade and tried to keep the origin of the cinnamon a secret from the European countries (Ravindran & Nirmal Babu, 2005). The arrival of the Portuguese in 1505 marked the end of the Arabian-controlled trade (De Silva, 1967), followed by the Dutch in 1658, and the British in 1796 (Suriyagoda et al., 2021). In 1948 Ceylon gained independence and is now known as Sri Lanka (Ananthavinayagan, 2019).

During the VOC period, the cinnamon production in Ceylon changed dramatically. Before the arrival of the VOC cinnamon was harvested in forests while during the later part of the 18th century, the VOC established cinnamon plantations (Kanapathypillai, 1969). There was a strong belief that cinnamon could not be cultivated in plantations, which was only debunked in experiments between 1765 and 1785, and the first cinnamon plantations were therefore established after 1770 (Dewasiri et al., 2021).

Three maps of Ceylon from 1793 show these plantations (see Figure 2 for one of the maps and Appendix II for all). The maps specify six types of agricultural LU: cinnamon plantations (yellow), coffee plantations (red), areca palm plantations (orange/red), pepper plantations (black), areas that are cleared for planting (green), and areas for which it is unclear what is planted there (on the old maps specified as "*kyaten*", amber). These maps are a valuable source of information on LU and therefore play an important role in this study.

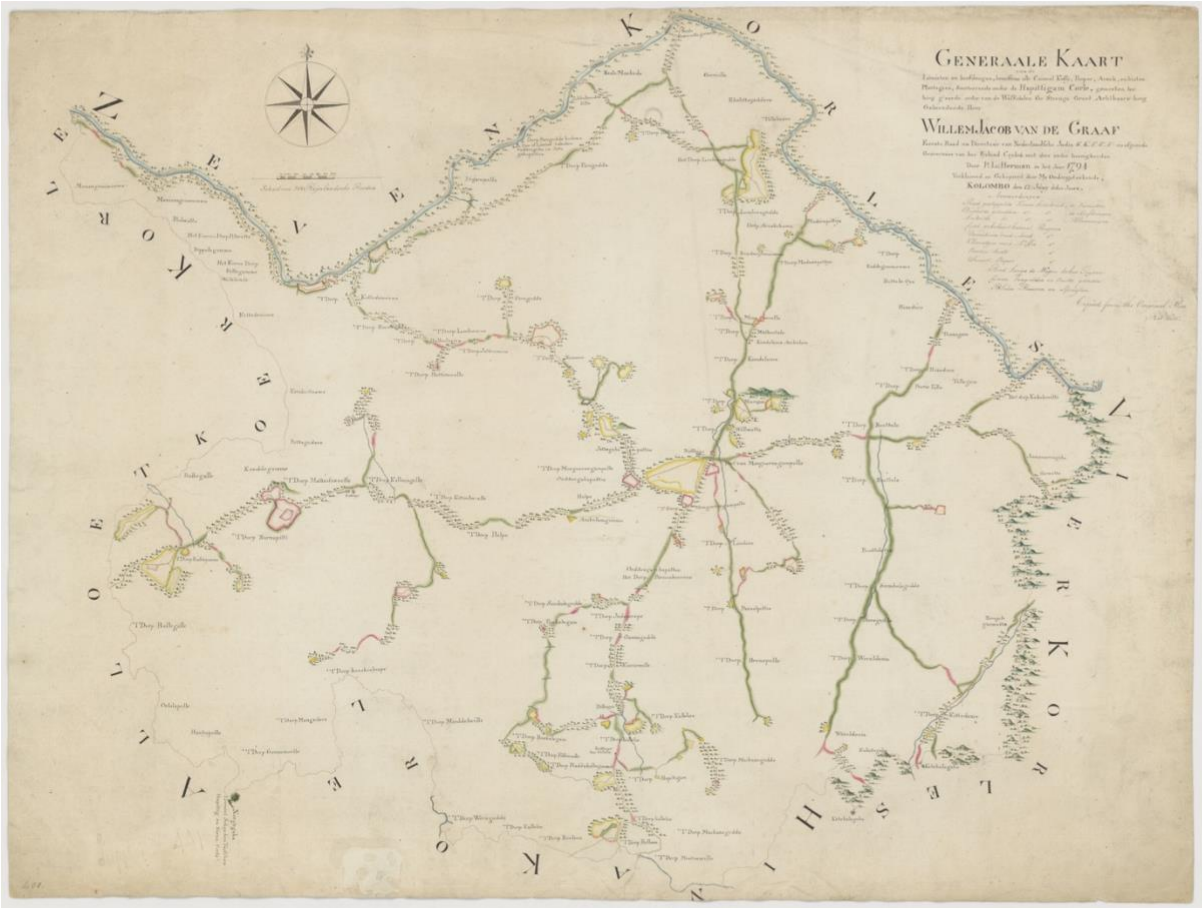


Figure 2: LU map from the Happitigam Corle in Ceylon.

3 Methods

To reconstruct LUC in Ceylon, the methods shown in the flowchart in Figure 3 are used. The method is divided into three aspects: analysis of the trade data (3.1), the text mining approach (3.2 & 3.3), and a combined approach using the outputs of the text mining and analysis of maps (3.4). Since no previous research has investigated LUC based on the archives of the VOC, the research was done using an inductive approach, where the outputs of previous steps informed the development of the methods. Therefore, to understand the methods, some results are briefly touched upon in the methods. For the text mining approach, both the considerations for developing the methodology and the used methods are presented. All data analysis was done in Python and the scripts and data used in this study can be found in Appendix VI, and the scripts can also be found on GitHub (https://github.com/Yegberink/VOC_land_use/tree/main).

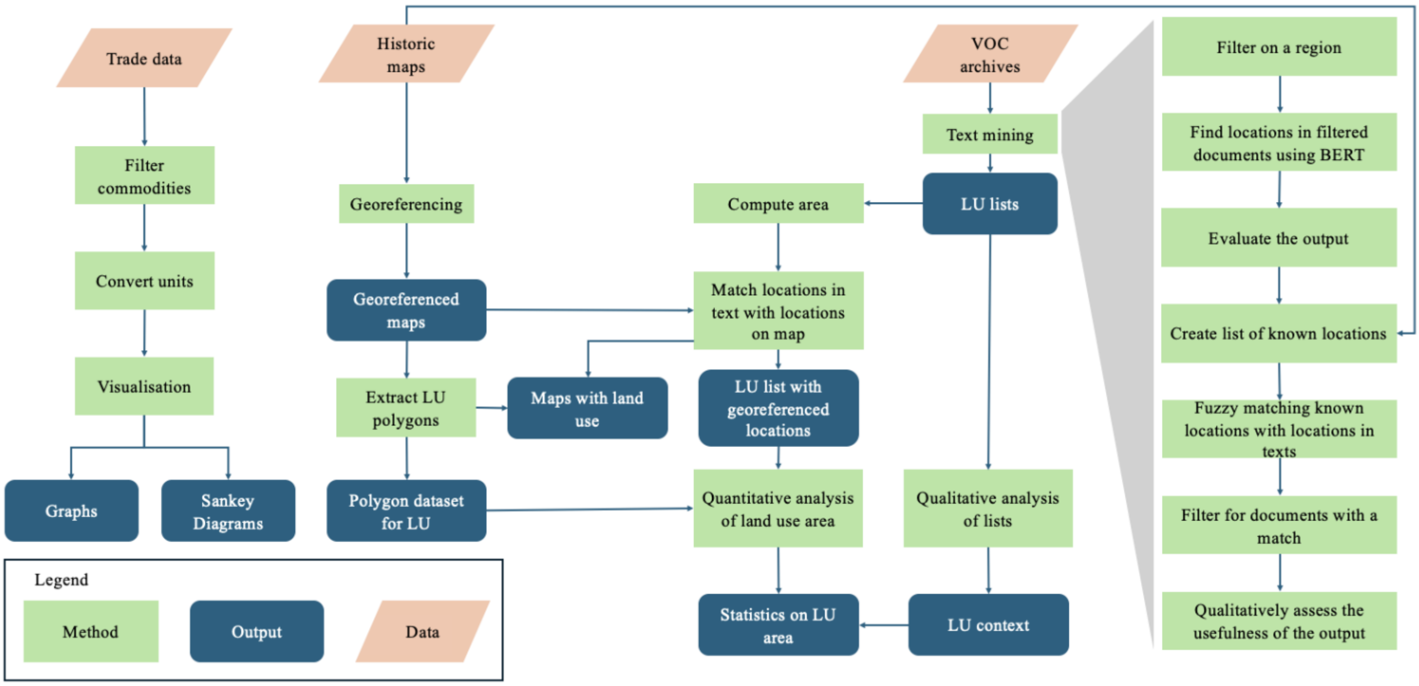


Figure 3: flow chart for the methods for reconstructing land use (LU) and trade patterns in former colonies of the VOC. Three data sources can be distinguished: trade data, historical maps, and the VOC archives. The text mining approach is depicted in more detail in Figure 4.

3.1 Trade Reconstruction

The VOC held extensive archives on the voyages their ships undertook, the cargo these ships carried, and the profit the voyages made (Huygens, 2022). These archives have been digitised and can offer insights into the trade flows of goods over time (Huygens, 2022). To add to the information in this dataset, in this study commodity names are standardised, the units are converted to kg, and the data is visualised using two different visualisations. Firstly, the overall size of the trade is visualised using line graphs, and secondly, the direction and size of the flows are presented in Sankey diagrams.

To make these visualisations, the total weight of the commodities transported on ships is used. This gives a dataset which specifies the year of the travel, the weight, the origin, and the destination of the traded commodity. Nine commodities (Table 1) were selected as it was not feasible to research all commodities in the trade data. The trade reconstruction is not restricted to the three regions that are the focus points in the rest of the thesis, because the trade data might show interesting trends for other commodities and locations and can help understand the historical context. Within the dataset, different words are used to describe the same commodity and by searching for the commodity name in the VOC glossary (Instituut voor Nederlandse Geschiedenis, 2000), synonyms for these commodities were found. A synonym in this thesis concerns either a linguistic synonym for the commodity name or if it is produced from the same source, for example “foelie” (mace) for nutmeg. The weights for the synonyms are simply added together to give an overall weight of the commodity. Products made from the commodities are not considered, for example clothing from silk fabric.

Table 1: Commodities and their different Dutch names as specified in the VOC glossary (Instituut voor Nederlandse Geschiedenis, 2000). *A synonym here concerns either a linguistic synonym or if the commodity is produced from the same source.*

Commodity	Dutch synonyms
Cinnamon	["kaneel", "boskaneel", "kaneel de matte", "candela"]
Nutmeg	["nootmuskaat", "muskaatnoten", "brouwersnoten", "foelienoten", "foelie", "rompen"]
Cloves	["antoffel", "blom", "caplet", "garioffelnagel", "nagel", "kruidnagel"]
Pepper	["peper", "staartpeper", "cubebe", "piper cubebe", "ruagiepeper"]
Opium	["amfioen", "opium"]
Silk	["zijde", "ablako", "agoni", "allegias", "ardasse", "ardassina", "armozijn", "baa", "bariga", "bogy", "cabessa", "cannagy", "capitoen", "casagy", "chauls", "coetchiaal", "cora", "dom", "floretzijde", "floszijde", "fora", "gangali", "gasen", "gert-kerckerie", "hittou", "kannegie", "kedgoda pessend", "ketsier", "khan baffy", "legia", "matiaal", "moghta", "mongo", "pangia", "parelzijde", "patteni", "pee", "peling", "poilzijde", "poolzijde", "poilzijde", "potti", "quetchoda-passant", "selvatica", "serge", "sjesum", "tabijnen", "taffachelas", "tafta", "tamut", "tanna-banna", "tanny", "tantianozijde", "thujas", "vloszijde", "volte corte", "zeem"]
Tea	["thee", "bing", "boei", "congo", "joosjesthee", "pecco", "songlo", "souchong", "tscha"]
Coffee	["koffie", "bun", "cauwa", "kauwa", "kitscha"]
Sugar	["suiker", "gula", "jagersuycker", "jagru", "poeyer", "stoksuiker", "kandijsuiker", "tamarinde"]

There were no standard units of weight during the time of the VOC and therefore different units of weight are used in the trade data. The weights thus have to be converted to compare the various commodities. The conversion of these old weights to kg is also based on the VOC glossary and the conversion table can be found in Appendix I. For some units, the glossary did not directly give useable numbers. In that case either another source is used, or an assumption is made on the conversion of these commodities. In both instances, this is specified in Appendix I. Some data points have been disregarded if no good conversion is possible, or if the unit is used less than three times in the dataset. This led to the inclusion of >98% of the data points for each of the commodities except for coffee (±85%) because it contains only limited data points, and silk (±30%) because for silk many data points have a unit for which no conversion was found.

For the line graphs, the total transported weight each year is used regardless of the direction of the trade. For example, if cinnamon is first transported from Ceylon to Jakarta and later imported back to Ceylon, this would add to the total weight transported. A more detailed analysis would be needed to disentangle these kinds of specificities.

3.2 Text Mining Methodology

The VOC archives are extensive, comprising millions of documents. The transcribed portion comprises 5 million pages, representing only a small part of the entire collection. Text mining is therefore used to identify documents that could contain information on LU. By extracting certain criteria from the text, large amounts of VOC documents can be filtered to allow for a qualitative assessment of the usefulness of the resulting documents. Ultimately the method should achieve the following goals: 1) filter out information that is not useful; 2) find quantitative data; 3) be fast enough to use on a large number of documents, 4) be applicable to multiple regions in which the VOC was active. Additionally, some consideration is given to the applicability of the approach to historical documents in other languages, although this is not tested in this thesis.

To filter the documents three criteria were evaluated: the occurrence of locations, the occurrence of commodities, and the occurrence of numbers. It was hypothesised that if these three occur together, they might contain either information on LU (e.g. sentences such as: in

Colombo there is a field with 900 cinnamon trees), or information on trade (e.g. sentences such as: From Colombo 900 kg of cinnamon was transported to the Netherlands).

The first criterion used for filtering is location. Since the VOC documents were generated using an AI model for HTR, the data contains many misspellings. Additionally, the texts contain different spellings of the same location. Therefore, it is impossible to simply match words with a known list of locations. To deal with this, two different approaches were assessed. The first is to replace the misspellings with a correct spelling from a list of known locations, and the other is to extract locations using Named Entity Recognition (NER) and then use fuzzy matching to find the known locations in a list.

For the first approach, a generative large language model (LLM) was used. LLMs are artificial neural networks trained on large quantities of text data (Meyer et al., 2023). By applying Transformer architecture these models can capture the meaning of words in the specific context of a sentence and the whole corpus (Vaswani et al., 2017). For example, the Transformer architecture makes it possible to identify whether *land* relates to landing a plane or a piece of land. Recently, generative LLMs have become available to the wider public with the main example being ChatGPT (OpenAI, n.d.). Generative LLMs can generate a textual output based on a textual input and could therefore be useful for correcting spelling mistakes in the texts. Llama2 is the best-performing open-source language model (Touvron et al., 2023), can be downloaded for free, and runs on a local device. Therefore, Llama2 was used to assess the potential of generative LLMs.

In contrast to generative LLMs, the second approach uses an LLM developed for one specific task, in this case NER. NER and especially Bidirectional Encoder Representations (BERT) are an often-applied model in text mining. The first BERT model was developed by Devlin et al. (2018) and has been adapted for other languages. The model ner_dutch_large (Schweter et al., 2020) from the Flair library was used in this study as it is an easy-to-use package, with high performance on Dutch texts. The model can be substituted with another model in the flair library so that the method can also be applied to documents in other languages. The ner_dutch_large model has a relatively large number of parameters meaning that it takes more time to run than smaller models, but it increases the performance of the model which is preferable for this thesis given the many misspellings in the VOC documents. The model runs with the Flair tokenizer, which splits sentences into words.

The output of the NER model is matched with the list of known locations using fuzzy matching. Fuzzy matching uses Levenshtein distance to compute the difference between two strings. One difference is described as one insertion, deletion, or substitution. For example, uninformed and uniformed have a Levenshtein distance of 1. With fuzzy matching, this is then transformed into a percentage of difference which can be used as a parameter for the maximum difference allowed to be considered a match. For example, two words can be considered a match when they are similar for more than 98%, which is computed in fuzzy matching.

To find the most useful text mining approach (NER or generative LLMs), a small sample (n=20) of the VOC documents was taken and the different approaches were applied. The outputs of this can be found in Appendix III. Initially, the approaches would be assessed on multiple criteria, however, most of the approaches were directly deemed impractical due to the long run time or gave clearly wrong results and therefore it was not necessary to assess these further. The run time of Llama2 was around 30 seconds per document, and with the large number of documents that had to be analysed, this was not a viable option. Furthermore, the output of Llama2 was very inconsistent and did not give useful results. The NER model performed better and gave appropriate results for the sample. Therefore, the NER approach was selected to develop the methodology further.

To check for the occurrence of numbers and commodities in the sentence, the tokenised sentences from the previous step are used. Since there are many mistakes in the transcriptions and differences in interpunction between current and historic Dutch, it is difficult to accurately split the texts into sentences. Therefore, in this thesis, a sentence is considered to be a string of 30 words around the occurrence of a location (15 on each side of the locations). To check for the occurrence of numbers, regular expression (regex) is used. This allows Python to quickly check if there are numbers in the sentences. It turned out that almost all sentences (±90%) contained some number either through a transcription mistake or for example a date. Therefore, this was not the most useful criterion. Finding commodities is more difficult given the different spellings and synonyms. Fuzzy matching was applied to identify the commodities in the text; however, this did not yield useful results. Depending on the fuzzy matching threshold either too many or no useful sentences were found. This criterion was therefore also not used for the final methodology.

Since two of the three criteria were not deemed useful, the approach of finding interesting sentences based on the co-occurrence of these was not applied. The final method uses the locations and filters these locations based on a list of known locations that are potentially interesting for LUC.

3.3 Applied Methodology

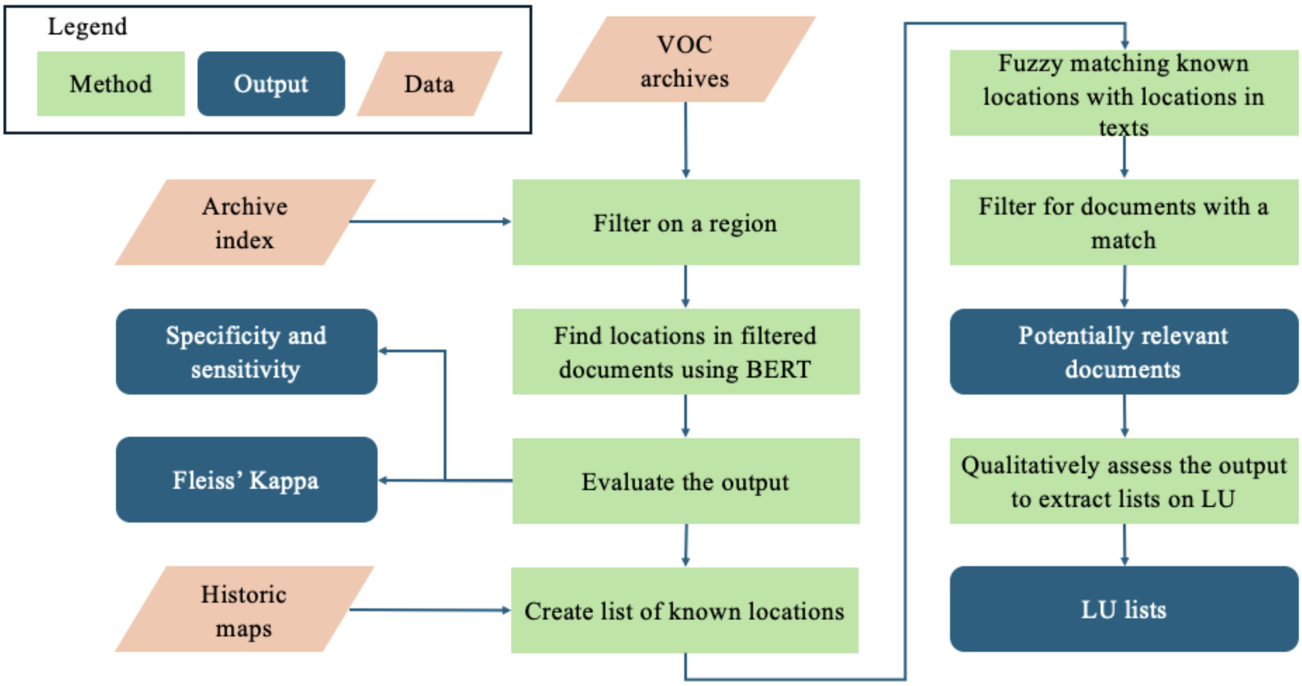


Figure 4: Detailed overview of the text mining methodology used to identify lists on land use (LU) in former VOC colonies with Python. An overview of the whole methods used for this thesis can be found in Figure 3.

The first step in the used methodology is to identify documents concerning one of the regions of interest (i.e. Ambon, Ceylon, and the Banda islands). The VOC documents are filtered according to a certain region based on the indexing done by the Nationaal Archief. All documents under the header of a certain region as specified in the index of Meilink-Roelofsz (2023) are used. This gives around 21000 documents for Ambon, 114000 for the Banda islands, and 72000 for Ceylon. By filtering the VOC archives in this way, it is possible to apply the large NER model without it taking too long to run.

The second step is to identify locations in the VOC documents using the BERT-based NER model ner_dutch_large, as explained before. For Ambon the runtime was around 10 hours, for Ceylon two days, and for Banda three days. After that, the output of the model is evaluated on the sensitivity and specificity of the results. Sensitivity relates to the percentage of occurrences a text mining algorithm picks up on, whereas specificity relates to the percentage of occurrences it categorises wrongly. For example, when creating the dataset of misspellings, a sensitive Llama2 algorithm might pick up on 95% of the misspellings of a certain location meaning 5% of misspellings were missed (false negative), however, 30% of the identified misspellings might be wrongly regarded as a misspelling (false positive). To minimise human error during the evaluation, multiple (4) Dutch-speaking evaluators are used, which is an often-applied approach in evaluating model outputs (Hendriks et al., 2021). These evaluators read over the sentences and assessed whether the algorithm missed and/or incorrectly classified an occurrence in this sentence. Afterwards, Fleiss' Kappa is calculated to indicate if the evaluation is likely to reflect the true sensitivity and specificity of the model. Fleiss' Kappa is a statistical measure used to measure agreement among observers (Landis & Koch, 1977). It therefore offers insights into whether the evaluation can be used to make claims about the performance of the model.

To apply the fuzzy matching in step five, a list of known locations is needed. For many of the locations important to the VOC, there are readily available lists (for example on vocsite.nl). However, these are mostly in the coastal regions, since the VOC was mostly interested in the coastal areas for trade and seafaring navigation (Guleij, 2017). In contrast, the production of the goods, which is interesting for LUC, happened more inland. Maps are therefore used to create a more detailed list of inland locations that can be used to match the locations found in the NER. These maps were found using literature review (Ceylon) or are taken from the Nationaal Archief section 4.VEL (Ambon and Banda), and all placenames on the maps are transcribed. The created lists and used maps can be found in Appendix II.

To filter the locations in the texts for the locations on the maps, fuzzy matching from the fuzzywuzzy package is used. The fuzzy matching is applied to the found locations and parts of the found locations using a sliding window. The sliding window ensures that if two words are pasted together due to mistakes during the HTR, this is still picked up. The sliding window takes each possible word embedded in the location without switching letters (for example for Banda: b, ba, ban, band, banda, anda, and, a). These elements are then compared to the locations in the list. The thresholds for all locations were set to 95% for the direct match and 98% for the matching of the elements in the sliding window. These thresholds were set high to ensure that only a few matches were returned, making it possible to qualitatively compare the outcomes. The second threshold is higher than the first because otherwise many wrong matches were returned. Some locations from the location list returned many matches, mainly obviously wrong ones. These were then excluded from the location list.

The sentences containing a relevant location are then linked back to the documents they occur in. This gives a dataset with documents that are potentially interesting for LUC. This dataset is assessed using the transcribed documents and images of the original handwritten texts. No pre-specified criteria were used, as it was unclear what the content of the documents could be, but it aimed to identify tables with numbers in the transcribed documents, and then check what these tables are about in the original documents.

3.4 Land Use Change Reconstruction

The analysis of the textual data for Ceylon yielded LU lists that contain information on LU in specific locations. To best use these lists, the locations mentioned in the text are matched with locations on a map. Placenames often change over time meaning it can be difficult to match a location on an old map or a historical text with a location on a modern map. The historical LU maps for Ceylon are therefore used to provide the background for matching the placenames from the texts. Consequently, the geographic location is limited to two regions: Hina Corle and Alloetcoer Corle, because only these two were both mentioned in the texts and there are historical LU maps preserved.

To extract most information from the maps, they are georeferenced onto a base map in a certain coordinate reference system (CRS). This is done by finding Ground Control Points (GCPs), points on the historical map that can be linked to a point on the base map (Affek, 2013). The maps (Appendix II, Figures 1 & 2) are then transformed using a Thin Plate Spline (TPS) transformation. TPS is an interpolation technique which can be used for scattered data (Keller & Borkowski, 2019). TPS is often used for georeferencing historical maps as the image is not scaled as a whole but is scaled variably depending on the distortion of the map in that specific region (Affek, 2013). It therefore corrects potential inaccuracies in the distances on the old maps. The maps are georeferenced onto the Kandawala / Sri Lanka Grid (EPSG: 5234) CRS, which allows for calculating polygon areas. From the georeferenced historical maps, a layer containing LU polygons is constructed. This is done by simply tracing the polygons on LU onto the CRS. The size of the polygons is then computed within QGIS.

From the VOC archives, lists on LU are extracted using the text mining methodology presented in 3.2. These lists contain both quantitative data on the size of the growing area and qualitative information on for example why a piece of land is cultivated (see Table 2). From the qualitative data, the context of the LU in the regions is assessed. This is again done using an inductive approach since the contents of these lists were not known beforehand. The size of the growing area presented in the lists is expressed in old local units: ammonam (written as ām) and a smaller unit the coernix. These units therefore need to be converted to hectares to be used in the further analysis. The ammonam concerns the area that can be sowed with a certain amount of wheat (Abkouw, 1765). Because this depends on the type of land, no direct conversion is possible but Abkouw (1765) specifies it to be between 750 and 900 vierkante roeden, which is between 1.06 and 1.27 hectares (Meertens Instituut, 2024). For the conversion rate, the average of these values is used (1 ammonam = 1.165 hectares). 44 coernix make one ammonam, which then equals 0.026 hectares (Abkouw, 1765).

To extract the most value from the LU lists, the locations found in the LU lists are matched with locations on the georeferenced historical maps. To start, fuzzy matching is used to identify some potential matches. The fuzzy matching uses the same list of known locations created in the text mining step (see Appendix II for the lists and 3.3 for the method) as it contains all locations on the georeferenced maps. To check the results of the matching, the matches are qualitatively assessed. The qualitative assessment is done by checking the location of a matched placename to potential matches in the vicinity. Since the LU lists for the regions are constructed by a person who travelled past these places, it is assumed that villages that are geographically close to each other are also listed close to each other and in most cases, there was a qualitative clustering of the proximity of placenames in lists and spatially on the historical maps. Furthermore, apart from the overarching regions (Alloetcoer Corle, and Hina Corle) the lists also mention a subregion. These subregions are plotted on the map to visually check potentially wrong matches.

From the resulting datasets (dataset from the map, and dataset from the texts) some statistics on the LU area are calculated. Firstly, a comparison is made between the growing area successfully matched with locations on the map and the total area mentioned in the texts. This shows how much textual data can be georeferenced using the abovementioned approach. Secondly, with the calculated growing area on the maps, a comparison is made with the area mentioned in the texts. This allows for a numerical comparison of the growing area in two years without losing data due to incomplete matching. Lastly, some descriptive statistics for the textual data are calculated, combining the qualitative information found in the lists and the quantitative information.

4 Results

In the following section, the results from the analysis of the various data types are presented. Firstly, the results from the text mining and trade reconstruction are presented, and lastly, a deeper dive is taken into the Ceylon region.

4.1 Text Mining

For each of the three locations, maps have been found that contain locations that are potentially interesting for LUC. As mentioned in 2.2, three maps containing information on LU are available for Ceylon. For Ambon and Banda, the maps in the collection of the Nationaal Archief 4.VEL were assessed. Most maps showed Ambon as a part of the larger archipelago, but one map was detailed enough to use. Lastly, for the Banda islands, a less detailed map than the ones used for Ambon and Ceylon was used, because there was no small-scale map available. All maps and the corresponding index number from the Nationaal Archief can be found in Appendix II. All locations mentioned on the map have been transcribed and collected in a list for each region which can also be found in Appendix II.

The NER step of the text mining resulted in a data frame of around 40000 sentences with a location for Ambon, 180000 for Banda, and 90000 for Ceylon. The evaluators received a sample of 100 sentences in which the NER model found 49 locations. The evaluators noted that in this sample, on average, the model missed 22 locations (with a spread of 18-29) and incorrectly identified 8 locations (with a spread of 4-11). The considerable spread of the results is reflected in a Fleiss' Kappa value of 0.55, indicating moderate agreement among the evaluators (Landis & Koch, 1977). In a later discussion of the results from the survey, the evaluators all expressed that they struggled with identifying locations themselves, which is an explanation for this lower Fleiss' Kappa value. This is probably due to the quality of the transcribed HTR texts.

Table 2: LU lists found in the text mining from the documents in Nationaal Archief 1.04.02. Two lists (7927, pp. 222-225 & 8937, pp.343-375) are transcribed and can be found in Appendix IV.

Document number	Page(s)	Year	Page found in text mining	Region	Content
7914	714-745	1740	742 & 743	Ambon	Number of trees sold per village and region
7927	222-225	1758	222	Ambon	Number of clove trees per village and region
8937	343-375	1723	351-354	Ceylon	Cultivated land area with units. Additionally contains textual information on the state of cultivation.
8941	80-134	1724	103 & 109	Ceylon	Number of coffee trees planted per village and region.
8954	778-806	1732	781	Ceylon	Cultivated land area with units. Additionally contains textual information on the state of cultivation.
8937	311-342	1723	331	Ceylon	Number of coffee trees planted per village and region.
8968	1052-1188	1735	1170	Ceylon	Number of trees cultivated per land parcel, village, and region. Additionally contains textual information on the state of cultivation. Does not contain the size of the parcels.
7967	329-336	1692	333	Banda	Number of nutmeg trees cultivated per village and region.
8013	109	1754	109	Banda	Number of nutmeg trees cultivated in total on Rosingain.
8016	66	1757	66	Banda	Number of nutmeg trees cultivated in total on Rosingain.
8018	81	1759	81	Banda	Number of nutmeg trees cultivated in total on Rosingain.
8027	490-505	1780	492 & 496	Banda	List of inhabitants and nutmeg trees per land parcel, village, and region. Additionally contains textual information on the state of cultivation.
8028	318-326	1782	320	Banda	List of inhabitants and nutmeg trees per land parcel, village, and region. Additionally contains textual information on the state of cultivation.
8031	385	1787	385	Banda	List of total nutmeg trees per region in 1786 and 1787.

The text mining yielded the documents presented in Table 2. Most documents specify the number of trees planted in a certain area. For the lists found in Ceylon, these are mainly coffee trees, which is interesting since coffee is not exported from Ceylon according to the trade data (see Figure 12 in section 4.2). For Ambon, one of the lists specifies the number of clove trees, and the other list specifies amounts of trees sold. Lastly, for Banda seven lists were found, all specifying how many trees are cultivated. Two of these lists additionally contain some qualitative information on the state of the cultivation. For all the identified lists only one or two pages of the lists were found with the text mining methodology. This means that many of the interesting pages were not identified by the methodology and therefore there are potentially additional interesting lists in the VOC archives.

For the further analysis of LU, two datasets (8937, pp. 343-375; 8954, pp. 778-806) were deemed especially interesting as they contain quantitative information on cultivated land area. In these lists, the village, region, and amount of cultivated land are specified. The LU list found in 8973 (pp. 343-375) additionally contains a qualitative assessment of the state of the cultivation in certain areas. To give an example:

"Door een der dorpelingen op eigen privéterein gecultiveerd (...) omdat dat de grond niet veel bijzonders is"

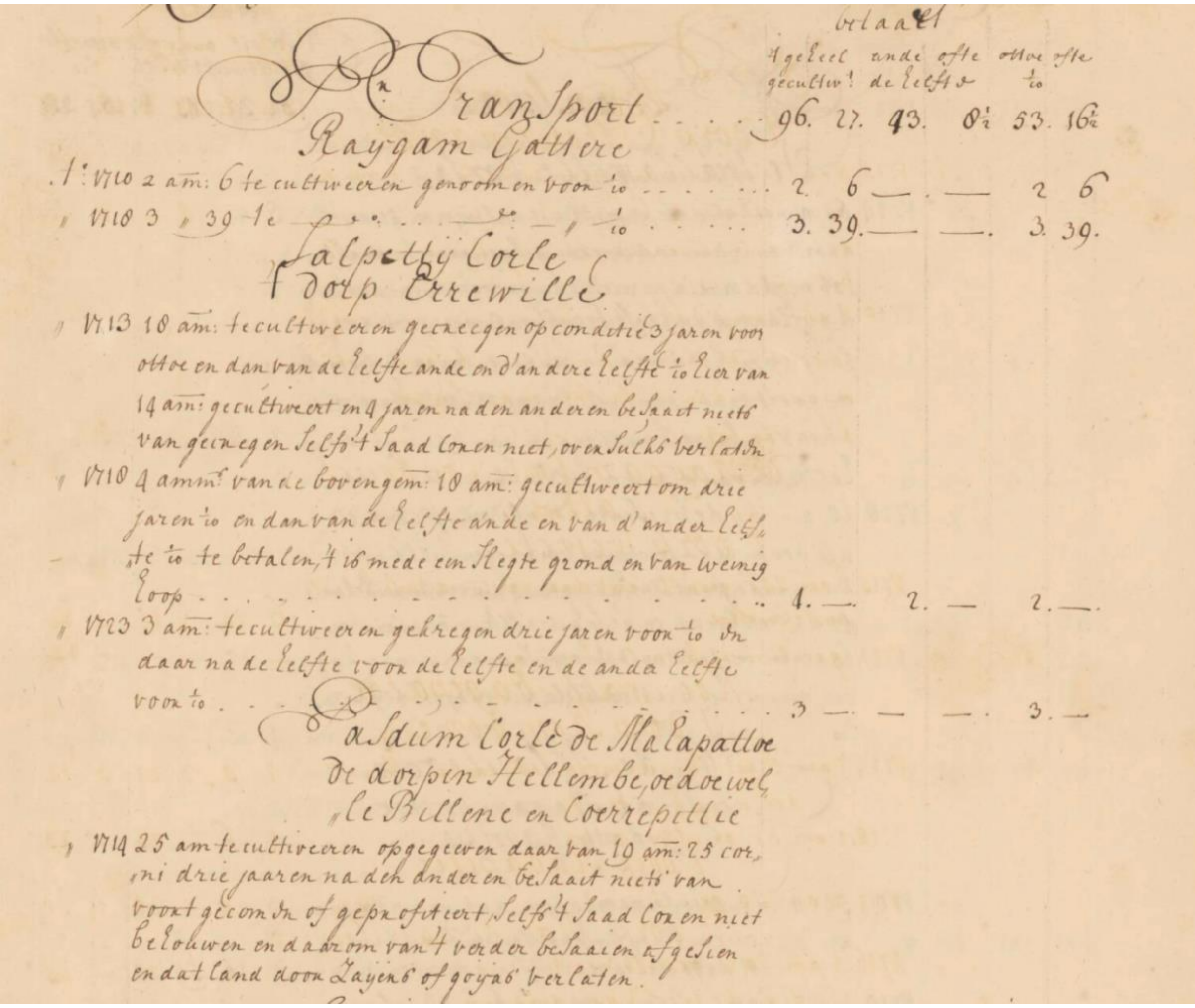
"Privately cultivated by one of the citizens (...) because the ground is nothing special"

The list also differentiates between the area that is still cultivated and the area that was previously appointed to be cultivated. This difference can be quite large and then often an explanation is given for the lack of cultivation. For example, an area of 25 ām (29 hectares) that is no longer cultivated:

"25 am te cultiveren opgegeeven daarvan zijn 19 ām: 25 corni drie jaaren na den anderen besaait niets van voortgecomen of geprofiteerd. Selfst 't Saad conen niet behouden en daarom van 't verder besaaien afgesien en dat land foor zaijers of goijas verlaten."

"25 am given to be cultivated of which 19 ām: 25 corni have been sowed three years apart. Nothing came from it or generated profit. Even the seeds have not been preserved and therefore further sowing of the land has not happened and the land has been abandoned."

In Figure 5 a part of the handwritten list can be found as an example. This gives an idea of the structure of the list and the information that can be found in it. Two lists were transcribed and can be found in Appendix IV.


belaast
4 geheel ande efte ottoe efte
gecultiv: de Eelfte to

Transport 96. 27. 43. 0½ 53. 16½

Raygam Gattere

1: 1710 2 ām: 6 te cultiveeren genoomen voor to . . . . . 2. 6 — . — 2 6
1718 3 „ 39 te „ „ „ to . . . . 3. 39. — —. 3. 39.

Salpettij Corle
t dorp Errewille

1713 18 ām: te cultiveeren gekreegen op conditie 3 jaren voor ottoe en dan van de Eelfte ande en d'andere Eelfte to Eler van 14 ām: gecultiveert en 4 jaren na den anderen besaait niets van geenegen selfs 't Saad Coren niet, over Sulks verlaten
1718 4 ammn: van de bovengem: 18 ām: gecultiveert om drie jaren to en dan van de Eelfte ande en van d'ander Eelfte to te betalen, 't is mede een Slegte grond en van weinig Loop . . . . 1. — 2. — 2. —.
1723 3 ām: te cultiveeren gekregen drie jaren voor to en daar na de Eelfte voor de Eelfte en de ander Eelfte voor to . . . . 3 — — —. 3. —

Pasdum Corle de Malapattoe
de dorpen Hellembe, oedoewel, le Bellene en Coerrepitie

1714 25 am te cultiveeren opgegeeven daar van 19 ām: 25 corni drie jaaren na den anderen besaait niets van voort gecomen of geprofiteert, selfs 't Saad Coren niet behouwen en daarom van 't verder besaaien afgesien en dat land door Layens of goyas verlaten.

Figure 5: a small part of a page from LU list 8937, p. 347. The list contains information on the area that is cultivated and the area that was appointed to be cultivated. The text contains information on the state of cultivation and the amount of tax owed.

4.2 Trade Reconstruction

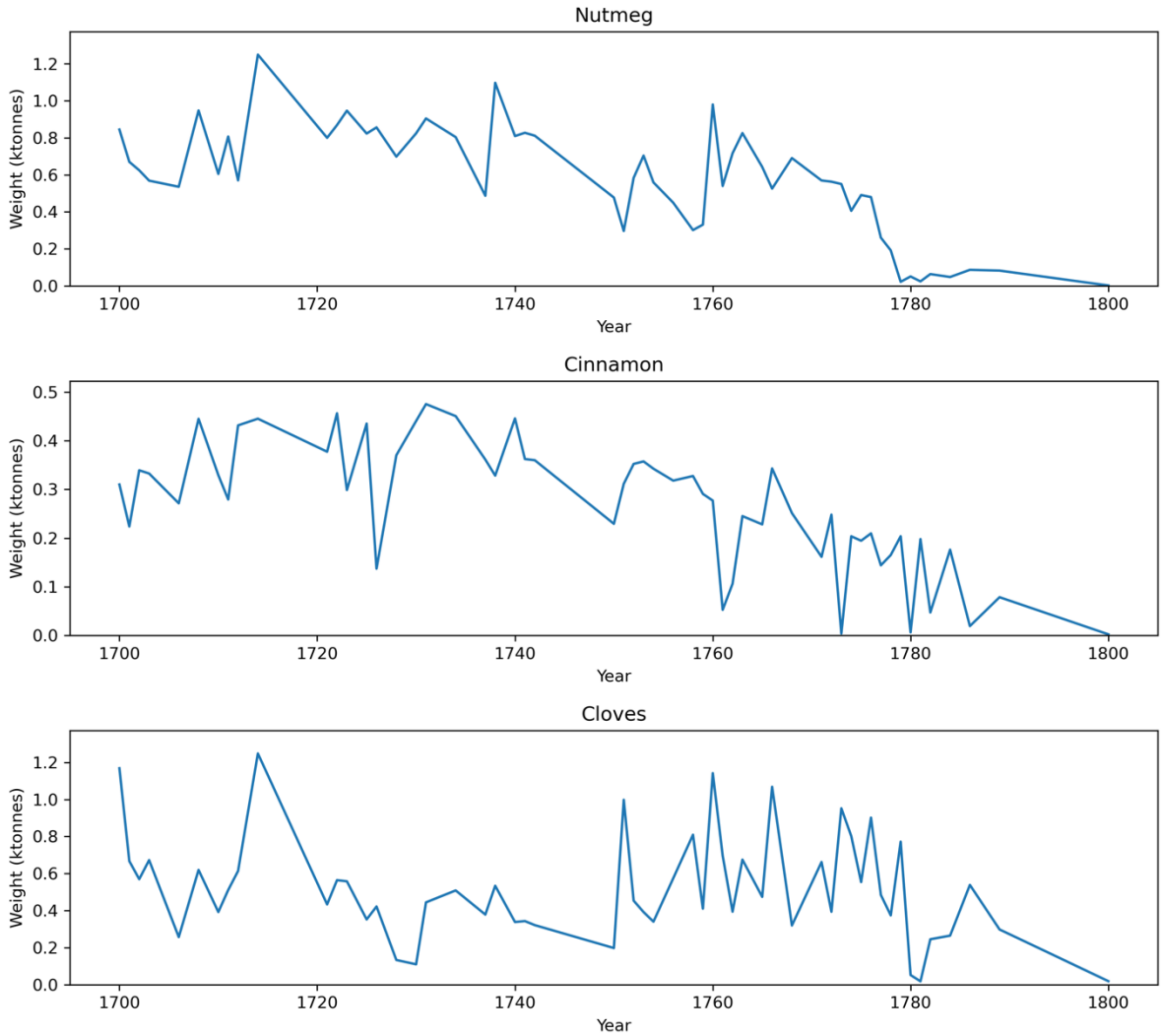


Figure 6: Line charts with the trends of the trade of nutmeg, cloves, and cinnamon by the VOC between 1700 and 1800. The y-axis changes per graph.

Figure 8 shows that pepper is the most traded commodity by weight, followed by sugar and tea. Cinnamon, cloves, and nutmeg are traded much less than these commodities, however, they are considered important commodities for the VOC in terms of economic value (Van Rossum, 2020). Diving deeper into these three commodities, Figure 6 shows that trade in these commodities peaked in the early years of the 18th century and decreased rapidly after 1780, the starting year of the 4th Anglo-Dutch war. For cloves, after 1720 there was a clear reduction in the transported weight until around 1750 when the trade picked up again. This is probably due to the reduction in the amount of clove trees that the VOC brought about to increase the prices of cloves (Knaap, 1996). For nutmeg and cinnamon, the graphs show a more gradual reduction in the amount traded.

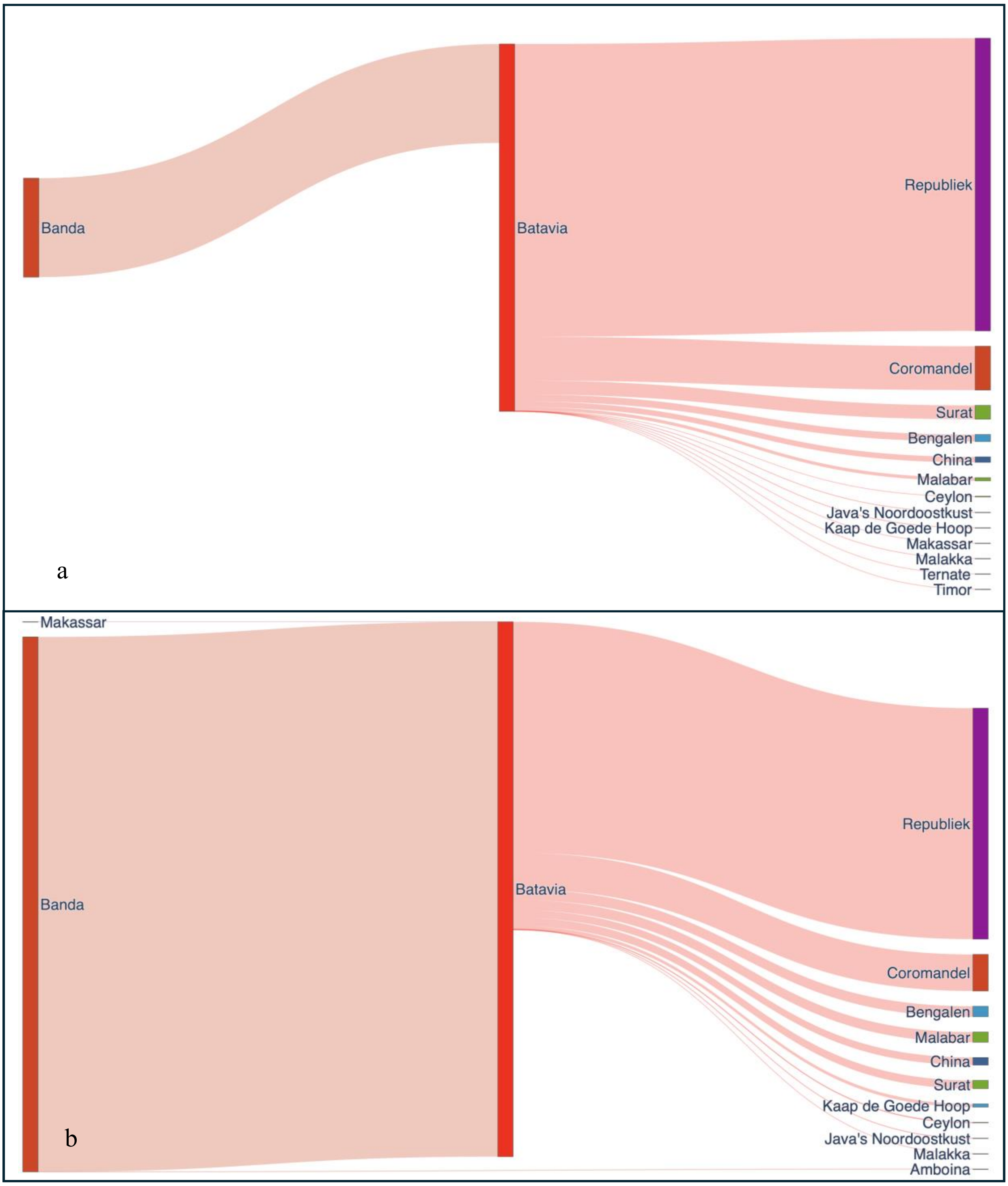


Figure 7: Sankey diagrams of the trade flows of nutmeg in the VOC for 1766 (a) and 1768 (b). Nutmeg was only exclusively cultivated on the Banda islands, and therefore the difference between import and export from Batavia means that there must have been some stock changes.

The Sankey diagrams (see https://voc-trade-visualisation.onrender.com for all diagrams and Figure 7 for two examples) show that cinnamon, cloves, and nutmeg were produced in specific regions. As discussed before, the VOC did this to manage the monopolies on the transport of these goods. The commodities were mainly imported to the Netherlands (Republiek); however, a substantial part of the cloves and nutmeg was transported to other regions, mainly within Asia.

Comparing the two Sankey diagrams in Figure 7, it can be seen that in 1766 the export of nutmeg from Batavia is larger than the import from Banda, and in 1768 this is the other way around. Because cloves were almost exclusively cultivated in Banda, there must have been some stock changes in Batavia during that period. These two Sankey diagrams are not the only occasions that this is clear. Going through some of the Sankey diagrams over time more occasions can be identified. As with the line charts, this can be explained by the data not reflecting the source of the commodities but rather the amount that is transported.

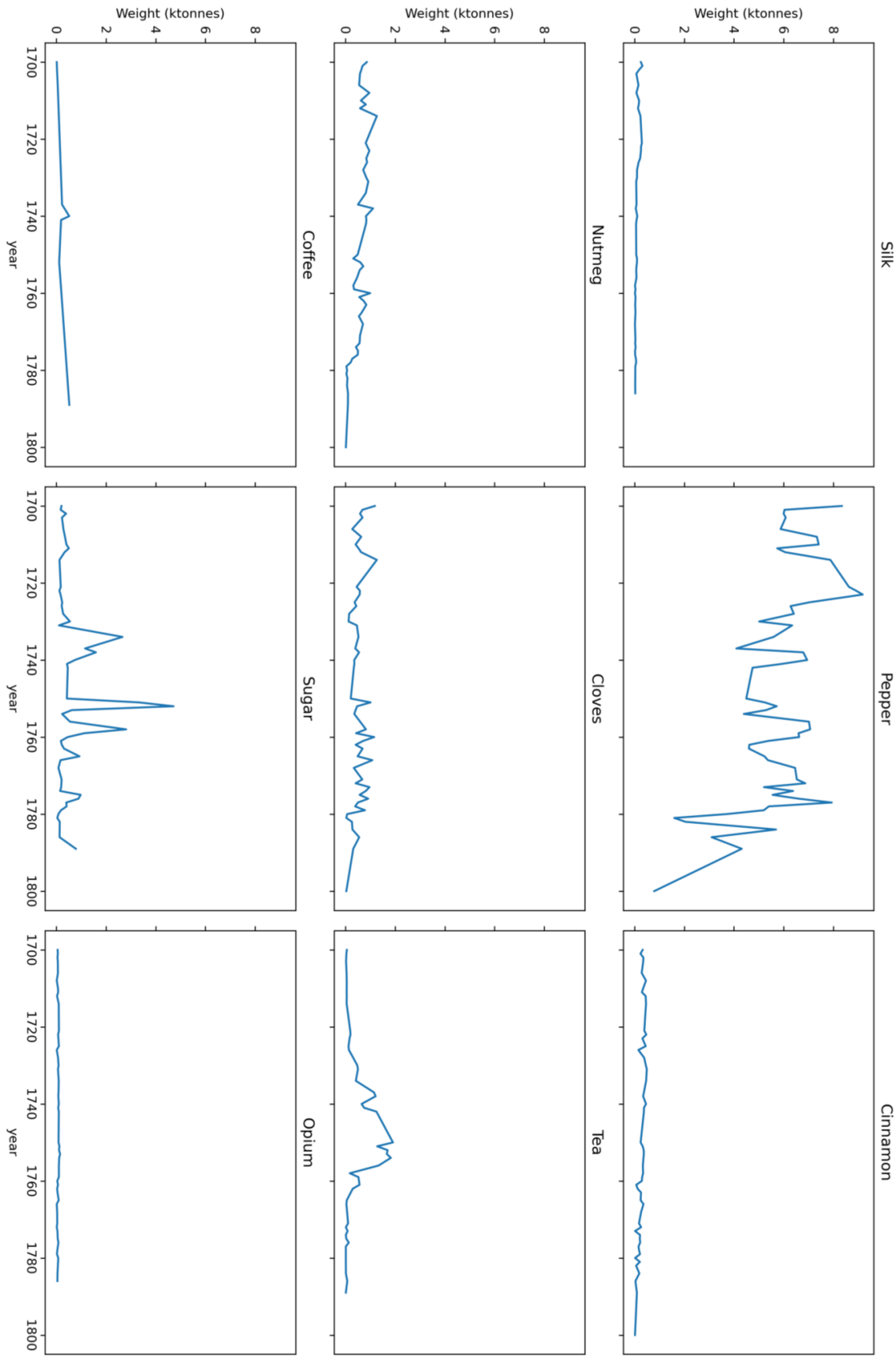


Figure 8: Trade in 9 different commodities conducted by the VOC over time with rigid axis. The rigid axis allows for comparison between the total amount of trade for each commodity.

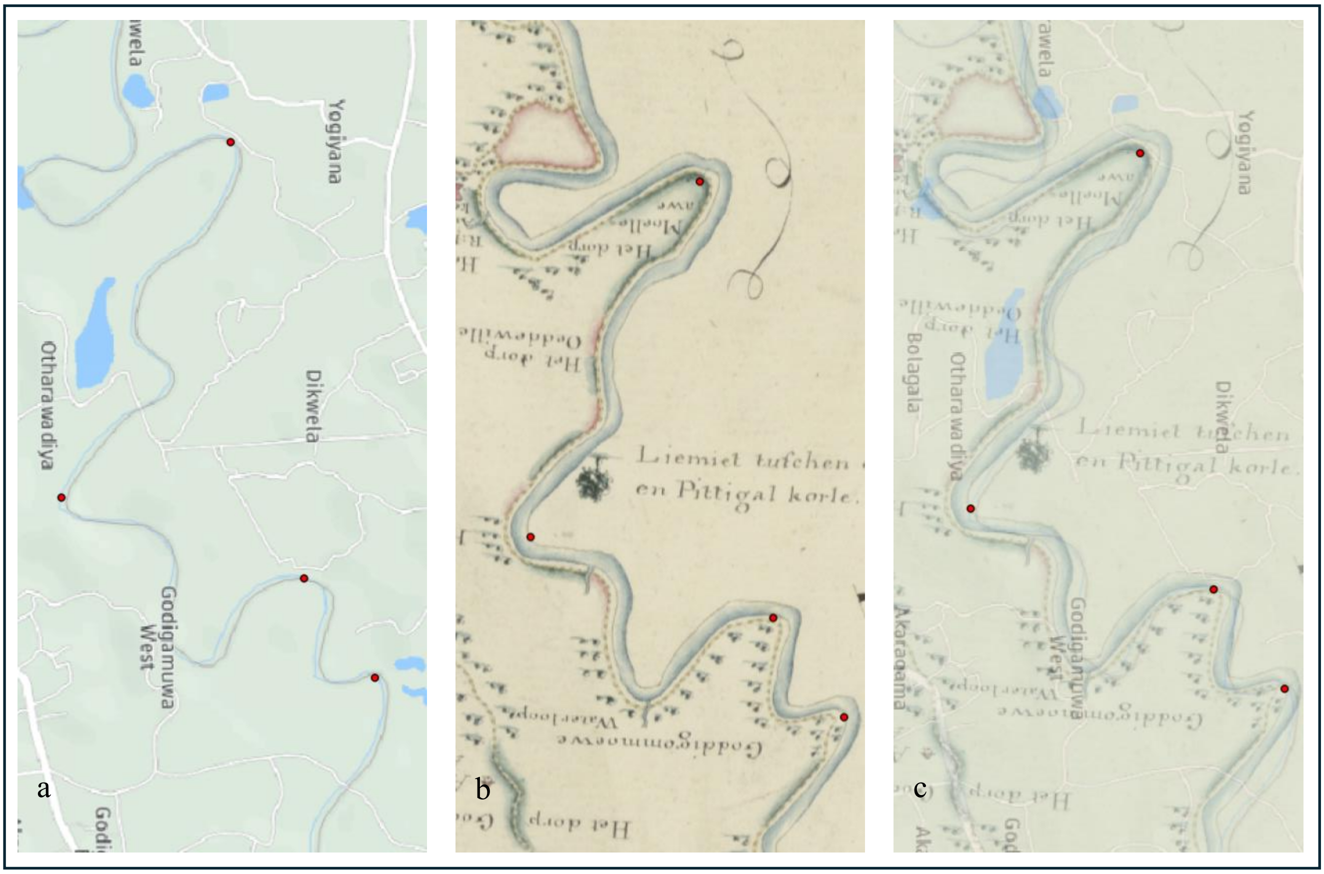


Figure 9: excerpt from the georeferenced map of Alloetcoer Corle, where a) shows the baselayer, b) shows the historical map, and c) shows an overlay of both. The GCPs are represented with the red dots. The full maps can be found in Appendix V.

The first step in the land use reconstruction is georeferencing the maps. The georeferenced maps of Ceylon from 1793 can be found in Appendix V. The tie points mainly follow the large rivers and the sea since the historical maps show very different road layouts and there are few other landmarks on the historical map. Figure 10 shows a small section of the georeferenced maps to give an impression of how the georeferencing came about. For the map from Alloetcoer Corle, 20 GCPs were found, and for Hina Corle 24. The rivers on the historical maps were mainly at the edges of the map, and therefore there are fewer GCPs in the inland section of the map. The GCPs, however, can be found all around the map.

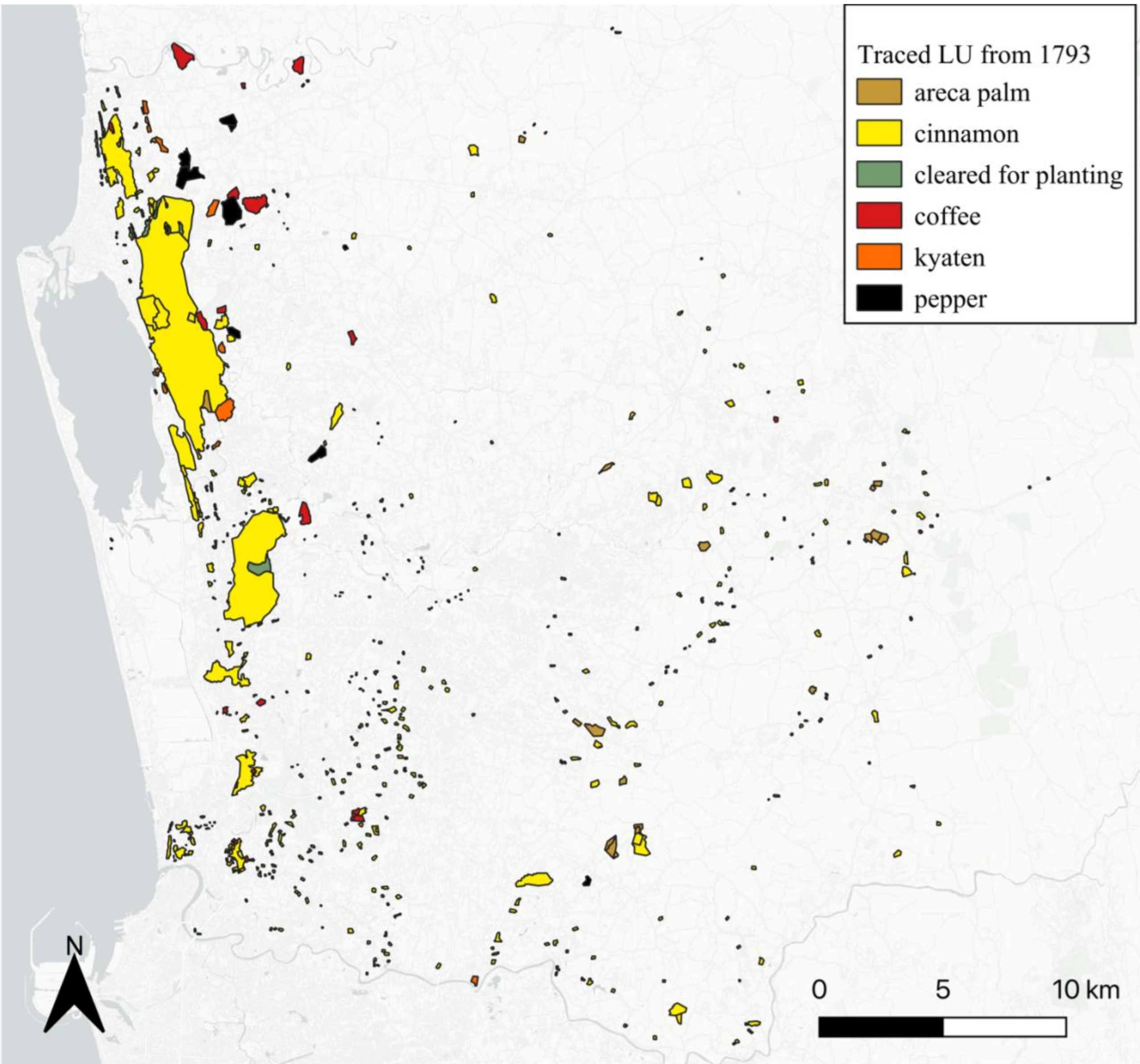


Figure 11: different cultivation types in Hina Corle and Alloetcoer Corle in 1793 traced from historical maps (for the maps see Appendix II). Most cultivation happened in areas close to the sea.

Using the georeferenced maps, the polygons as shown in Figure 11 are created. Cinnamon is the most widely cultivated crop, followed by coffee and pepper (Table 3), however, pepper cultivation is mostly concentrated on the northwestern side of the map and therefore the spread over the map is minimal. According to Figure 12, pepper is the most exported product by weight, and coffee is not exported at all. The exports as shown in Figure 11, however, are for the whole island and not only for the two regions presented on the map. Therefore, there could be much more widespread cultivation of pepper in other places on the island. The coffee not being exported is interesting as it shows that either the coffee is consumed locally, or the export does not appear in the trade data.

Table 3: Total area from the polygons in Figure 11 that were taken from georeferenced maps on LU during the VOC period in Hina Corle and Alloetcoer Corle in 1793.

Land use	Area (ha)
Cinnamon	4057
Coffee	264
Pepper	181
Areca Palm	152
Unclear	105
Cleared for planting	86

The planting areas as presented on the VOC maps show that they do not necessarily follow the paths of the larger rivers, as was hypothesised. All planting areas are located around villages, which are mainly connected through roads rather than waterways. In addition, the closer to the coast, the larger the planting areas are. These planting areas closer to the coast mainly surround the lake of Negombo, which is currently one of the most highly populated areas of Sri Lanka. This could have something to do with the area being flatter or having higher quality soil than the more inland area, and therefore being easier to cultivate. Another explanation could be that the VOC focused more on areas closer to the coast as these were closer to ports used to transport goods.

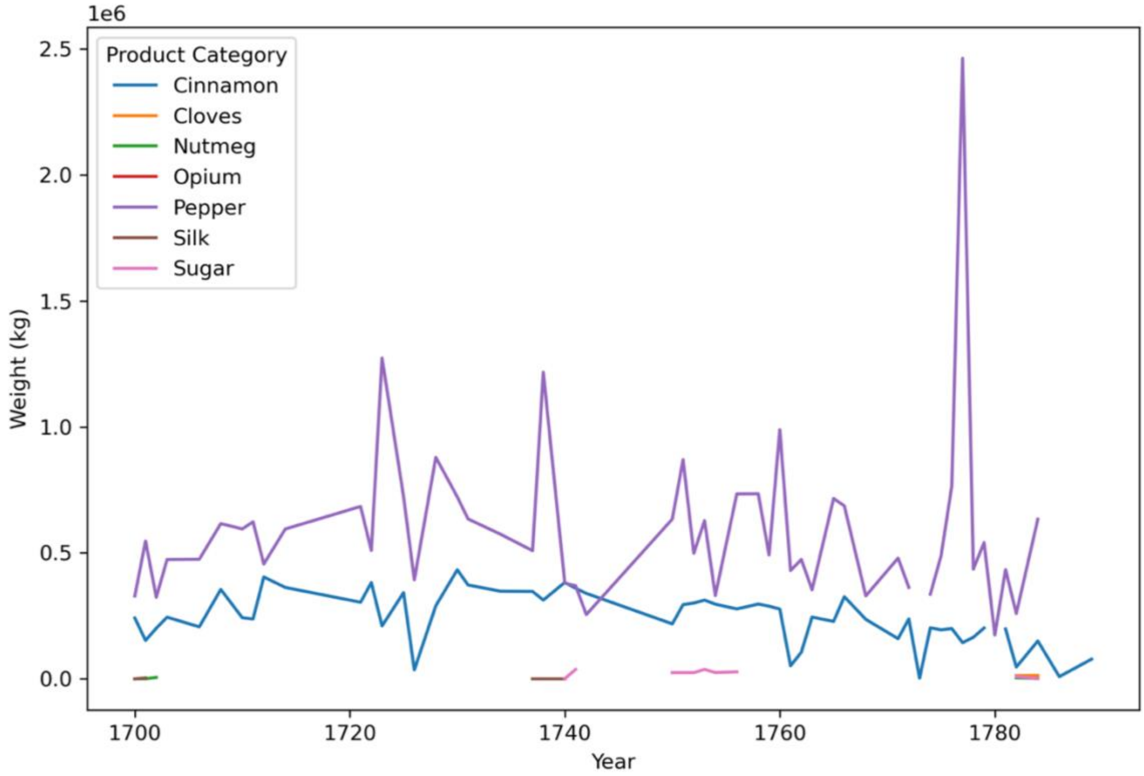


Figure 12: exports from Ceylon over time by weight. Pepper is for almost all years the most exported commodity, followed by cinnamon. Other commodities are only sporadically exported. Coffee is not exported from Ceylon based on the trade data.

With the LU list found in the text mining data from 1723, the map shown in Figure 13 is constructed. The texts distinguish between what was cultivated in 1723 specifically and what was appointed by the VOC to be grown. This means that for statistics on the growing area, two numbers could be used. For this study, mostly the appointed growing area is used because, for most occasions, this land is cleared to be planted even if it is no longer used. When comparing the year 1723 to other years, however, the actual cultivation in that year is used. For the maps, the appointed growing area is used, as it most clearly reflects where LUC has occurred.

From the texts 36 villages (out of 113 mentioned in the LU list) were successfully georeferenced to a location on the map, meaning that from the LU list, 32% of the locations could be connected to a location on the map. These locations contain around 34% (172 ha) of the total appointed cultivation area (506 ha) in Alloetcoer Corle and Hina Corle mentioned in the LU list. Figure 13b shows the size of the cultivation area on the map and Figure 13a shows the placenames of the georeferenced points and the region in which they are categorised according to the LU list. It can be seen that the locations found in the texts are all a bit more inland than the polygons in Figure 11, however, they too show larger cultivation areas closer to the coast. Furthermore, locations close to each other on the map are also in the same region as described in the LU list, indicating that the georeferencing was successful.

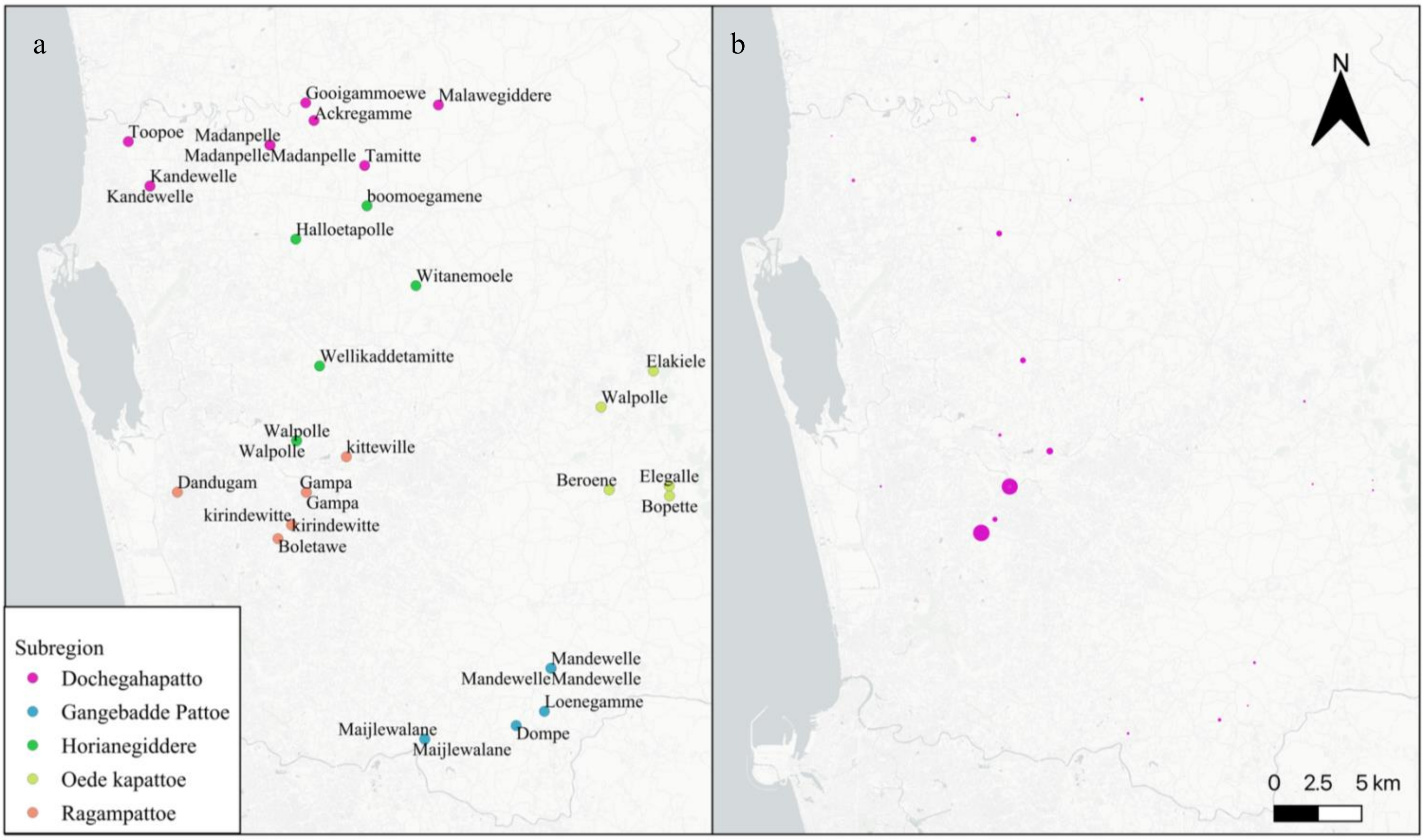


Figure 13: Maps created from a LU list from 1723 (see Appendix IV for the transcribed list). a) shows the locations names that have been matched to a location on a georeferenced map, and the subregion as mentioned in the LU list. b) shows the size of the growing area. The growing areas are relatively further inland compared to the maps from 1793 (Figure 11). The colour points in a) show that georeferenced points that are in the same region are also close to each other on the map. The areas in b) are larger towards the coast.

The remarks made in the texts give some additional interesting results. There is a big difference between the amount of land the VOC allocated to be cultivated, and what was actually cultivated. For all regions mentioned in the LU list (not only for Hina Corle and Alloetcoer Corle) in 1723 around 48% (430 ha) of the allocated land was not, or no longer cultivated. Sometimes the texts include an explanation for this which can be categorised into three main explanations: the land is poor (8 times), there are not enough workers (4 times), or the planting did not return a profit (10 times). Additionally, the texts mention areas cultivated "*without consent*" from the VOC. These areas contain around 10% of the total cultivated land, but around

40% of the cultivated plots mentioned in the text. These areas are often described with an additional notion that the ground is poor (69 times).

Using the data on the maps and the data in the texts the map in Figure 14 is constructed. It can be seen that the cultivated area in Hina Corle and Alloetcoer Corle has grown greatly from the 326 ha (actually cultivated area) mentioned in the texts from 1723 to the 4845 ha presented on the maps from 1793. There is also limited overlap in the growing areas, meaning that areas cultivated in 1723 do not appear on the maps from 1793 (Figure 14). Although there could be overlap in the locations from the texts that could not be georeferenced, it at least shows that the location of the planting areas changed over time.

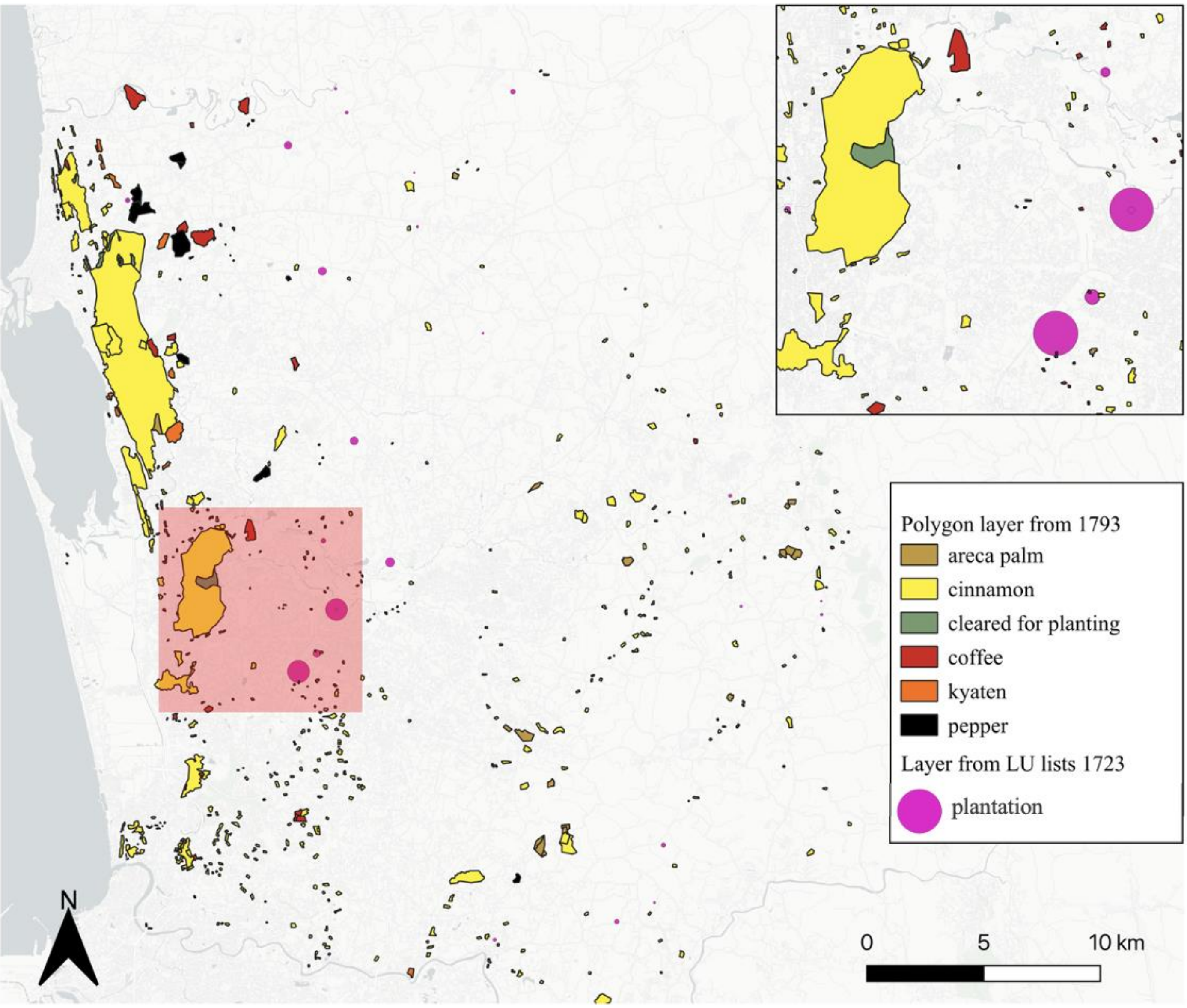


Figure 14: combined map with traced polygons from historical maps from 1793 (for historical maps see Appendix II) and information found in textual documents from 1723 (for the lists see Appendix IV). Much more land was cultivated in 1793 than in 1723 and the location of the plantations was more inland.

5 Discussion

5.1 Text Mining

The use of LLMs made it possible to sift through the large archives of the VOC, and by combining it with information found on maps, allowed for the identification of LU lists. It is fast enough to be useable for research on large archives, and sensitive enough to identify some interesting documents without identifying too many useless ones. The approach is applicable to various regions (as long as there are maps) and can easily be adapted to other languages.

Because the methodology can identify LU lists, by definition it is successful. However, the NER evaluation and the small number of pages of the found lists being identified with the text mining approach show improvements can be made. It is hard to quantify the percentage of potentially interesting documents the method can identify. There are potentially many more LU lists that might contain useful information. The NER model assesses the occurrence of a location based on the surrounding words and in a list the locations can be too isolated to be picked up on (Li et al., 2022). This means that some locations are missed by the NER and therefore these documents are disregarded in the method. It indicates that although the approach works regarding the direct outcome, it does not give a comprehensive overview of all interesting documents.

With this method being able to go through documents with little to no ex-ante knowledge of the content, it can serve well as a starting point for further analysis. To take the text mining of historical documents for the identification of LU lists a step further, the structure of the identified documents can be used as input for refining the identification process. For example, the lists found by the method mostly contain a high amount of interpunction, which is how the HTR model coped with the table structure of the texts. This would allow the identification of additional LU lists, making more information available for LU reconstructions.

Advancements in AI and LLMs could make this type of research much more structured. AI models for table recognition are being developed and work well on modern printed tables (Smock et al., 2021). Such a model for handwritten texts would be a more direct way of identifying lists in the documents and would not require the transcripts of these documents.

Additionally, developments in HTR and LLMs can improve the quality of the transcripts, and NER models for historical Dutch can improve the identification of locations (Hendriks et al., 2021). Additionally, the push for digitising and transcribing archives in other countries and languages opens up more documents for analysis (e.g. Karnadipa et al. (2023)).

5.2 Trade Reconstruction

The trade data in Huygens (2022) was used to reconstruct trade patterns. This required unit conversions and the grouping of commodities, which was mostly possible with information from the VOC glossary (Instituut voor Nederlandse Geschiedenis, 2000). The visualisations of the trade data can be used to easily grasp the size and direction of the trade.

The trade data used, however, are not always in accordance with information from other data sources. Van den Belt (2008) calculates that the total export of cinnamon from Ceylon between 1700 and 1760 adds to 36700000 pond, and with the same conversions used in 3.1 this equates to 1813 tonnes. They calculate this by assessing the archives of the VOC and not the trade data in Huygens (2022). The total export during the same years found in this thesis is only 967 tonnes, meaning that almost half the exports are missing. The conversion rates used by Van den Belt (2008) are similar to the ones used in this thesis, meaning that either the grouping of commodities is wrong, or the trade data is incomplete.

Additionally, the trade data does not always reflect the source or origin of the commodity. For example, the differences in the "origin" of cloves seen in Figure 7 and cloves being almost exclusively cultivated in Banda (Brixius, 2018) show that there must have been some stock changes in Batavia disguising the actual origin of the commodities.

Lastly, apart from the trade data not showing the actual trade, there can also be a difference between what is traded and what is produced. For example, during the 4th Anglo-Dutch War in 1780, the trade collapsed, only to rebound slightly at the end of the century (see Figure 6). Production of commodities on the other hand is unlikely to have ceased and therefore this does not reflect an actual decline in the production of goods.

5.3 Land Use Change Reconstruction

The analysis of the texts and maps shows a large increase in the cultivation of different commodities, mainly cinnamon. The texts from 1723 mostly paint a picture of experimentation with cultivation, whereas the maps from 1793 show that cultivation is much more formalised. The changing locations of the plantations over time, the many notions of poor soil quality, and the many abandoned plots mentioned in the texts give the impression that the VOC initiated LUC and that this did not go without setbacks. Cinnamon was not cultivated in plantations before 1770 (Dewasiri et al., 2021), therefore it is logical that the identified texts, which are all from before 1770, do not regard cinnamon plantations. The large scale of cinnamon cultivation on the maps shows that a large part of the LUC must have happened after 1770.

LUC is often quantified using a modelling approach and only sporadically uses historical documents. The analysis of historical documents, on the other hand, often aims to quantify the economic situation of Ceylon (Van den Belt, 2008), or looks at the situation of workers in Ceylon (Dewasiri et al., 2021). The approach taken in this study allows the use of historical documents for reconstructing and analysing LUC, albeit on a small scale. The documents contain information on LU and even specify units of area, and this information can be linked to locations on a map with relative certainty. Furthermore, the textual documents offer a wealth of qualitative information that can be used to determine why certain areas were cultivated and others were not.

This study does not explain the patterns of LUC found in the documents and maps. For example, it was not assessed whether the trade patterns correlate with the actual land use found in lists. It would be interesting to assess how much trade or other factors such as precipitation and elevation influence the spatial and temporal distribution of LU as described in the LU lists. Additionally, the lists of the number of trees found in a certain region can be converted to an area by making assumptions about the spacing of the trees. Doing this would provide additional information for reconstructing LU and could be used to estimate LU in years for which no area is specified.

5.4 Future Research

As mentioned before, the inductive and pragmatic research approach taken in this thesis allowed for the identification of LU lists and the georeferencing of placenames in these lists to locations on the map. The methodology, however, is not always very elegant and often requires much interpretation from the researcher. This makes it very time-consuming to apply the method to other regions in which the VOC was active and therefore limits the usefulness of the approach for adding to models such as HYDE3.2 (Goldewijk et al., 2017), LUH2 (Hurtt et al., 2020), and SAGE cropland (Ramankutty & Foley, 1999). Preferably, a method would be developed with the text data as an input and georeferenced land use lists as an output, but this proved impossible with currently available LLMs and other AI models. With the absence of a comprehensive overview of the LU for the whole of the VOC area, it would be interesting to see if other indicators can be used to inform where and when LU occurs.

To disentangle these, more data is needed regarding both a larger region and a larger temporal range. As mentioned in 5.1, there are probably many more interesting lists in the documents. In the long term, these could be identified with some table recognition software, however, in the short term, a more pragmatic approach such as a search for a high percentage of interpunction in sentences could also help to identify more lists. Additionally, the lists with amounts of trees can be converted to LU with some assumptions, offering another source of data for the analysis.

With the identified and transcribed lists, a correlation between the amount of LU and trade over time can be assessed. This can be done for a sample of regions specified in the trade data making it possible to narrow down the potential areas for cultivation. Additionally, using trade data would not require georeferencing of locations, and therefore if there is a correlation this would be a very useful indicator for LU. Based on the results from this thesis, there are indications that the trade data from Huygens (2022) is not complete, however, a correlation would comprehensively prove its usefulness.

For a spatial analysis of the found lists, the placenames in the lists need to be linked to a location on the map, preferably with a more formalised linking of placenames from the texts to placenames on the maps. This thesis used a qualitative assessment of the vicinity of the point to the placenames and the points occurring in a similar region. However, the approach could be formalised by using some statistical measures that consider both the fuzzy matching outcome and the vicinity of the points to another point. This approach could lead to a higher percentage of correctly matched locations than what was achieved in this study (34%), meaning more data is available for the spatial spread of the LU. Furthermore, this approach makes it possible to do the matching much faster and eliminates the need for manual intervention.

Lastly, with the georeferenced LU lists, LU models can be assessed against actual historical LU data. In this thesis, LU lists are found from as early as 1692, which means that the accuracy of the models in predicting LU over 300 years ago can be assessed. Consequently, the model parameters can be adjusted according to the outcome of such an analysis, potentially improving the models for a much longer period.

6 Conclusion

LLMs can be used for the analysis of historical documents, however these do need input from the researcher. The runtime of generative LLMs means that they are not useful for analysing large quantities of text data, and therefore, for now, task-specific models are still the way to go. Additionally, the documents transcribed with HTR are of such a low quality that scans of the texts must be used to qualitatively analyse the output of the Text mining process. This makes it impossible to completely automate the process and requires the researcher to acquire knowledge of the research area.

A quantitative reconstruction of trade can be made using the trade data from Huygens (2022), however, there are indications that this data is incomplete. Both maps and textual data can be used to reconstruct colonial land change patterns. The texts contain information on the size and location of planting areas, either in terms of actual quantification of the size or information on amounts of trees. For Ceylon, LU maps are available adding another layer of information. A part of the locations mentioned in the LU list from the textual data has been successfully georeferenced onto a map. This allowed comparing the textual data and data on a map. The usefulness of the trade data as a proxy for LU was not assessed in this study, however by further analysing the texts, this can be done in the future.

This thesis can be used as a basis for further research both on the LUC patterns in Ceylon, Banda, and Ambon, and research on the further development of a methodology for text mining. To disentangle the potential drivers of historical LUC, a broader dataset across larger regions and time frames is needed. This dataset can be used to quantify the correlation between

LU and trade, which could pinpoint cultivation areas without requiring georeferencing. Additionally, for spatial analysis a more formalised statistical approach for the linking placenames could improve accuracy and efficiency, surpassing the current 34% matching rate. Lastly, the georeferenced LU lists from historical data can refine LU models, enhancing their accuracy for predictions over the past 300 years. In this way this thesis can be a small step towards offering transparency into where historical LUC happened, on what scale, and who is accountable.

7 Acknowledgements

First and foremost, I would like to thank Brian Dermody for his counsel during every phase of the research. I enjoyed our feedback meetings, and I would not have been able to complete my research without your guidance. I would also like to thank Janpieter van der Pol and the digital humanities group for their help with the development of the text mining methodology. Lastly, my discussion with Marcel van Egmond on historical maps of the VOC helped me understand the difficulties in georeferencing historical locations to their modern counterpart. I enjoyed the research process a lot, and without the help and knowledge of the aforementioned people, the end result would not have been the same.

8 References

Abkouw, F. V. (1765). Te Colombo. In *Memoriën van munten, maten en gewichten, in Ned. Indië, Voor-Indië*. https://books.google.nl/books?id=85FiAAAAcAAJ&pg=RA1-PA5&dq=ammonam&hl=nl&newbks=1&newbks_redir=0&sa=X&ved=2ahUKEwiFmdWmyfSGAxVL3AIHHZe3AtAQ6AF6BAgKEAI#v=onepage&q&f=false

Affek, A. (2013). Georeferencing of historical maps using GIS, as exemplified by the Austrian Military Surveys of Galicia. *Geographia Polonica*. https://doi.org/10.7163/GPol.2013.30

Ananthavinayagan, T. V. (2019). *Sri Lanka's History: Colonialism, Independence and Conflict*. 17–67. https://doi.org/10.1007/978-981-13-7350-3_2

Arima, E. Y., Barreto, P., Araújo, E., & Soares-Filho, B. (2014). Public policies can reduce tropical deforestation: Lessons and challenges from Brazil. *Land Use Policy*, *41*, 465–473. https://doi.org/10.1016/J.LANDUSEPOL.2014.06.026

Austin, K. G., Schwantes, A., Gu, Y., & Kasibhatla, P. S. (2019). What causes deforestation in Indonesia? *Environmental Research Letters*, *14*(2), 024007. https://doi.org/10.1088/1748-9326/AAF6DB

Barceloux, D. (2008). Cinnamon (Cinnamomum Species). *Medical Toxicology of Natural Substances: Foods, Fungi, Medicinal Herbs, Plants, and Venomous Animals*, 39–43. https://doi.org/10.1002/9780470330319.CH4

Brixius, D. (2018). A hard nut to crack: nutmeg cultivation and the application of natural history between the Maluku islands and Isle de France (1750s–1780s). *The British Journal for the History of Science*, *51*(4), 585–606. https://doi.org/10.1017/S0007087418000754

Cervera, T., Pino, J., Marull, J., Padró, R., & Tello, E. (2019). Understanding the long-term dynamics of forest transition: From deforestation to afforestation in a Mediterranean landscape (Catalonia, 1868–2005). *Land Use Policy*, *80*, 318–331. https://doi.org/10.1016/J.LANDUSEPOL.2016.10.006

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, *15*(1). https://doi.org/10.1145/3479010

De Silva, C. R. (1967). The Portuguese in Ceylon, 1617-1638 . *University of London, School of Oriental and African Studies (United Kingdom)* . https://www-proquest-com.proxy.library.uu.nl/openview/928133663aac88e0389d6aa79ad2e5e2/1?accountid=14772&cbl=2026366&diss=y&parentSessionId=%2B1iDpJ8myN%2BhQNFUxCHfUP

%2F9YIW0VY7NDGJhQPa14PY%3D&parentSessionId=Vg7kLJ4L4yvIZCLbApxPlG1tBj1SeIZjs04EwMht8f4%3D&pq-origsite=gscholar

Den Elzen, M., & Schaeffer, M. (2002). Responsibility for past and future global warming: Uncertainties in attributing anthropogenic climate change. *Climatic Change*, *54*(1–2), 29–73. https://doi.org/10.1023/A:1015750019028/METRICS

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Dewasiri, N. R., Wagenaar, L. J., & Uyangoda, J. (2021). Historical, ethno-botanical and social aspects of cinnamon cultivation in Sri Lanka. *Cinnamon: Botany, Agronomy, Chemistry and Industrial Applications*, 39–61. https://doi.org/10.1007/978-3-030-54426-3_2/FIGURES/5

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N., & Snyder, P. K. (2005). Global consequences of land use. *Science (New York, N.Y.)*, *309*(5734), 570–574. https://doi.org/10.1126/SCIENCE.1111772

Gilg, A. (2009). Perceptions about land use. *Land Use Policy*, *26*(SUPPL. 1), S76–S82. https://doi.org/10.1016/J.LANDUSEPOL.2009.08.018

GLOBALISE. (2023). *VOC transcriptions - GLOBALISE*. https://hdl.handle.net/10622/JCTCJ2

Goldewijk, K. K., Beusen, A., Doelman, J., & Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene - HYDE 3.2. *Earth System Science Data*, *9*(2), 927–953. https://doi.org/10.5194/ESSD-9-927-2017

Greene, T. J. (2024). Text and Data Mining for Pianists? Bringing Digital Humanities to a Graduate Music Research Methods Course Through Topic Modeling. *Music Reference Services Quarterly*, *27*(1), 10–25. https://doi.org/10.1080/10588167.2023.2293602

Guleij, R. (2017). In de kaart gekeken: Gerrit de Haan, baas-kaartenmaker van de VOC in Batavia: (Persoonlijkheden). *Groniek*, *213*. https://ugp.rug.nl/groniek/article/view/32560

Hendriks, B., Groth, P., & Van Erp, M. (2021). *Recognising and Linking Entities in Old Dutch Text: A Case Study on VOC Notary Records*. https://assets.amsterdam.nl/publish/pages/

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., & Coates, C. M. (2015). Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, *30*(suppl_1), i50–i75. https://doi.org/10.1093/LLC/FQV046

Hurtt, G. C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B. L., Calvin, K., Doelman, J. C., Fisk, J., Fujimori, S., Klein Goldewijk, K., Hasegawa, T., Havlik, P., Heinimann, A., Humpenöder, F., Jungclaus, J., Kaplan, J., Kennedy, J., Krisztin, T., Lawrence, D., … Zhang, X. (2020). Harmonization of Global Land-Use Change and Management for the Period 850-2100 (LUH2) for CMIP6. *Geoscientific Model Development*, 1–65. https://doi.org/10.5194/gmd-2019-360

Huygens. (2022). *Dutch-Asiatic Shipping in the 17th and 18th centuries*.

Instituut voor Nederlandse Geschiedenis. (2000). *voc-glossarium*. www.inghist.nl

IPCC. (2023). *Climate Change 2023 Synthesis Report*. 35–115. https://doi.org/10.59327/IPCC/AR6-9789291691647

Jadin, I., Meyfroidt, P., & Lambin, E. F. (2016). International trade, and land use intensification and spatial reorganization explain Costa Rica's forest transition. *Environmental Research Letters*, *11*(3), 035005. https://doi.org/10.1088/1748-9326/11/3/035005

Kanapathypillai, V. (1969). *Dutch Rule in Maritime Ceylon 1766-1796*. https://www-proquest-com.proxy.library.uu.nl/openview/38588c8fed71ad5570b038b509c919aa/1?accountid=14772&cbl=2026366&diss=y&parentSessionId=%2Fr8DqAlmPpin%2BCfCSd72Skg9LqX1PgZOscU1HxvVMr0%3D&parentSessionId=9Ns4OGESau9ym4keXD%2FLKh7w4UQ7K2wH%2Fec7zd0aZX4%3D&pq-origsite=gscholar

Karnadipa, T., Safitri, K. A., Vita, D., Swasto, W., Grataridarga, N., Mardiati, W., & Putri, N. R. (2023). Digitization of the 17th and 18th Centuries' Dutch East India Company (VOC) Archives for The Archives' Preservation. *Proceedings 2022, Vol. 83, Page 60*, *83*(1), 60. https://doi.org/10.3390/PROCEEDINGS2022083060

Keijser, L. (2018). *Projectplan De ijsberg zichtbaar maken*. http://read.transkribus.eu/

Keller, W., & Borkowski, A. (2019). Thin plate spline interpolation. *Journal of Geodesy*, *93*(9), 1251–1269. https://doi.org/10.1007/S00190-019-01240-2/FIGURES/13

Knaap, G. (1996). *The economic policy of the VOC*. Brill. https://brill-com.proxy.library.uu.nl/display/book/9789004454491/B9789004454491_s005.xml

La Nasa, J., Mattonai, M., Modugno, F., Degano, I., & Ribechini, E. (2019). Comics' VOC-abulary: Study of the ageing of comic books in archival bags through VOCs profiling. *Polymer Degradation and Stability*, *161*, 39–49. https://doi.org/10.1016/J.POLYMDEGRADSTAB.2019.01.001

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

l'Honoré Naber, S., & Boink, G. (2023). *4.VELH Inventaris van de verzameling buitenlandse kaarten Leupe, supplement, 1621-1813 (1925)* . https://www.nationaalarchief.nl/onderzoeken/archief/4.VELH/invnr/593

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Luyssaert, S., Jammet, M., Stoy, P. C., Estel, S., Pongratz, J., Ceschia, E., Churkina, G., Don, A., Erb, K., Ferlicoq, M., Gielen, B., Grünwald, T., Houghton, R. A., Klumpp, K., Knohl, A., Kolb, T., Kuemmerle, T., Laurila, T., Lohila, A., … Dolman, A. J. (2014). Land management and land-cover change have impacts of similar magnitude on surface temperature. *Nature Climate Change 2014 4:5*, *4*(5), 389–393. https://doi.org/10.1038/nclimate2196

Medagoda, N., Shanmuganathan, S., & Whalley, J. (2013). A comparative analysis of opinion mining and sentiment classification in non-english languages. *International Conference on Advances in ICT for Emerging Regions, ICTer 2013 - Conference Proceedings*, 144–148. https://doi.org/10.1109/ICTER.2013.6761169

Meertens Instituut. (2024). *De oude Nederlandse maten en gewichten*. https://mgw.meertens.knaw.nl/maat/12

Meilink-Roelofsz, M. A. P. , R. R. , S. H. (2023). *Inventaris van het archief van de Verenigde Oost- Indische Compagnie (VOC), 1602-1795 (1811)*.

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P. C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, *16*(1), 1–11. https://doi.org/10.1186/S13040-023-00339-9/METRICS

Meyfroidt, P., Rudel, T. K., & Lambin, E. F. (2010). Forest transitions, trade, and the global displacement of land use. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(49), 20917–20922. https://doi.org/10.1073/PNAS.1014773107/SUPPL_FILE/PNAS.201014773SI.PDF

Nationaal Archief. (2021, June 4). *Kijk symposium De IJsberg zichtbaar maken terug | Nationaal Archief*. https://www.nationaalarchief.nl/beleven/nieuws/kijk-symposium-de-ijsberg-zichtbaar-maken-terug

Nijman, J. (1994). The VOC and the expansion of the world-system 1602–1799. *Political Geography*, *13*(3), 211–227. https://doi.org/10.1016/0962-6298(94)90027-2

OpenAI. (n.d.). *Introducing ChatGPT*. Retrieved February 29, 2024, from https://openai.com/blog/chatgpt

Paranavitana, K. D. (1999). Custody of the Archives of the V.O.C. in Ceylon — A Review (1640-1948) on JSTOR. *Journal of the Royal Asiatic Society of Sri Lanka*, *44*, 1–30. https://www.jstor.org/stable/23731417

Plieninger, T., Höchtl, F., & Spek, T. (2006). Traditional land-use and nature conservation in European rural landscapes. *Environmental Science & Policy*, *9*(4), 317–321. https://doi.org/10.1016/J.ENVSCI.2006.03.001

Ramankutty, N., & Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles*, *13*(4), 997–1027. https://doi.org/10.1029/1999GB900046

Ravindran, P. N., & Nirmal Babu, K. (2005). *Cinnamon and Cassia: The Genus Cinnamomum* (P. N. Ravindran, K. Nirmal Babu, & M. Shylaja, Eds.). CRC press. https://books.google.nl/books?hl=en&lr=&id=KZa8aPxR_-wC&oi=fnd&pg=PP1&ots=poJhtfwb5-&sig=QUvXcNT1VBRDCQHsOUS2xrWiiOE&redir_esc=y#v=onepage&q&f=false

Rijs Textiles. (n.d.). *Stretch Zijde Satijn Stof Marine*. Retrieved July 1, 2024, from https://www.rijstextiles.com/nl/modestoffen/stretch-zijde-satijn-stoffen/stretch-zijde-satijn-stof-marine.html

Rossum, M. van. (2011). De intra-Aziatische vaart: Schepen, "de Aziatische zeeman" en ondergang van de VOC? *Tijdschrift Voor Sociale En Economische Geschiedenis*, *11*(3), 31–68. https://research.vu.nl/en/publications/de-intra-aziatische-vaart-schepen-de-aziatische-zeeman-en-onderga

Schweter, S., Akbik, A., & De, Alan. A.-B. (2020). *FLERT: Document-Level Features for Named Entity Recognition*. https://arxiv.org/abs/2011.06993v2

Scott, G., & Hewett, M. L. (2008). Pioneers in ethnopharmacology: The Dutch East India Company (VOC) at the Cape from 1650 to 1800. *Journal of Ethnopharmacology*, *115*(3), 339–360. https://doi.org/10.1016/J.JEP.2007.10.020

Smail, R., Donaldson, C., Govaerts, R., Rayson, P., & Stevens, C. (2021). Uncovering Environmental Change in the English Lake District: Using Computational Techniques to Trace the Presence and Documentation of Historical Flora. *Digital Scholarship in the Humanities*, *36*(3), 736–756. https://doi.org/10.1093/LLC/FQAA047

Smock, B., Pesala, R., & Abraham, R. (2021). PubTables-1M: Towards comprehensive table extraction from unstructured documents. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 4624–4632. https://doi.org/10.1109/CVPR52688.2022.00459

Sterk, B., van Ittersum, M. K., & Leeuwis, C. (2011). How, when, and for what reasons does land use modelling contribute to societal problem solving? *Environmental Modelling & Software*, *26*(3), 310–316. https://doi.org/10.1016/J.ENVSOFT.2010.06.004

Suriyagoda, L., Mohotti, A. J., Vidanarachchi, J. K., Kodithuwakku, S. P., Chathurika, M., Bandaranayake, P. C. G., Hetherington, A. M., & Beneragama, C. K. (2021). "Ceylon cinnamon": Much more than just a spice. *Plants, People, Planet*, *3*(4), 319–336. https://doi.org/10.1002/PPP3.10192

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*.

Twidle, H. (2013). Writing the Company: From VOC Daghregister to Sleigh's Eilande. *South African Historical Journal*, *65*(1), 125–152. https://doi.org/10.1080/02582473.2013.763399

Van den Belt, A. (2008). *Het VOC-bedrijf op Ceylon : een voorname vestiging van de Oost-Indische Compagnie in de 18de eeuw*. Walburg Pers. https://hdl.handle.net/1887/12688

Van Garen en Stof. (n.d.). *Hoe meet je stof? | Van Garen & Stof en Donna Colori Restyle.* Retrieved July 1, 2024, from https://www.vangarenenstof.nl/c-3879338/hoe-meet-je-stof/#

Van Rossum. (2020). *De VOC, van Amsterdam naar Azië* (P. Brandon, G. Jones, N. Jouwe, & M. Van Rossum, Eds.). https://www.spectrumnonfictie.nl/sites/default/files/books/issuu/9789000372867.pdf

van Vliet, J., Magliocca, N. R., Büchner, B., Cook, E., Rey Benayas, J. M., Ellis, E. C., Heinimann, A., Keys, E., Lee, T. M., Liu, J., Mertz, O., Meyfroidt, P., Moritz, M., Poeplau, C., Robinson, B. E., Seppelt, R., Seto, K. C., & Verburg, P. H. (2016). Meta-studies in land use science: Current coverage and prospects. *Ambio*, *45*(1), 15–28. https://doi.org/10.1007/S13280-015-0699-8/FIGURES/4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Wagenaar, L. (2004). 'Eerst eenigen tijd versleeten met hunne dansers zien danssen en springen' Ceylonese Compagniesdienaren schrijven over kaneel en kaneelschillers in 1786. In Werkgroep Indisch-Nederlandse Letterkunde (Ed.), *Indische Letteren* (Vol. 19). dbnl. https://www.dbnl.org/tekst/_ind004200401_01/_ind004200401_01_0003.php

Weststeijn, A. (2014). The VOC as a Company-State: Debating Seventeenth-Century Dutch Colonial Expansion. *Itinerario*, *38*(1), 13–34. https://doi.org/10.1017/S0165115314000035

Wikipedia. (n.d.). *Amfioensociëteit - Wikipedia*. Retrieved July 1, 2024, from https://nl.wikipedia.org/wiki/Amfioensoci%C3%ABteit#cite_note-1

Winkler, K., Fuchs, R., Rounsevell, M., & Herold, M. (2021). Global land use changes are four times greater than previously estimated. *Nature Communications 2021 12:1*, *12*(1), 1–10. https://doi.org/10.1038/s41467-021-22702-2

9 Appendix I

Unit	Conversion	Commodity
lb	0.494	Cinnamon
baal	41.99	Cinnamon
lb	0.494	Pepper
pikol	60.268	Pepper
bahar	247	Pepper
lb	0.494	Nutmeg
pees	0.005	Nutmeg
pot	0.5	Nutmeg
aam	150	Nutmeg
lb	0.494	Cloves
potje	0.1	Cloves
pot	1	Cloves
lb	0.494	Tea
lb	0.494	Coffee
lb	0.494	Silk
rol	21	Silk
lb	0.494	Sugar
pikol	60.268	Sugar
lb	0.494	Opium

Below an explanation is given for each of the conversion factors above. If no source is mentioned, the is from Instituut voor Nederlandse Geschiedenis (2000).

Cinnamon:

Baal cinnamon = 85 pond (Wagenaar, 2004)

lb = pond = 0.494 kg

Pepper

lb = pond = 0.494 kg

Pikol = 122 pond

Bahar ~ 500 pond

Nootmuskaat

lb = pond = 0.494 kg

Pees = piece ~ 5 gram per piece

Pot = assumed to be about 100 pieces

Aam = barrel of 153.6 liter assumed to be ~ 300 pot so 300000 pieces

Kruidnagel

lb = pond = 0.494 kg

Potje = assumed to be 1000 pieces à 2 gram per 20 pieces

Pot = assumed to be 10000 pieces à 2 gram per 20 pieces

Tea

lb = pond = 0.494 kg

Coffee

lb = pond = 0.494 kg

Silk

lb = pond = 0.494 kg

Weight of silk = 70g per m^2 (Rijs Textiles, n.d.)

Rol ± 10 m length 1.2 m wide (Van Garen en Stof, n.d.)

Rol = 12 m^2 = 0.84 kg

El = 69 cm length, 1.2 m wide = 0.828 m^2 = 0.05796 kg

For pees (piece) no good transformation is available since there is a variety of products that use this unit that could all have different weights.

Sugar

lb = pond = 0.494 kg

Pikol = 122 pond

Opium

lb = pond = 0.494 kg

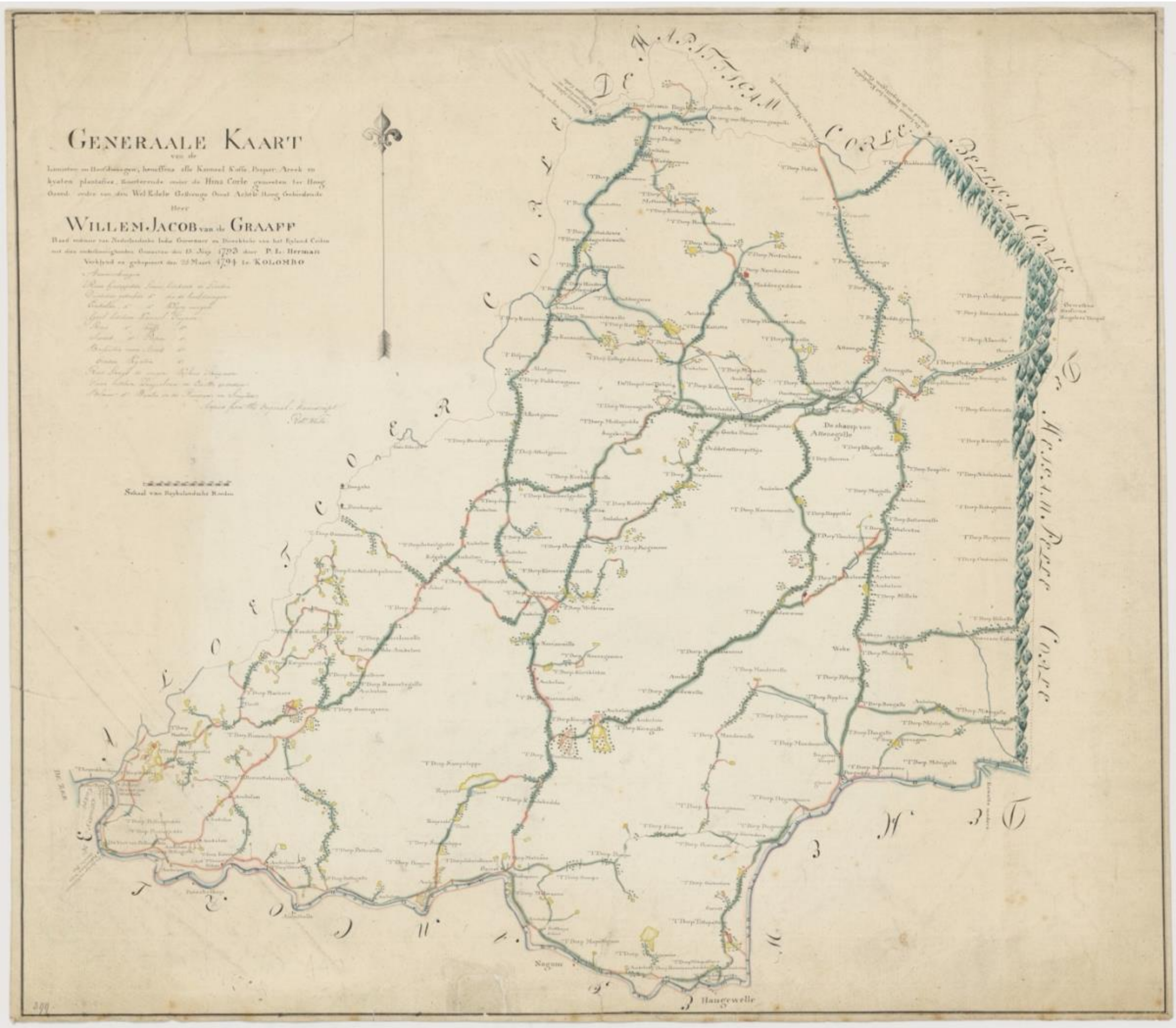
Kist = 60 kg (Wikipedia, n.d.)

10 Appendix II

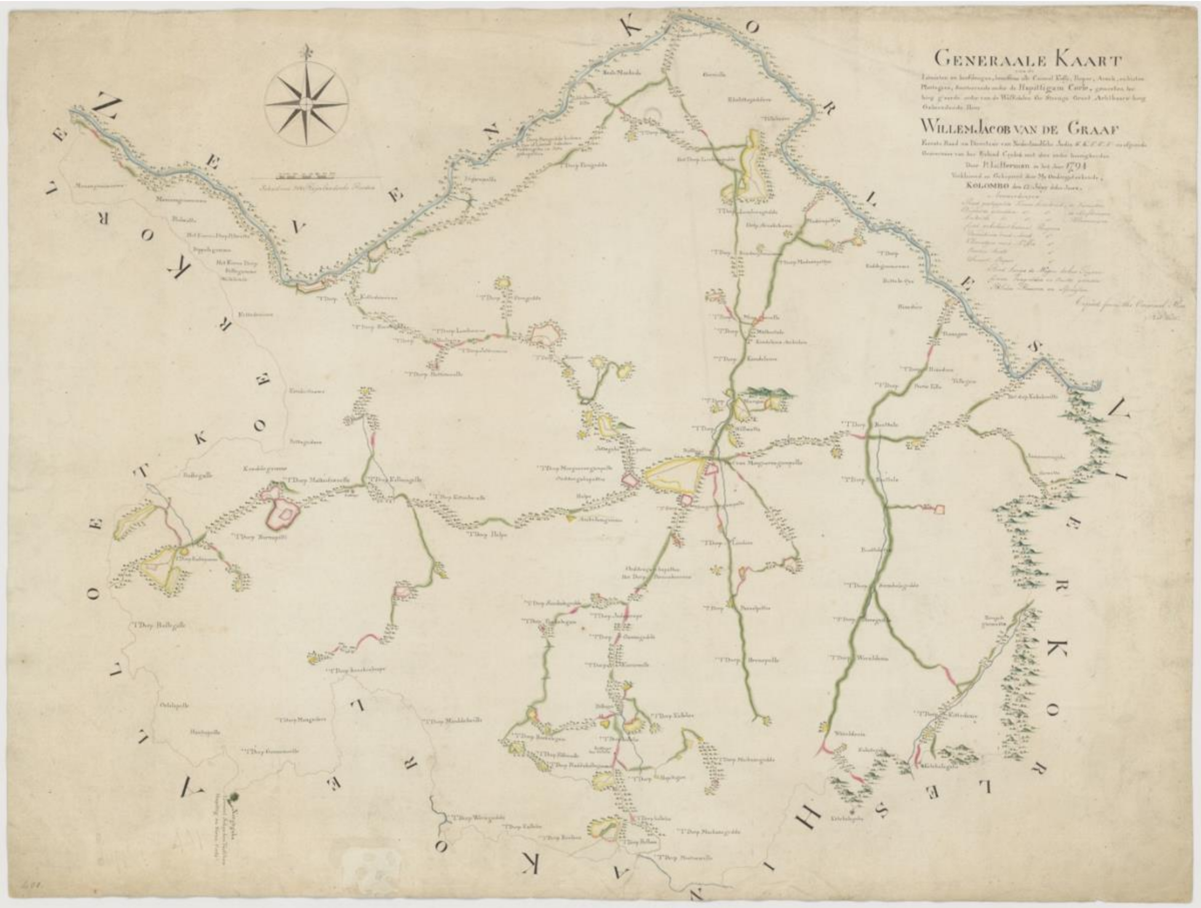
Ceylon

Figure 1: NL-HaNA_4.VELH_334

Figure 2: NL-HaNA_4.VELH_332

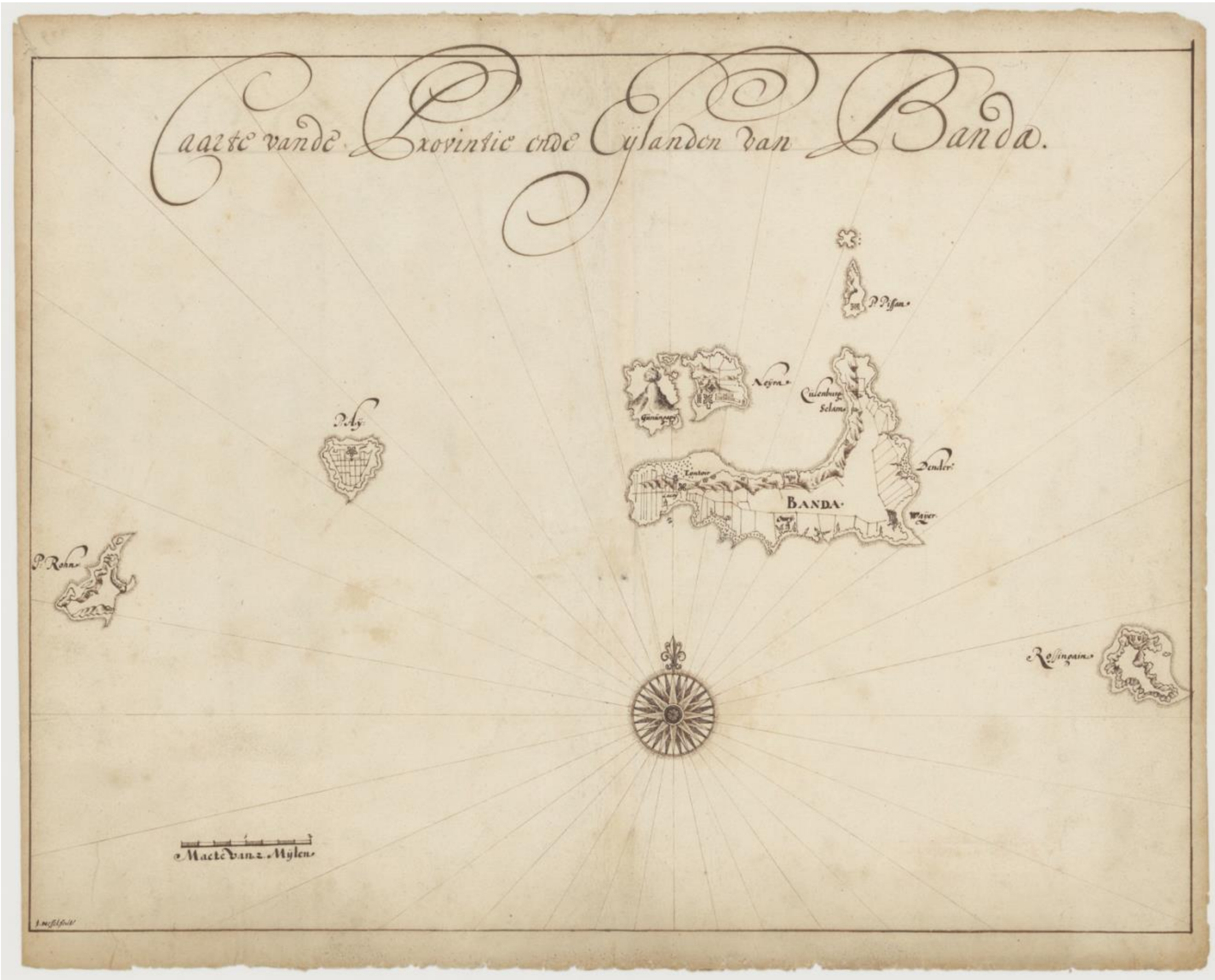


NL-HaNA_4.VELH_333



Ambon - NL-HaNA_4.VEL_1325

Banda - NL-HaNA_4.VEL_1356



The transcribed locations can be found in the accompanying excel files called:

- *placenames_sri* (Ceylon)
- *placenames_banda* (Banda)
- *placenames_ambon* (Ambon)

In the excel both the villages and the overarching region are mentioned. For Ceylon a different overarching region means a different map, and for Banda and Ambon this is a different island.

11 Appendix III

See the Excel called *output sample*. The different approaches taken are specified in the first row and mean:

- Llama2: identification of locations in the documents using Llama2 with output parcing through LlangChain. Output parsing is the cleaning of the output of an LLM. This makes it easier to use the output of the model for further analysis.
- Ollama: applies Llama2 without output parcing and using the Ollama package to make it easier to use on a local device.
- ner_dutch_large: Application of the NER model as specified in the methods.

12 Appendix IV

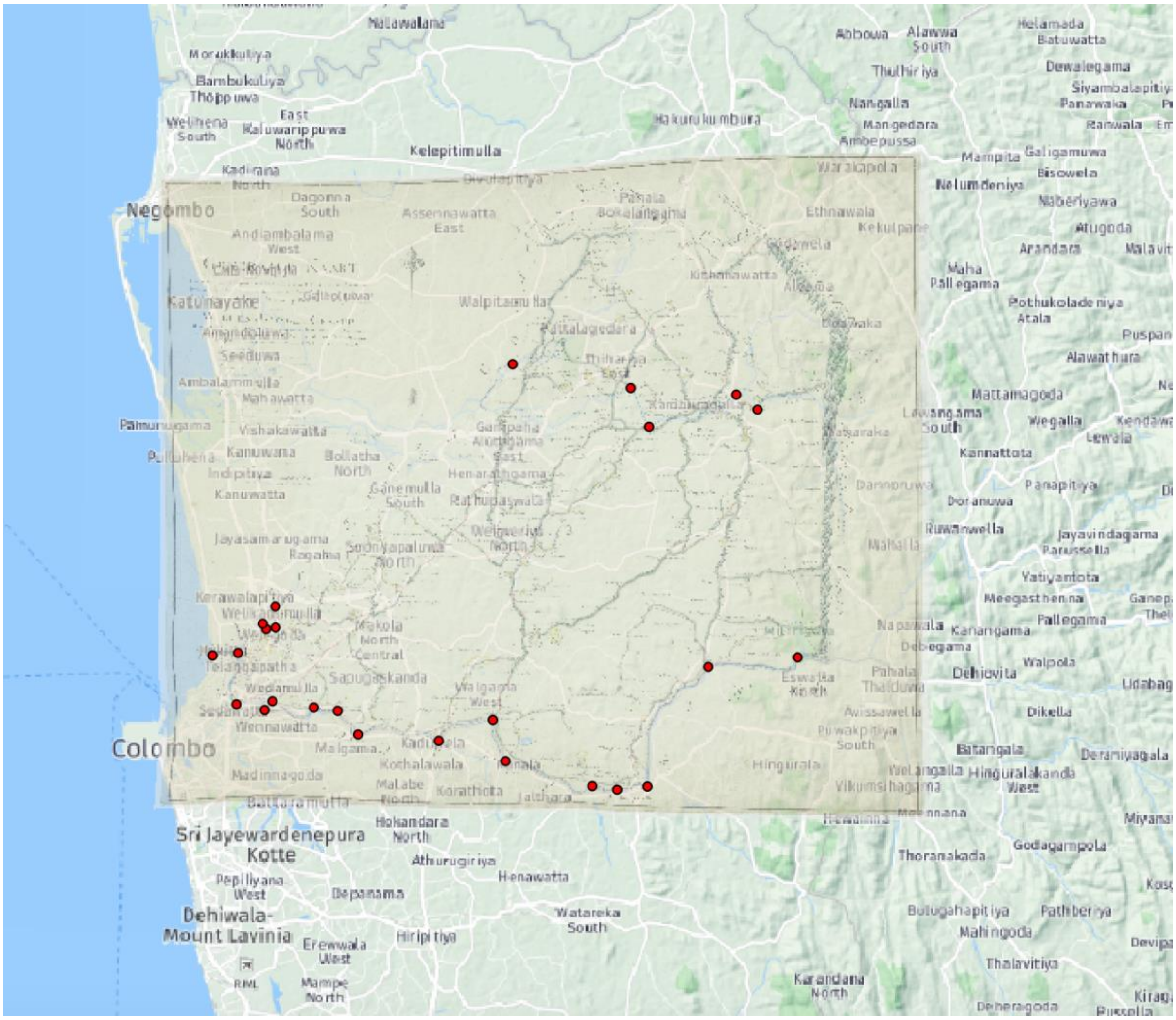
Transcribed documents can be found in accompanying excel files. These excel files are:

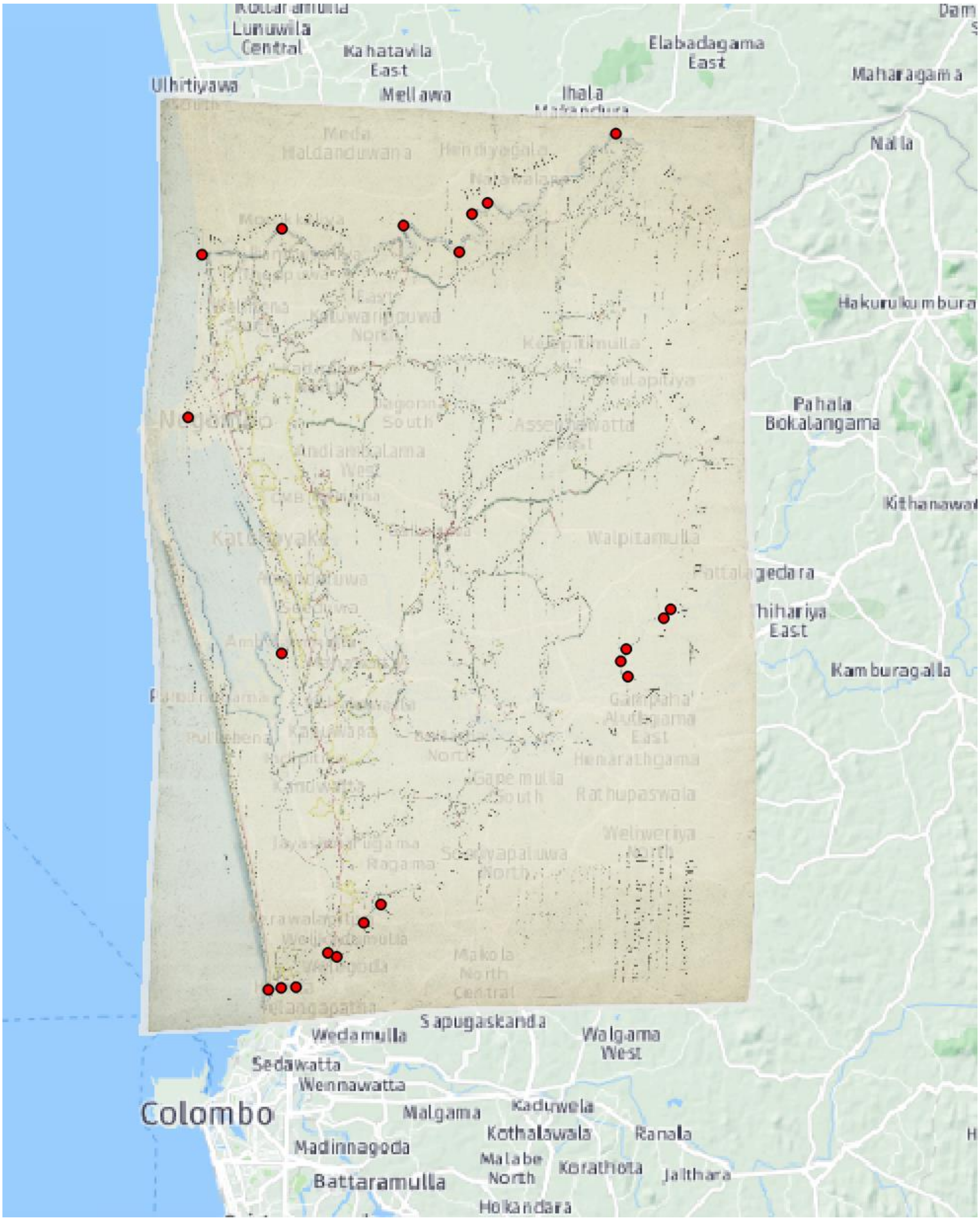
- *Dataset_ambon_1758* (transcript of index number 7927, pages 222-225)
- *Dataset_ceylon_1723* (transcript of index number 8937, pages 343-375)

The transcribed documents for Ceylon also contain some computations for the area in hectares and some coding is done based on the remarks made in the texts.

13 Appendix V

Georeferences maps of Ceylon. The CRPs are presented on the maps.





14 Appendix VI

For the scripts and data see the attached zip file.