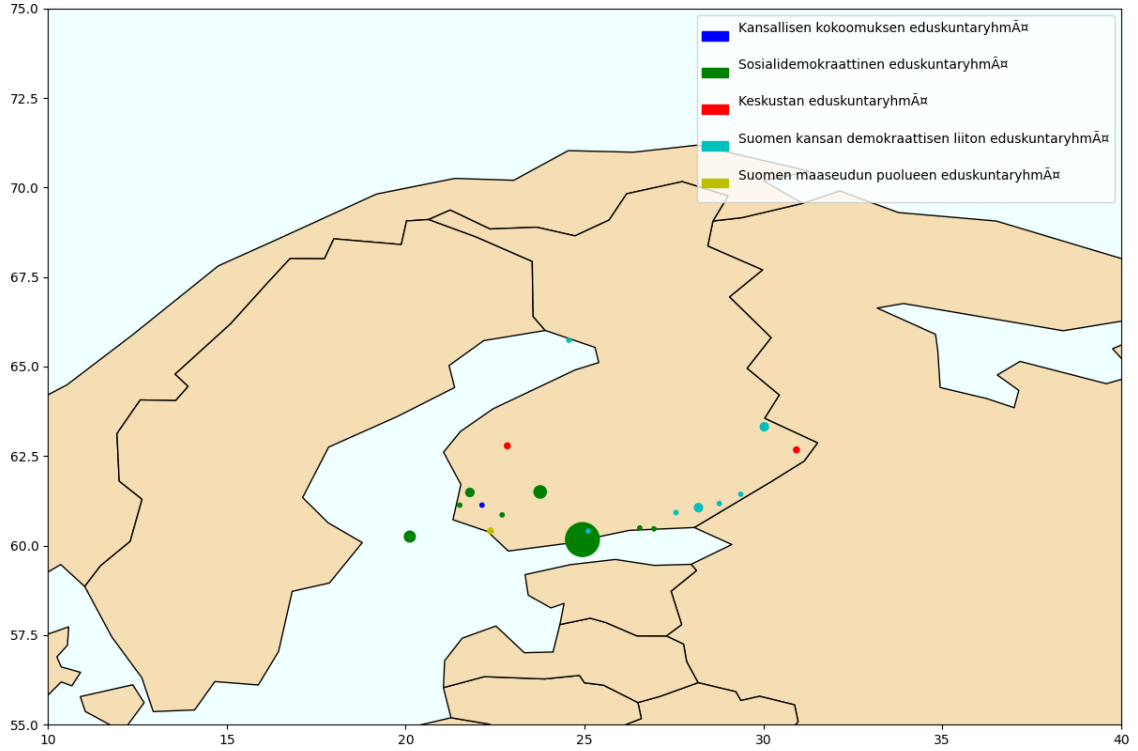
Final Project

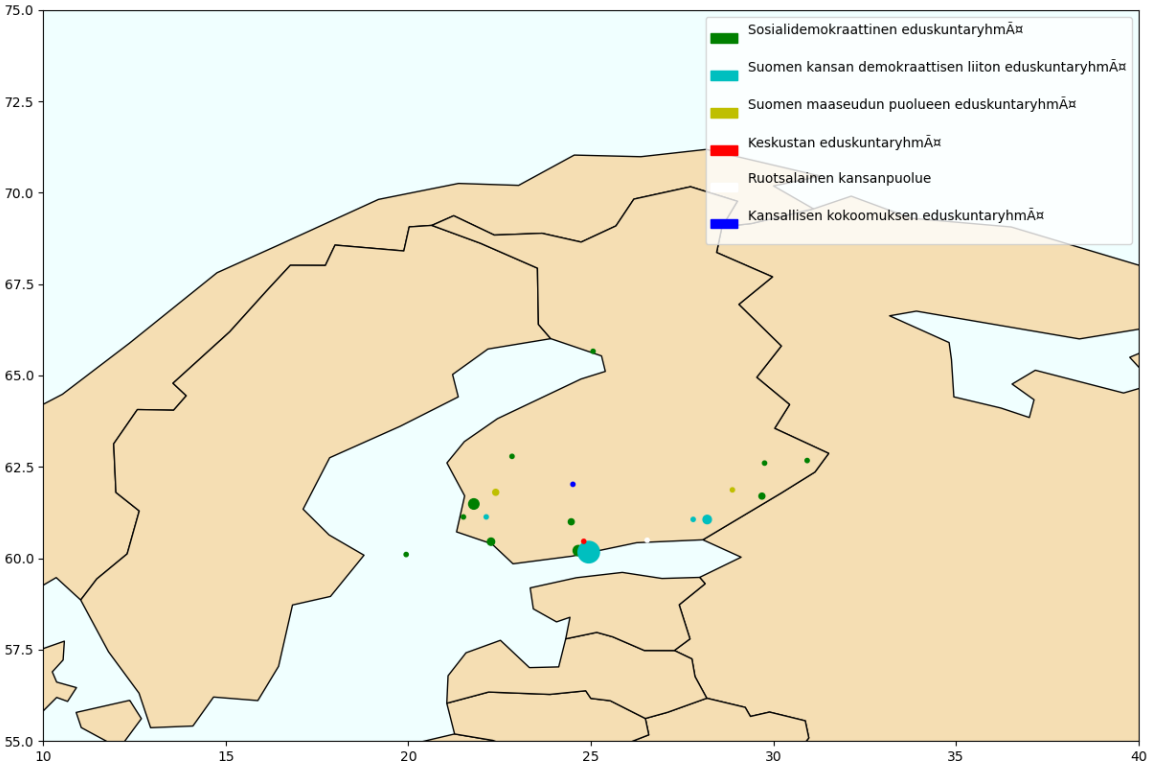
Mikko Kivistö

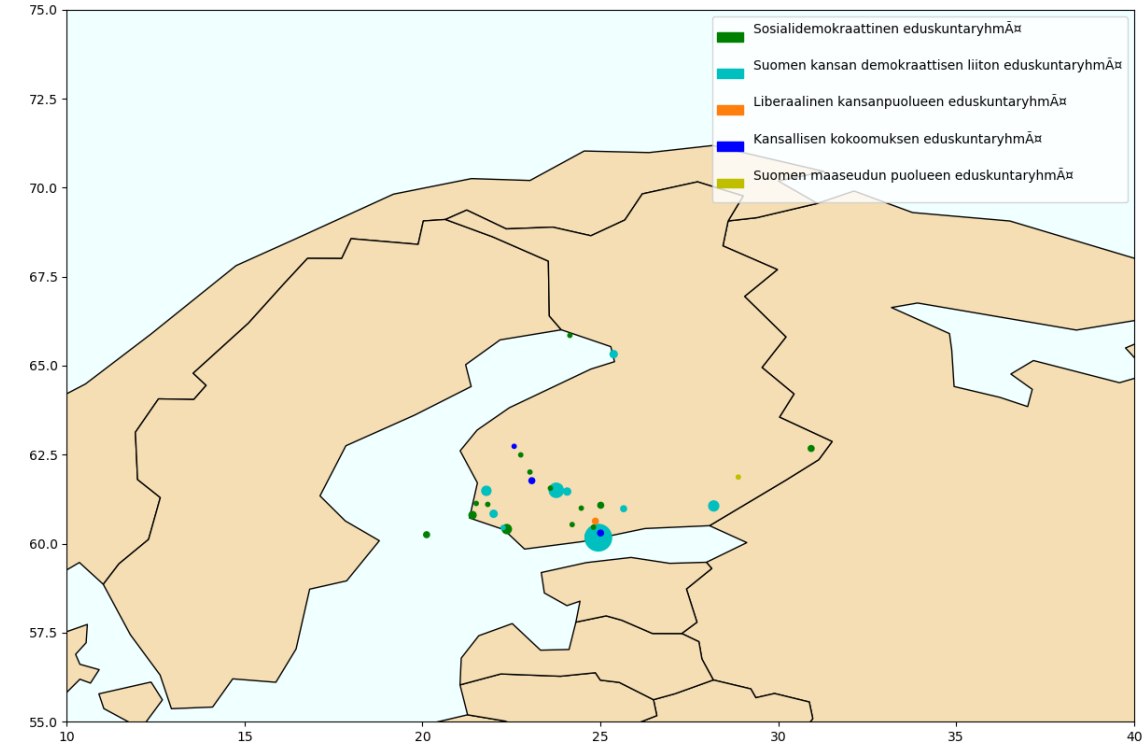
Finnish politicians in pictures – biases of Finna

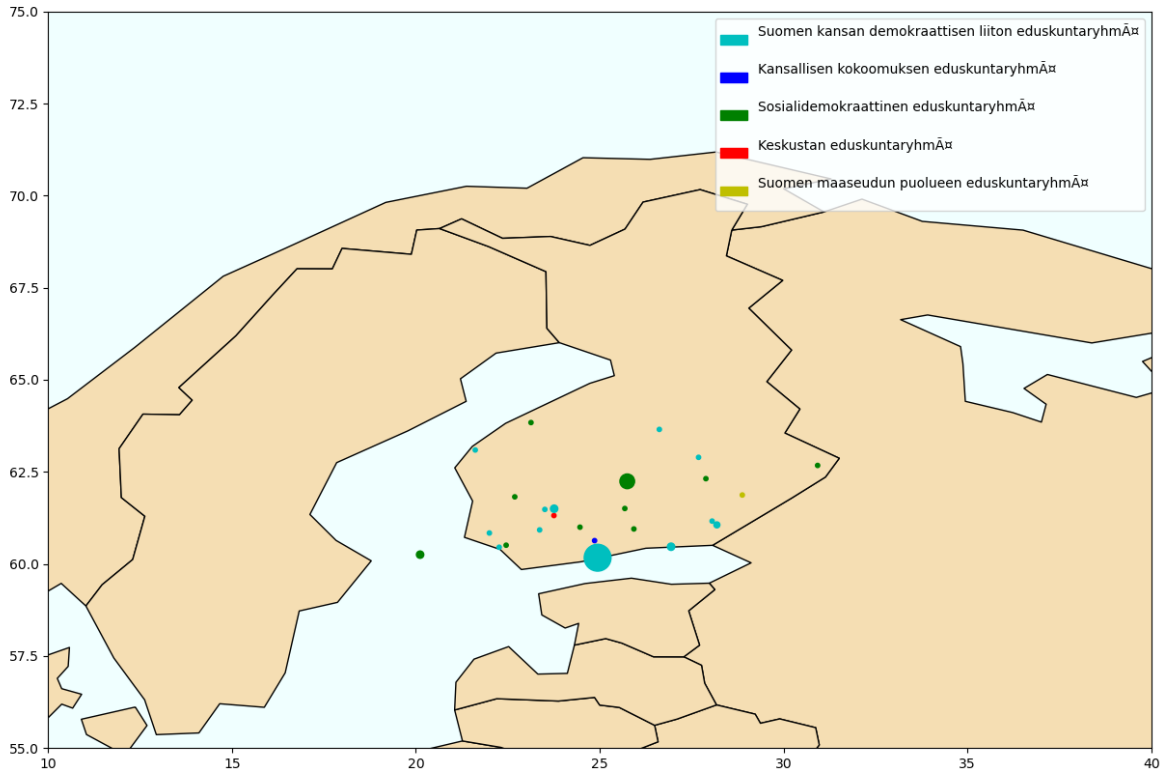
Initial research question

The inspiration for this project was the Finna database and my interest to see what could be done with it. Finnish history is not really my "own" field, but for the purposes of this project I thought I knew enough. I started the project with question: "Where did Finnish politicians travel during the latter part of 20th century?". My initial idea was to fetch all the pictures from a certain time period from Finna mentioning any politicians by name and then visualize where different politicians had travelled during different years. I tried to do this with two different timespans: 1972-1975 and 1991-1995, both featuring their own set of politicians in parliament. These are the results of that visualization:

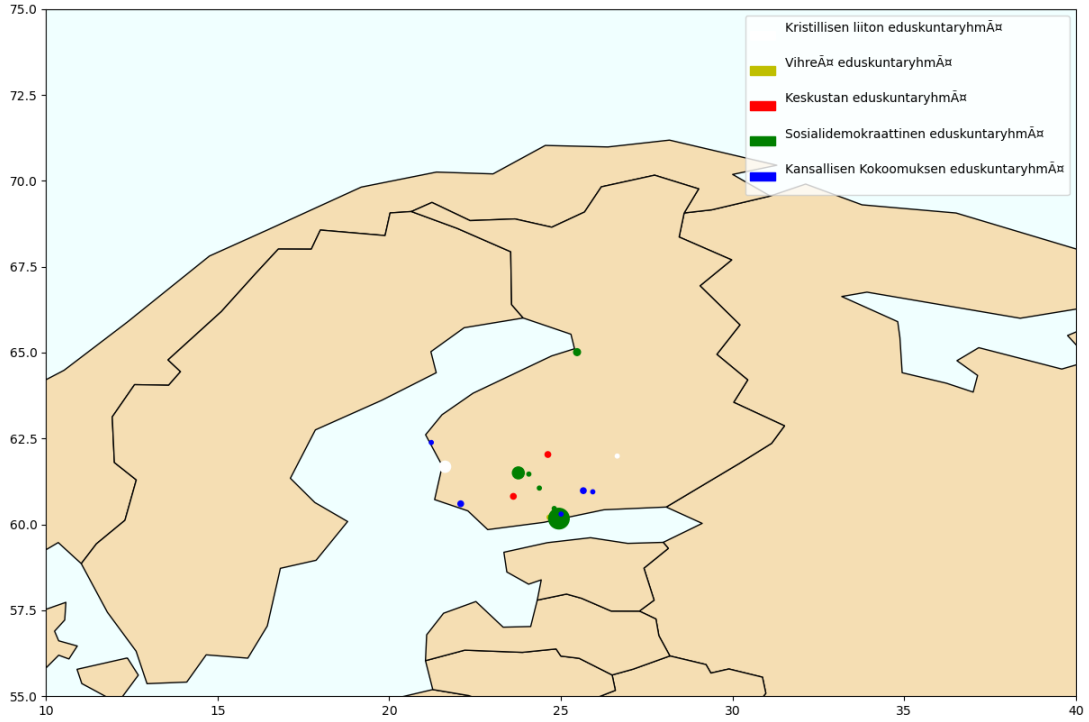

Kansallisen kokoomuksen eduskuntaryhmÃ¤
Sosialidemokraattinen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Suomen kansan demokraattisen liiton eduskuntaryhmÃ¤
Suomen maaseudun puolueen eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40

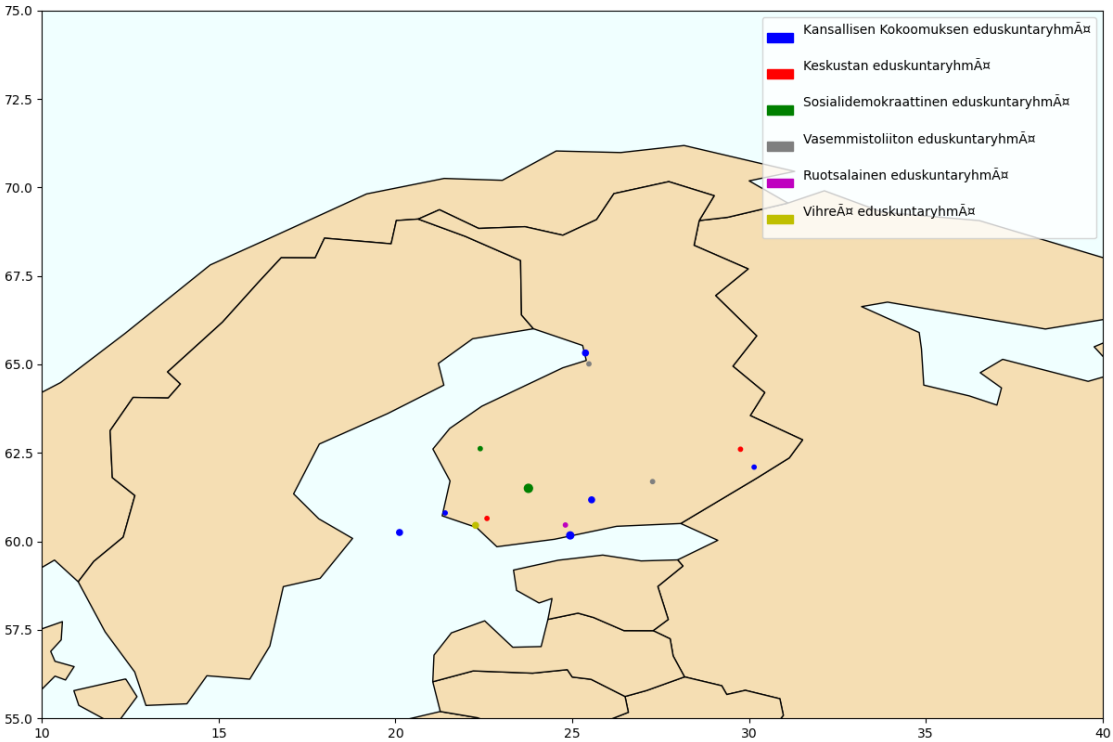

Sosialidemokraattinen eduskuntaryhmÃ¤
Suomen kansan demokraattisen liiton eduskuntaryhmÃ¤
Suomen maaseudun puolueen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Ruotsalainen kansanpuolue
Kansallisen kokoomuksen eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40

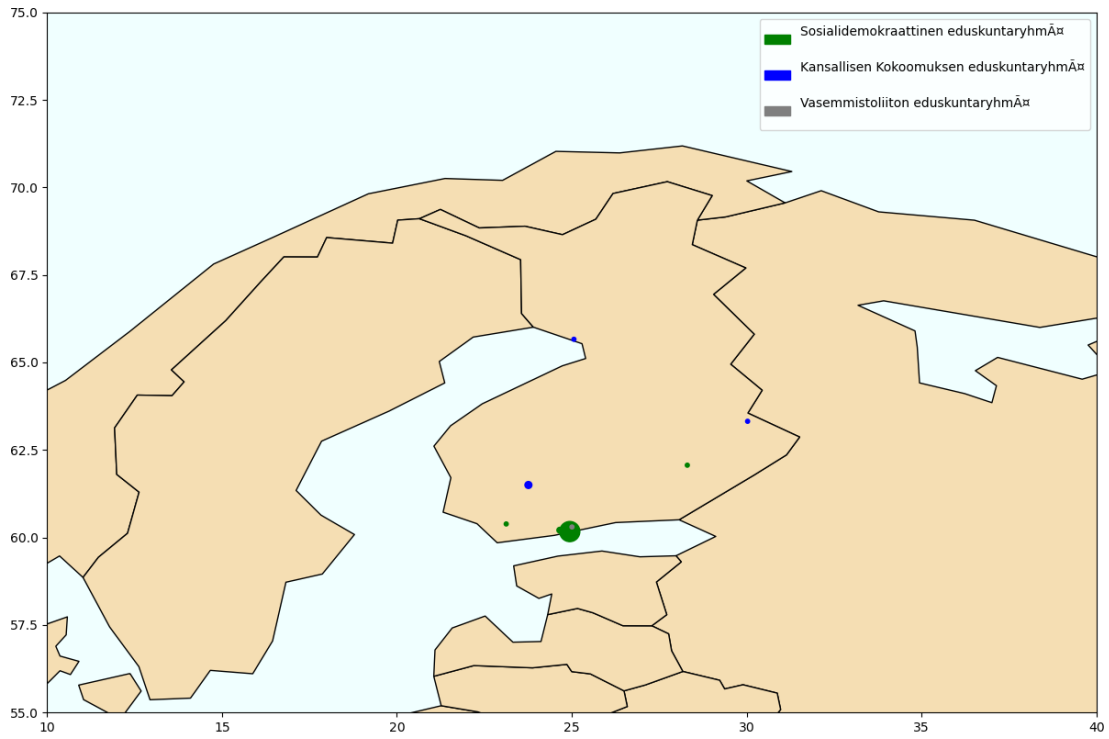


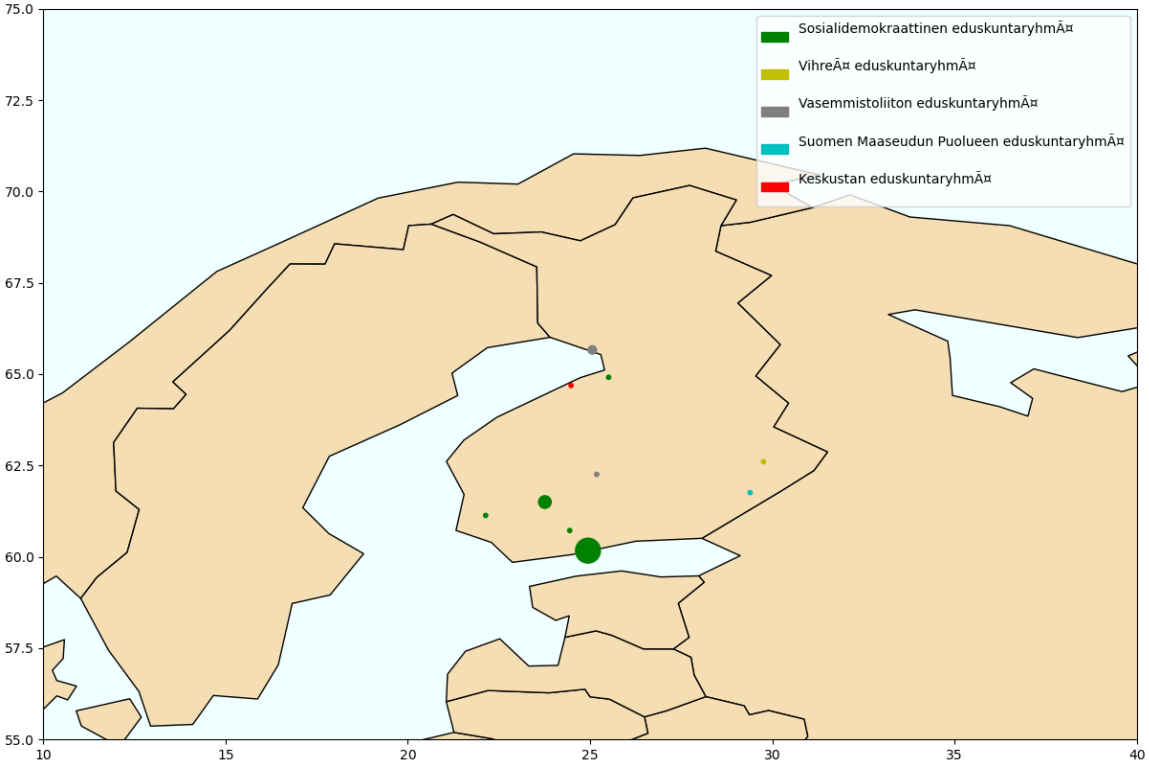


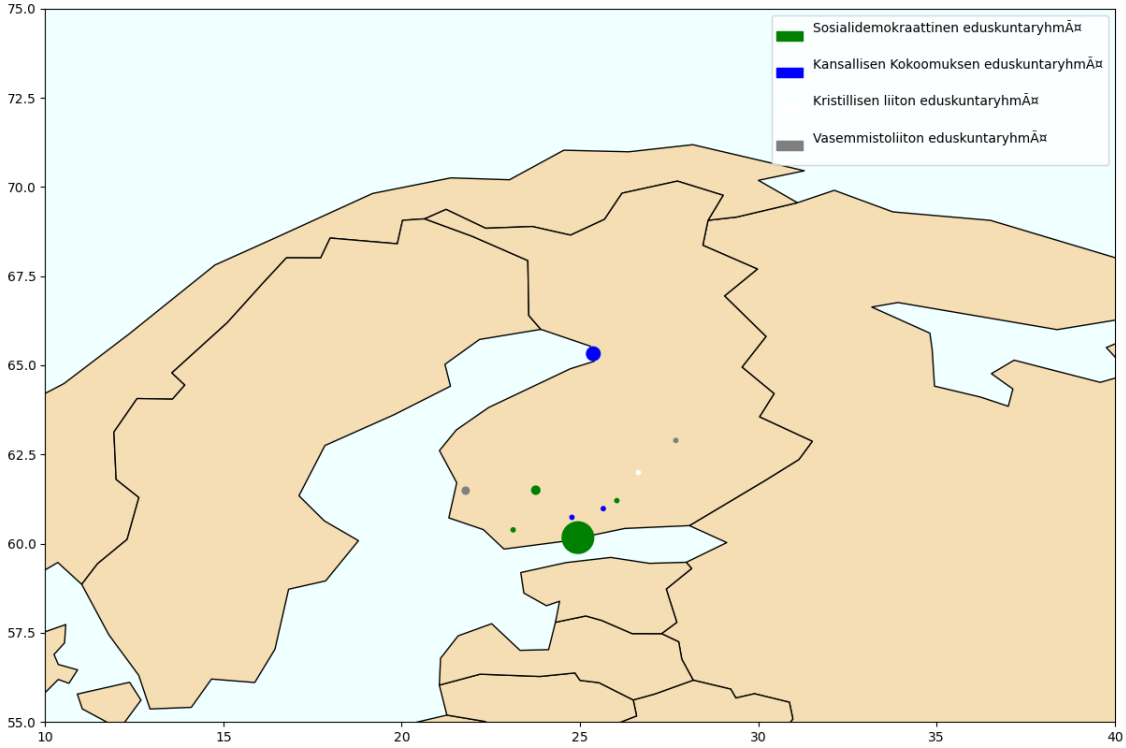
(Fig. 1 - 4.) 1972-1975


Kristillisen liiton eduskuntaryhmÃ¤
VihreÃ¤ eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Sosialidemokraattinen eduskuntaryhmÃ¤
Kansallisen Kokoomuksen eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40


Kansallisen Kokoomuksen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Sosialidemokraattinen eduskuntaryhmÃ¤
Vasemmistoliiton eduskuntaryhmÃ¤
Ruotsalainen eduskuntaryhmÃ¤
VihreÃ¤ eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40


Sosialidemokraattinen eduskuntaryhmÃ¤
Kansallisen Kokoomuksen eduskuntaryhmÃ¤
Vasemmistoliiton eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40


Sosialidemokraattinen eduskuntaryhmÃ¤
VihreÃ¤ eduskuntaryhmÃ¤
Vasemmistoliiton eduskuntaryhmÃ¤
Suomen Maaseudun Puolueen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
10
15
20
25
30
35
40



(Fig. 5 – 9.) 1991-1995

Interactive visualizations with more information can be generated by running the code provided, but one can see even with these the problem, when trying to answer the initial question. Firstly, there is not that much data and only cities which reliably have multiple hits are Helsinki and Tampere. Maybe one could find some patterns, for example, most of the hits are in the southern parts of Finland, but with so low amounts of hits and how the towns in Finland are dispersed, it is hard to actually claim anything historical about that.

Secondly, and more importantly, most of the hits come from the left-wing political parties. Each of the towns is coloured with the colour of the most prominent party in the data in that town. If almost all hits come from these politicians, one really can't claim that the data tells where the politicians at large travelled. Even if there were more diverse set of data and more political parties were represented, there still would have been problems. For example, it is highly unlikely, that all visitations of politicians would have been documented and there would have been a problem of which events (and maybe which politicians) were important enough to be photographed. I will circle back to these problems if and when they become important for the updated research question.

Updated research question

After discovering that the data really was not suitable for the initial research question, I thought that there were two ways that the research could go. Either stick with the initial question and focus only on the left-wing politicians or assess the biases of the database. I chose the latter. The updated research question is:

Which groups and individuals are prominent in Finna? How does that affect the way the political landscape of 1970-1995 is presented in Finna?

At this point I will delve into the data and tools used in my project and then come back to the results and analysis.

Data

The main data used in this project was fetched from Finna using their API. Finna is, in their own words, "a search service that collects material from hundreds of Finnish organisations under one roof".[1] Finna also holds information about many different formats of data from photos to objects. What makes the data interesting for this research question and challenging for the initial research question is the fact, that the data Finna contains is not theirs, but comes from different organisations which naturally collect and share different kinds of data. So, depending on which organizations have shared their data with Finna, the final dataset formed in Finna can be very different.

For my project data, I made queries to Finna API that searched for politicians' full names and returned all relevant metadata that might contain the date, the location or the providing institution of each picture. Since the code is available for reproducing this step, I will not in this paper go in greater detail about reproducing the queries, if they are not important for the discussion of analysis or biases. The code is decently commented so it should be somewhat readable also for those, who don't write Python themselves.[2] To get the data to analysable form, I parsed the metadata returned from Finna and created a data frame which included the politician in question, the location, the date, the party of the politician and the institution that provided the picture. If location or date could not be found, the picture could not be added to the data frame, since it would not add anything to the analysis. If a politician appeared in multiple pictures during the same day, only one was left to the data frame, since they were

[1] https://www.finna.fi/Content/about_finnafi. (referenced 8.12.2020)
[2] Code can be found from https://github.com/mikkosk/finna-metadata-mapper

probably from the same event. Unfortunately, some pictures only had the year, not an exact date, and in those cases, there can be some duplicates, because it was not obvious if the pictures were from the same event or not.

In addition to Finna's data, also smaller sets of data were used. To be able to visualize the politicians to the map I needed a set of data with names of Finnish towns and their coordinates, which I got from Wikipedia and formed into a .csv-file.[3] Similarly, I needed a file which contained the names of the politicians. I decided to look at two different sets of Finnish politicians in parliament. Those from 1972-1975[4] and those from 1991-1995[5]. The .txt-files I created follow a structure were a line with *::* followed by a parties name starts a new group, where all the politicians until the next such line will be placed.[6] Some politician changed parties during these years or left the parliament altogether. For this project's purposes, I deemed enough to just remove duplicates of politicians in different parties, but for more accurate results, one would probably have to change the program so that it stops counting the politician after they have left the parliament.

Reproducing the data should be quite easy. Only thing you need to do is run the Python code provided and fill in the locations where you have saved the .txt and .csv-files provided. Since the program fetches the data from Finna directly, it can of course change during in the future and on the hindsight, the data version used to produce this exact project should have probably been saved in its own file. To see what the code does in detail and to reproduce it, one can take look at the .py-file itself where everything is visible and shortly described.

Analysis

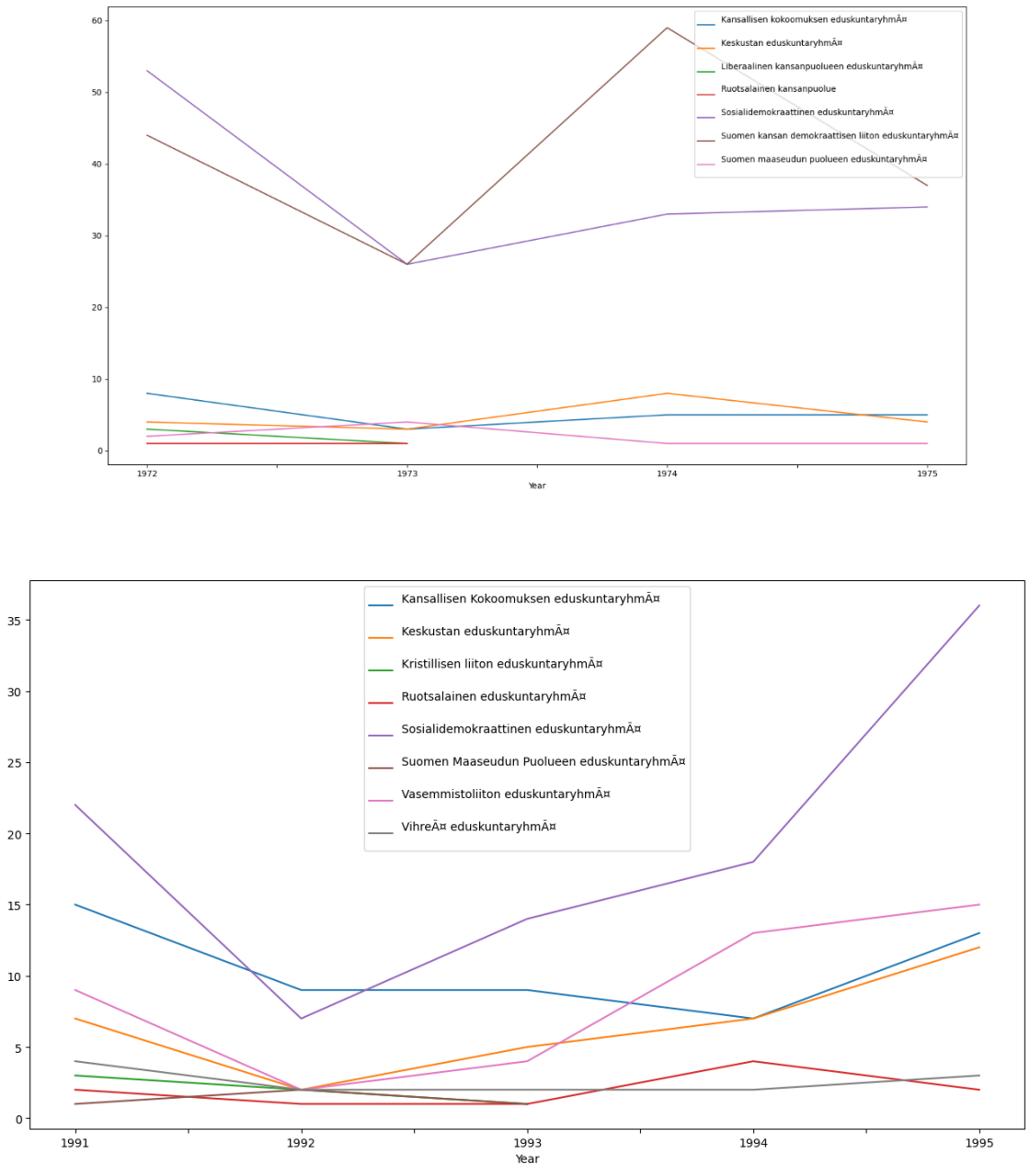
I already covered the short analysis of the initial research question. For the updated research question some new visualizations were needed since, the maps did not cover in detail how bad the bias of the data was. For that reason, I drew visualizations for following things:

1. The number of pictures of party members
2. How much pictures were from each institution?

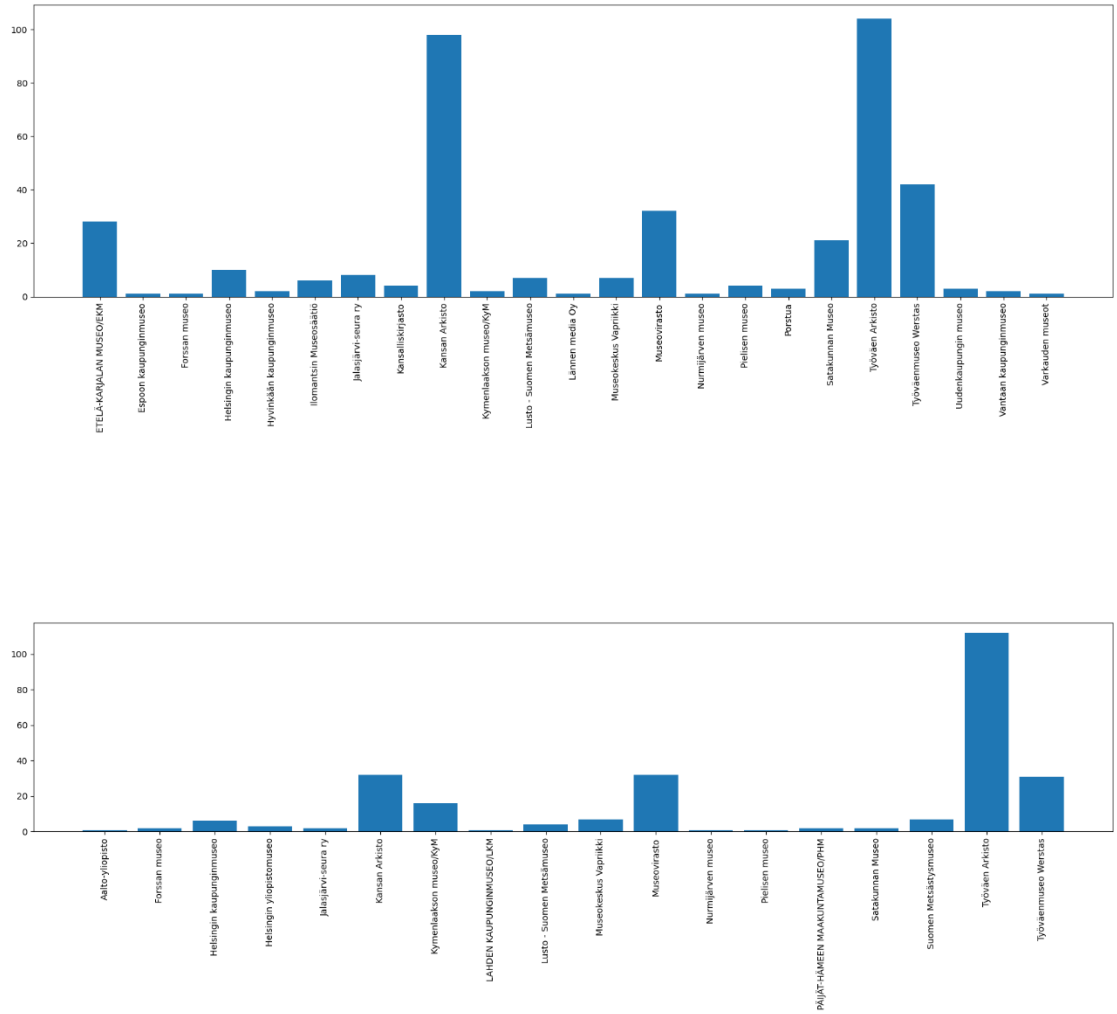
[3] https://fi.wikipedia.org/wiki/Luettelo_Suomen_kuntien_koordinaateista
[4] https://fi.wikipedia.org/wiki/Luettelo_vaalikauden_1972%E2%80%931975_kansanedustajista
[5] https://fi.wikipedia.org/wiki/Luettelo_vaalikauden_1991%E2%80%931995_kansanedustajista
[6] The same tool can thusly be used to visualize other themes and groups or even countries by changing these files. Some small changes will also probably have to made to the visualization.

3. Which parties did the pictures of the institutions feature?
4. Which individuals are present?

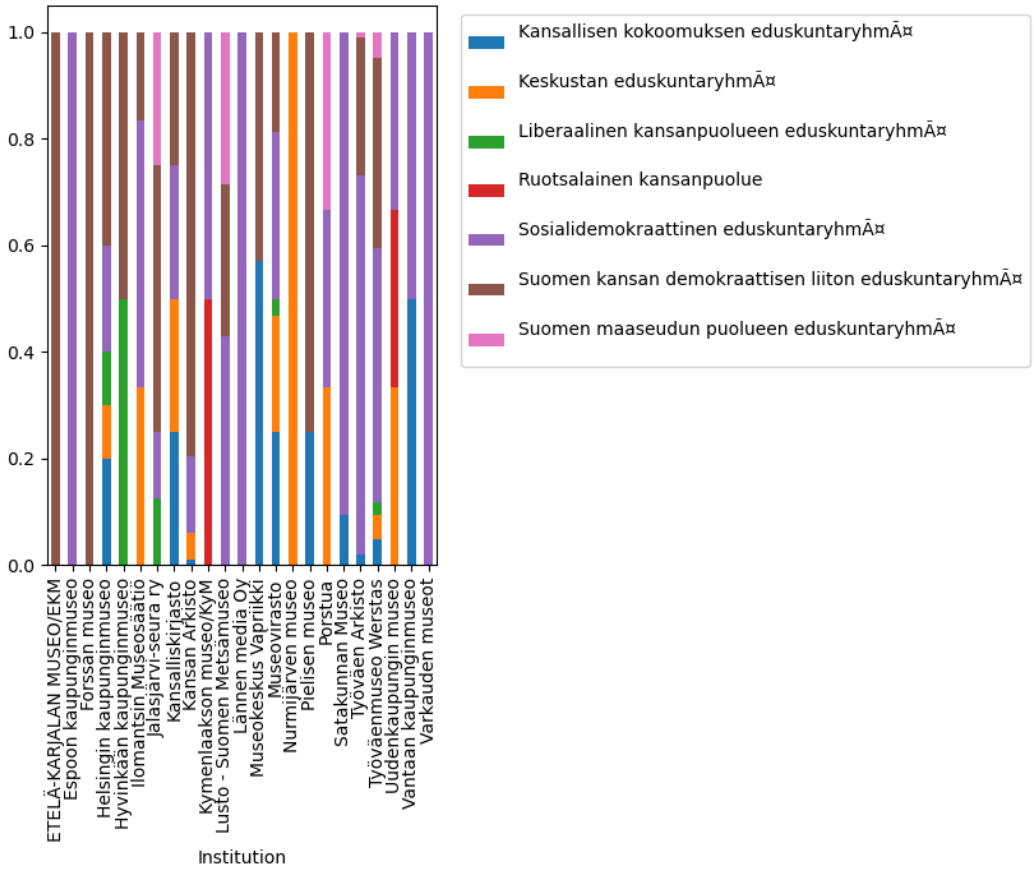
For comparison, there is still visualizations for both time periods, 1972-1975 and 1991-1995. The pictures can also be found in the project folder, where they are in bigger size.

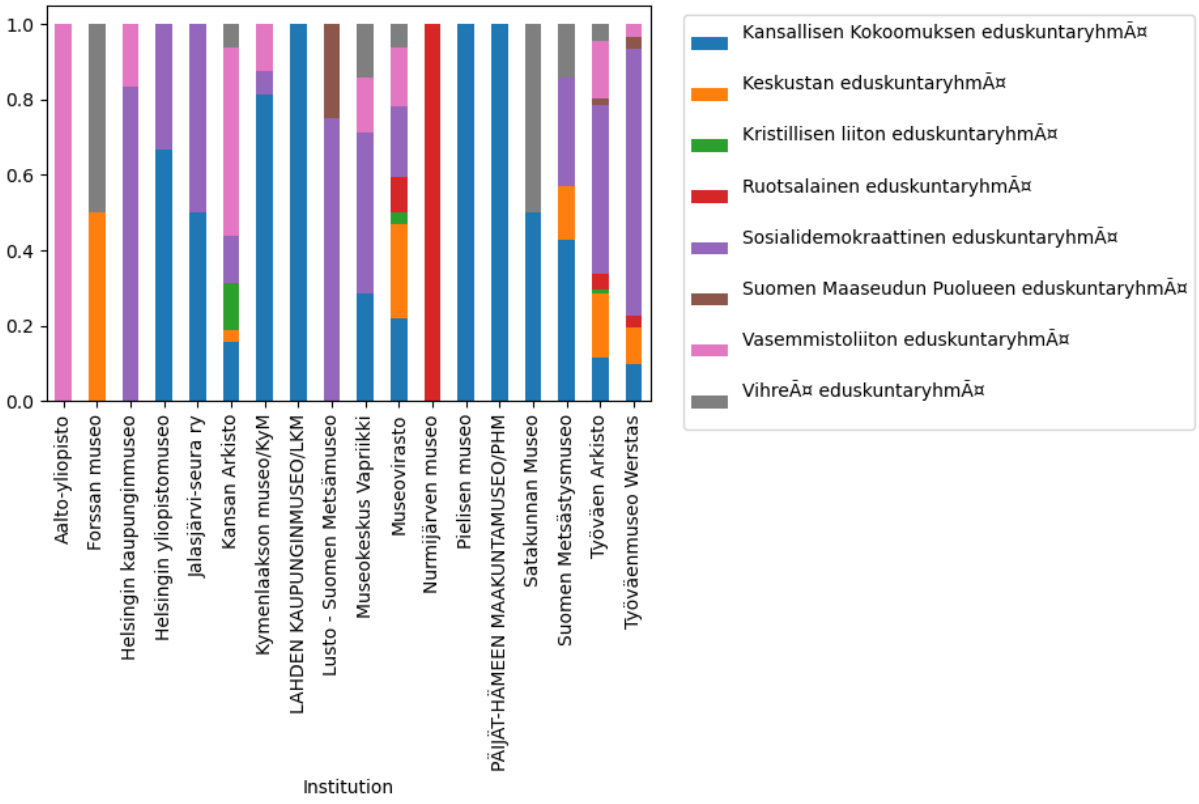


(Fig. 10-11.) Count of pictures by party

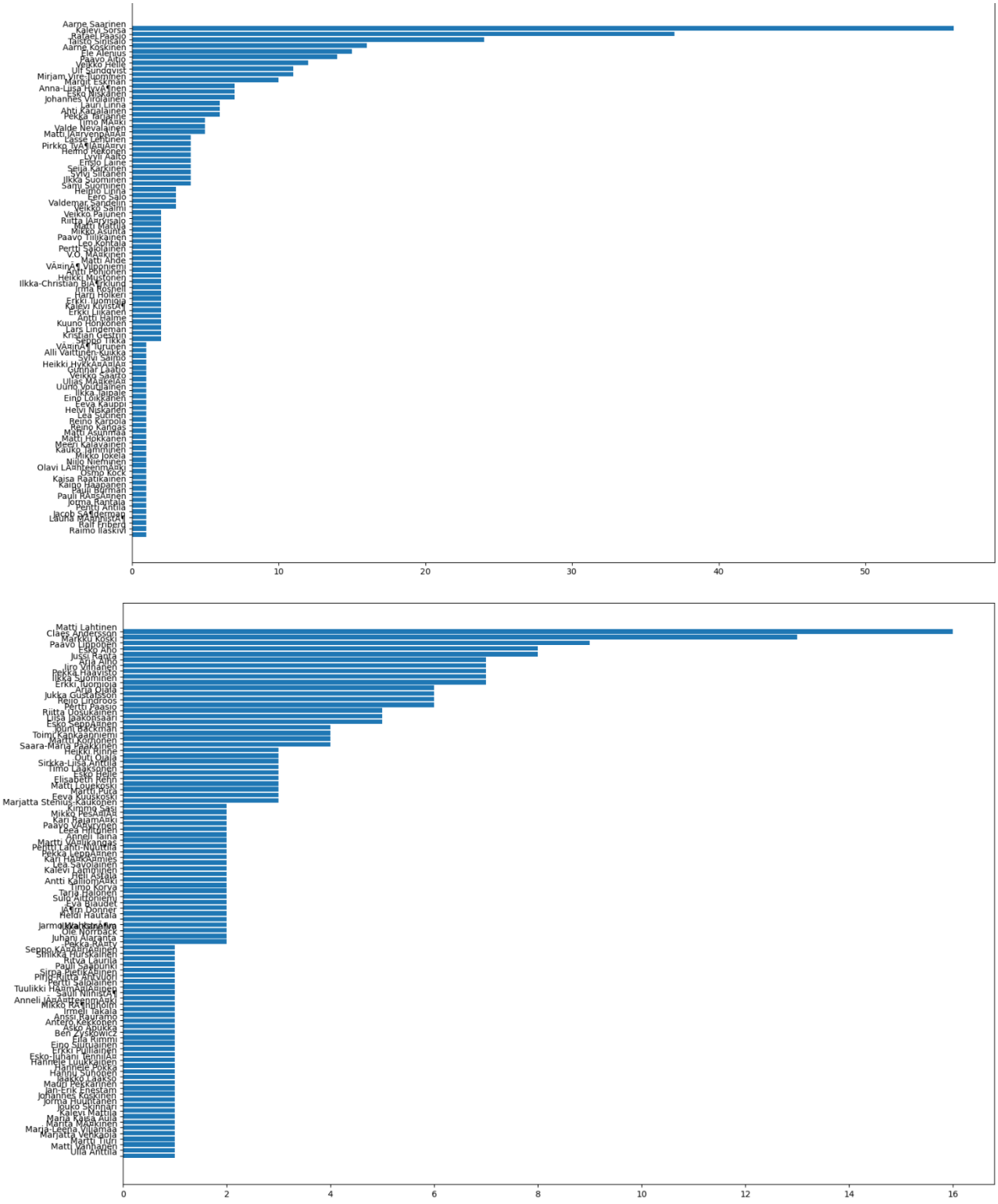


(Fig. 12-13.) Count of pictures by institution


1.0
0.8
0.6
0.4
0.2
0.0
ETELÄ-KARJALAN MUSEO/EKM
Espoon kaupunginmuseo
Forssan museo
Helsingin kaupunginmuseo
Hyvinkään kaupunginmuseo
Ilomantsin Museosäätiö
Jalasjärvi-seura ry
Kansalliskirjasto
Kansan Arkisto
Kymenlaakson museo/KyM
Lusto - Suomen Metsämuseo
Länsnen media Oy
Museokeskus Vapriikki
Museovirasto
Nurmijärven museo
Pielisen museo
Porstua
Satakunnan Museo
Työväen Arkisto
Työväenmuseo Werstas
Uudenkaupungin museo
Vantaan kaupunginmuseo
Varkauden museot
Institution
Kansallisen kokoomuksen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Liberaalinen kansanpuolueen eduskuntaryhmÃ¤
Ruotsalainen kansanpuolue
Sosialidemokraattinen eduskuntaryhmÃ¤
Suomen kansan demokraattisen liiton eduskuntaryhmÃ¤
Suomen maaseudun puolueen eduskuntaryhmÃ¤


1.0
0.8
0.6
0.4
0.2
0.0
Aalto-yliopisto
Forssan museo
Helsingin kaupunginmuseo
Helsingin yliopistomuseo
Jalasjärvi-seura ry
Kansan Arkisto
Kymenlaakson museo/KyM
LAHDEN KAUPUNGINMUSEO/LKM
Lusto - Suomen Metsämuseo
Museokeskus Vapriikki
Museovirasto
Nurmijärven museo
Pielisen museo
PÄIJÄT-HÄMEEN MAAKUNTAMUSEO/PHM
Satakunnan Museo
Suomen Metsästysmuseo
Työväen Arkisto
Työväenmuseo Werstas
Institution
Kansallisen Kokoomuksen eduskuntaryhmÃ¤
Keskustan eduskuntaryhmÃ¤
Kristillisen liiton eduskuntaryhmÃ¤
Ruotsalainen eduskuntaryhmÃ¤
Sosialidemokraattinen eduskuntaryhmÃ¤
Suomen Maaseudun Puolueen eduskuntaryhmÃ¤
Vasemmistoliiton eduskuntaryhmÃ¤
VihreÃ¤ eduskuntaryhmÃ¤

(Fig. 14-15.) Percentage of parties by institutions



(Fig. 16-17.) Count of pictures by individual

These visualizations further strengthen the observations about the data bias that one could see even in the vaguer map visualizations. One can see that in both time periods the number of pictures of left-wing parties (SDP, SKDL, Vasemmistoliitto) are multiple when compared to others. Even though these parties had remarkable share of the seats in the parliament – 92/200 in 1972 and 67/200 in 1991 – it does not explain such high numbers for these parties and low amount for other big parties like Kokoomus – 34 seats in 1972 and 40 in 1991.[7]

It seems like the data is vey biased towards these left-wing parties because of the institution that are providing the data. During both time periods three of the five biggest institutions compared in pictures provided, were *Kansan arkisto, Työväen arkisto and Työväenmuseo Werstas* all of which can be described as labour archives. Other archives have relatively little impact overall since these three archives provide the clear majority of all pictures.

The fact that the most important archives are labour archives does not necessarily mean that the data is biased, but when we look at the percentages of which parties the pictures of these institutions feature, we can see the bias. Most of the pictures from these institutions feature pictures of politicians from one of the left-wing parties. Other big parties, Kokoomus and Keskusta, are also featured, especially in 1991, but not in realistic proportions. The volume of these institutions accompanied with their diversity leads to Finna providing, in this sense, a very biased set of data, which maybe could be used to say something about these parties, but not to examine the Finnish field of politicians as a whole.

The labour archives seem to be the only ones that have enough material to at least entertain a thought about using computational methods. Having said that, it seems like that some institutions share its problem with having pictures from only certain parties, but some institutions could probably form a quite good set of data, if more pictures exists and would be added to Finna. For example, *Etelä-Karjalan museo* is one of the biggest institutions in 1972-1975 period, but it also has only pictures of SKDL, even though it does not have similar labour archive status as the other mentioned institutions. Then again, the pictures of *Museovirasto* seem to follow – for both time periods – follow quite well the actual proportions of the parliament. *Museovirasto* also has more than few pictures, so it is likely that it is not only a coincidence. This, of course, does not instantly mean that the pictures are

[7] Tilastokeskus: *Suomen tilastollinen vuosikirja 2004*, s. 562–563. Helsinki: Tilastokeskus, 2004. https://www.doria.fi/bitstream/handle/10024/90078/yyti_stv_200400_2004_net.pdf?sequence=1&isAllowed=y (referenced 8.12.2020).

useful for purposes like my initial research question. Many of the pictures from Museovirasto seem to be portraits of the politicians and so do not tell anything about the travels of them.[8]

Lastly, when looking at the individuals that were prominent in the data, one can find surprising and not so surprising individuals on the list. For example, all the prime ministers (Paasio 1972, Sorsa 1972, Aho 1991) can be found on the list as well as party leaders (e.g. Andersson, Holkeri) and other important, especially left-wing, politicians (e.g. Taisto Sinisalo and Aarne Saarinen). These seem to be pretty much in line with the other data. Mostly featuring left wing politicians, but also small amounts of other big parties. The weird thing about this data is that there seem to be quite unimportant, non-left-wing politicians like Matti Lahtinen (Kokoomus) and Markku Koski (Keskusta) with very high number of hits. Also, when manually searching from Finna, one cannot find any pictures of the actual politician.[9] This leads to last part of this project, all the bias problems when going through Finna's data the way I did are not Finna's fault. There are also several problems with varying importance in my code, which will be discussed next.

Problems in the pipeline

I think my pipeline captures quite well the biggest problems with the data and its bias, but that is mostly because, to see – at least the most obvious – bias of the data, one really does not have to have very high tuned tools. Still, I think this works quite well for the updated research question. Although the research question got answered at least somewhat satisfyingly, there is still many problems with the pipeline. I will not focus anymore on the problems of the data, because that is intertwined with the research question and was already discussed.

The first major problem with the pipeline is creating the queries to Finna API and parsing through the returned data. To make queries about people one must search them by their names. Since names are not unique searching by name is not guaranteed to only return pictures with a certain person. That is why there was so many pictures of people like Markku Koski, who really should not have - and actually was not – featured in the fetched pictures.

---

[8] Example portrait. Similar portraits are prominent in the pictures provided by *Museovirasto*
https://www.finna.fi/Record/museovirasto.6AC93AC76530267B41A07F8A7B0AB4DD?lng=fi

[9] Example search for Markku Koski
https://www.finna.fi/Search/Results?filter%5B%5D=%7Eformat%3A%220%2FImage%2F%22&filter%5B%5D=search_daterange_mv%3A%22overlap%7C%5B1991+TO+1995%5D%22&join=AND&bool0%5B%5D=AND&lookfor0%5B%5D=Markku&lookfor0%5B%5D=Koski&type0%5B%5D=AllFields&type0%5B%5D=AllFields

Instead other people named “Markku” and places with “Koski” in their name were returned. Common names and names that can also be used in another context were not handled well. I probably should have been even more strict in what form the names should appear, but that would have missed some actual hits, which would not have been good, since the data was already small. The problem with full names, which I used, is also that, not everyone is always called with their full names. For example, searching with only last name, would probably yield more correct results than the current way, but it would also include so much wrong results that they would be useless.

Other problem which ties into a similar problem is that Finna’s metadata is flaky. Not every organisation includes same metadata and in same form. That is why I decided that I would not try to fetch the information about the date and the location from certain fields that were returned in JSON-form and rather I would just treat the whole JSON as a string and try to see, if I could match the date and the location from that string. This worked somewhat well, but there were once again problems, which not all are easy to solve. For example, some place names are so short or general that they are often included in other names or random words. These are places like *Ii* or *Sund*. Latter of which seems to, for example, have matched with the name of the politician *Ulf Sundqvist*. One solution could be to search the town names in lemma form and make sure that they are not followed by any alphabets. This would probably work with English, but with Finnish there comes a problem since the town name like Espoo can be followed with several different endings e.g. Espoo*ssa*. This was problem even with my less strict matching since some names, like Helsinki, don’t match with their lemma form if they are inflected e.g. Helsingi*ssä.*

There could also be other problems with how the town list was used. As far as I know, the data I used to create it included all the current and former towns in Finland, but some could still have been missed. The bigger problem is that towns are not the only way to determinate locations. I could imagine that there are pictures with even more precise location that are missing the town name, and so my code did not catch those. From the visualization point of view, more densely populated areas with more towns can probably look busier than their counterparts were towns are further away from each other. Even though, it was not really a problem with my research question, it should be noted, if tackling research question like my initial question.

After the part of fetching the data and turning it into a data frame, I think my pipeline is much more solid. There should not be any big problems when creating the visualizations and the visualization that were made all tell something about the research question. There could of course always be more analysis done, but for a project this size I hope this catches enough angles to answer the research question.