NSM 2022
NORDIC STATISTICAL MEETING

Profiling library users in Denmark – Combining daily reports with demographics

Maria Pedersen, Statistics Denmark, mrp@dst.dk

Paul Lubson, Statistics Denmark, pal@dst.dk

Abstract

This paper examines the Danish statistics on library services on an individual level. The paper outlies the cooperation established between Statistics Denmark and the public library sector in Denmark. The end result is automated daily reports provided from the public libraries to Statistics Denmark containing information on an individual level linking users with their library borrowing. The setup with regards to handling data is discussed as well as how privacy of the users is ensured. Data covers both the traditional public libraries and the digital ones.

Next, information on the users is combined with demographical statistics in order to profile the user. Age, sex, education, income and social status are among the demographics linked to the library user and the end result is a detailed profiling of the behavior of different demographic segments. This data is produced both on a national level and on a municipal level. Furthermore, the address of each library user is used to profile the user base of each public library. We profile both their local user demographics, but also through interference the demographics of the locals that are not library users.

Finally, examples of how the daily data mapped use of libraries during the pandemic are shown.

Keywords: Library statistics, Demograpic profiling. Automated big data.

1. Introduction

The Danish library statistics have recently been expanded to profile loaners of library materials and their demographics.This paper starts with an introduction to what motivated the expansion of the library statistics in Statistics Denmark. This is followed by a discussion on the actual data collection process and the link to the demographics in the DST registers. Afterwards our paper presents the results from the project including our findings. We also show how the libary statistics covered the COVID-19 crisis using daily reports. At last,we discuss how this data collection

method can be used on other types of data and discuss further expansions of the current statistics.

In Denmark there are just under 500 public libraries. These had almost 20 million. loans of different materials from the fysicial branches in 2020 and also 7 million loans featuring either electronic books or audiobooks from their digital platform, eReolen. In the following we wil present how the library statistics in Denmark went from a yearly data collection and a publication that only included information regarding take-outs from the libraries with no information about the loaners to a more dynamic data collection and publication that links the loans to the actual user of the library. The end-result is a statistic that profiles library users on a wide range of demographics, which can also be linked with metadata that describes exactly what materials have been taken out

2. About the study

We start by describing why the statistics came into play and then proceed to show how the data collection and data processing takes place. In addition, we will explain how the needs of the libraries have created new opportunities through data. Finally, we will describe how the demographic and geographical registers in Statistics Denmark are linked with library data.

2.1 GDPR and a need to look in new directions

What inspired this project was the new GDPR-guidelines in 2016, which made the protection of personal data to a top priority amongst all Danish and European enterprises. The consequence was that data about library loans on an individual level were not allowed to be stored more than 30 days at the libraries and a lot of valuable data would be lost, if no action was taken. An agreement was therefore made between the Danish libraries, the Ministry of Culture and Statistics Denmark. The aim was to secure the library data with in of Statistics Denmark that by law has permission to keep and process data about the individual loans for research purposes. The other part of the agreement was that Statistics Denmark would deliver high-frequency aggregated results to the individual Danish public libraries.

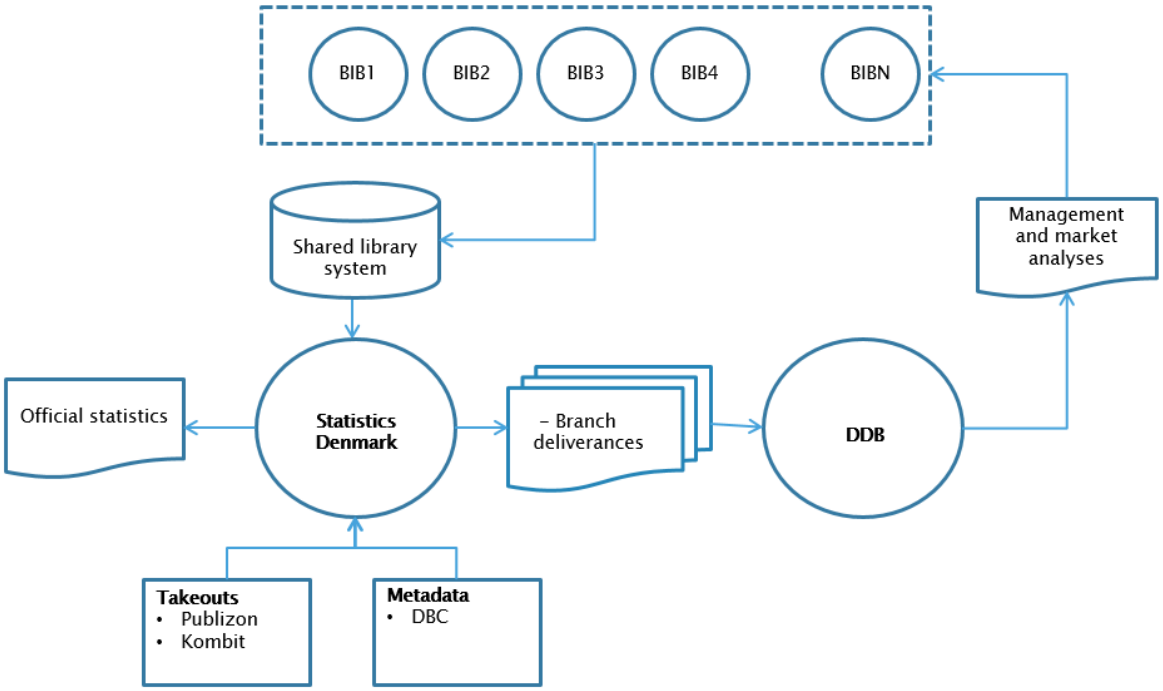
A pilot project conducted in 2017 investigated the possibilities of linking individual information about the users of two public libraries with demographic information from Statistics Denmark's registers. The project involved a survey of the users of libraries in Ballerup and Copenhagen municipality. The pilot project showed that it is relatively simple to link information about both users of electronic and physical resources with information in Statistics Denmark registers. The established database contained a number of tables that described the users in relation to selected background variables, such as age, gender, ethnicity, income and distance to library branch. Also an analysis of market shares was made measuring the amount of library users in relation to the inhabitants in the municipality.

2.2 The data collection

The data collection consists of daily reports with microdata on an individual level from each library system in Denmark and the digital platform, eReolen. A data transfer has been setup, so that data from both the public libraries and eReolen are transmitted to Statistics Denmark every night through a secure FTP gateway. This library data contains the social security number of the user coupled with time of the takeout and and identification number that identifies exactly what material the user loaned. This data is securely stored in Statistics Denmark and the socal security number is replaced with an anonomized identificator (Person ID), so that the people working with the data do not have access to the social security number.

The ownership of the library data is located with Danskernes Digitale Bibliotek (DDB). Statistics Denmark get data from their IT-provider Kombit for the traditional loans and Publizon for the the electronic loans. Apart from receiving library data nightly from DDB, we also have received access to the DBC metadata through an API in order to know how the library materials are coded and identified. DBC or Dansk Biblioteksсenter is the agency tasked with keeping track of metadata identifying the materials that are in the libraries. We will describe how we handle data later, but here it suffices to state that once we've received the data as described then we link it with demographic info on each library user and create reports from national level down to the level of each public library. These reports are then delivered to DDB and they

distribute them to the libraries. We also publish official statistics on a national level in the Statistics Denmark Statbank. That process is ilustrated below.



BIB1 to BIB4 are representing different public libraries. These share the same library system in Denmark. Since the first official publication of the new expanded library statistics in 2020, Statistics Denmark has built a database with over 50 miliion datapoints from the libraries. This database is on average expanded with 100.000 datapoints each day. Each datapoint consists of information regaring the place and time of the takeout , the metadata identification-number of the material and the social security number of the loaner.

2.3 Clarification of the needs from the libraries

What was of importance to the public libraries is being able to profile their users, so they know what demographics make use of them, but also to be able to profile their non-users. Therefore we not only profile the segments that have loaned materials, we also profile their local catchment area, so they can see if their userbase reflects their catchment area. We've defined that area as half the distance to the closest other public libray with a minimum of a 1 km radius. In their report they receive number of materials taken out with a demographic breakdown on the loaners together with the

number of loaners that made a loan and also the number of people in their local area in that demographic segment.

2.4 Linking of data to the demographic and geographic registers

Statistics Denmark has an extensive collection of register data which have been collected since 1970's which apart from being used to make official statistics als si made available to researchers and for special deliveries to forexample the ministries or the press. To link the registers within more than 250 different subject-areas together, a unique identification code for either the people or enterprises are used. For people it used to be their social security number, but these days that is hidden and replaced with an anonomous identifier (PersonID).

To produce the official statistics and branche deliverances the library data is linked with data from the population and demography registers by using their PersonID. The current links that we currently produce statistics on are age, gender and origin (Danish decent, immigrant or decendent of immigrants). Then we look at their education level, employment status and income. Finally we link to their residential adress, which allows to geografical stratas and also calculate distance from the user to the library.

Apart from statistics on the individual users we also do statistics on households.

2.5 Quality of data and coverage

We have full coverage of all takeouts from the libraries and receive data on a daily basis, therefore there are few concerns about the quality of the data in terms of quantity. Quality issues are centered mostly around delivery failures to Statistitics Denmark from our data providers. So far we've only experienced internal failures when moving the transmitted data into our databases, but no failures in receiving data. When establishing the statistic we put effort into ensuring that all the relevant libraries where accounted for, at some municipalities apart from the public libraries also transmitted take out from school libraries, which is outside the scope of this statistic. We are in ongoing dialogue with our users with regards to our coverage. DDB is not allowed to keep information about individuals due to GDPR regulation, but

they still keep track of library takeouts, so our data aggregates are regularly checked. Thus there is littel in terms of uncertainty in terms of measuring takeout. There is some uncertainty when linking to our demographic registers. This process will leave us with takeouts that can't be matched. We do not try to estimate the values of takesouts that can't be matched with our registers, but instead group them as „unknown". With regards to variables that are linked to our census (age and gender), in the first quarter of 2022 we were able to match 99.5 percent of our physical takeouts and 97.9 percent of our digital takeouts. The digital takeouts are harder to match as there are several options to logon to our digital platform Ereolen.

3. Results

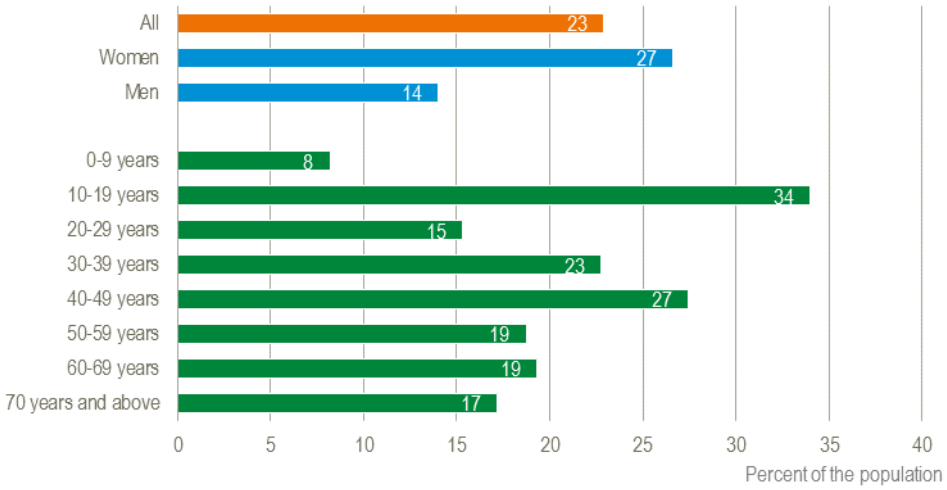
3.1 Basic results about the demography of the library users

In the following, we will show the results from 2021 and at the end of the chapter, we will unveil how library data helped to show hoarding effects during the COIVD-19 shutdowns in Denmark.

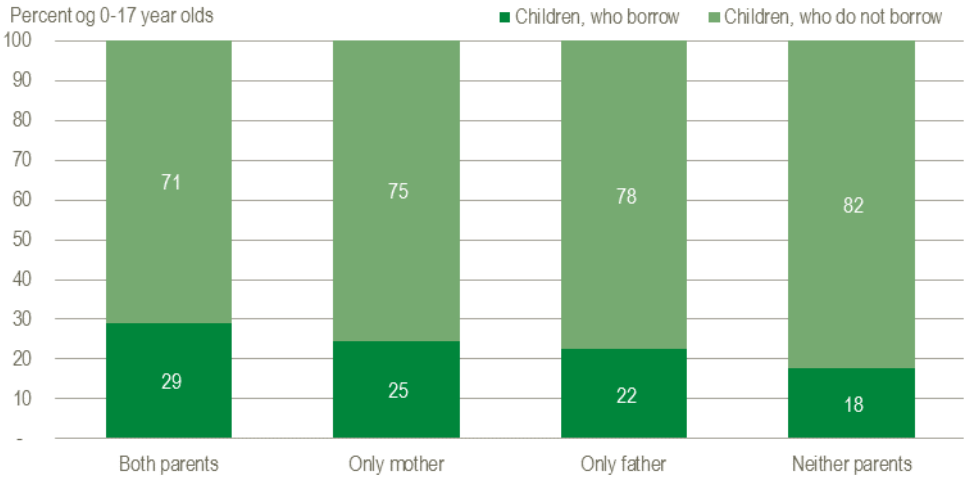
Results show that 23 percent of the danish population loaned library materials from the country's public libraries and eReolen during 2021. The library users loaned library materials a total of 24.7 million times This is equivalent to each borrower averaging 18 materials during the period.

Regarding gender, library takeouts are most prevalent among women. In 2021 there were 785,700 female loaners, corresponding to 27 percent of the women in the population. In the same period, there were 410,000 male loaners corresponding to 14 percent of the men in the population. In the physical libraries women would on average loan 22 materials, while men would loan 17. At eReolen, men and women borrow almost the same amount of materials. The female borrowers clicked for 12 materials, while the men settled for 11. The reduced difference may be due to municipal quotas on the amount of digital lending per person. There is a monthly cap on how many electornic matrials you can take out, which is typically a maximum of 3 materials.

Looking at the different age group the smallest share of library loans is seen among the 0-9-year-olds, where 8 percent of the age group lend materials. In comparison, that figure is 34 percent in the 10-19-year-olds group and15 percent of the 20-29 year olds. For the rest of the population, between 17-27 percent lend a material from the libraries in 2021, where it was most popular with the 40-49-year-olds. It is to a greater extent the younger library users who use eReolen, while the older part of the population sticks to the physical libraries.

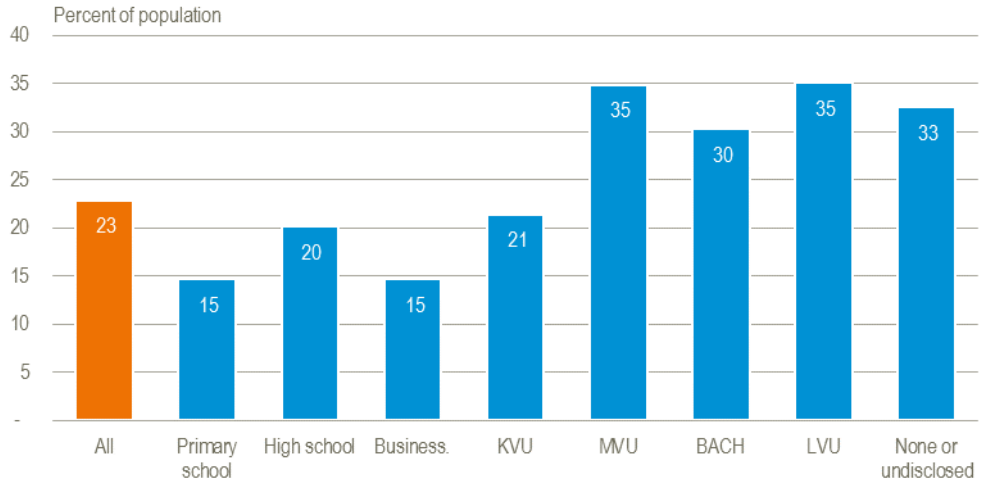


It is possible through our registers to link the individual with their respective mother and father, we can investigate whether a child under the age of 18 is affected by their parents' borrowing behavior. By examining data from 2021 coupled with the parents' own borrowings, the results showed that 29 percent of children lend materials at the library at least once during the year, if both their parents have also done so. If neither parent had visited the library then that number is 18 percent. In addition, there was a slight difference in the behaviour dependend on which parent went to the library. In the survey, we found that 25 percent of children borrow materials at the library if only the mother used the library, where the result was 22 percent if only the father was going for reading materials.



Loans of library materials are far more widespread among people of Danish origin than immigrants and descendants. During 2021 a share of 22 percent of the persons of Danish decent loaned library materials, while that dropped to 18 percent among descendants of immigrants and further down to 12 percent among first generation immigrants.

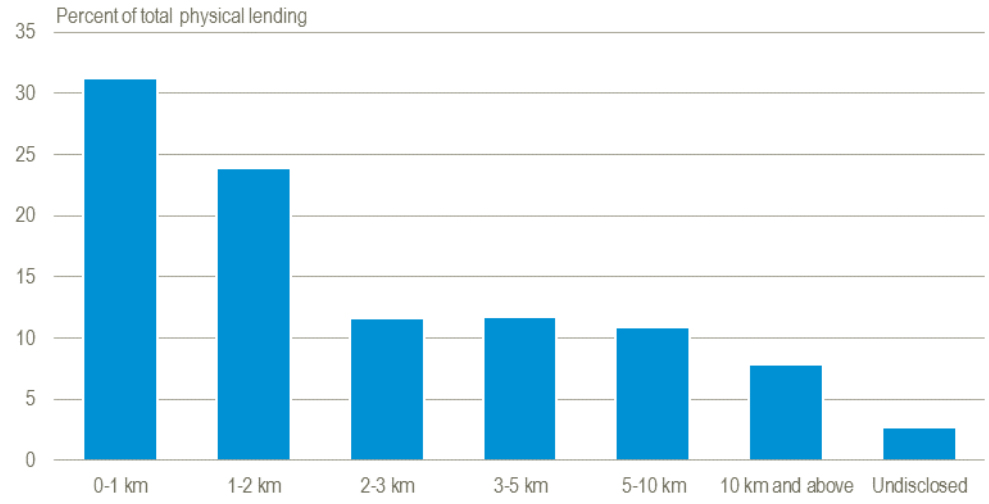
It is far more common to borrow library materials in the part of the population that has completed a bachelor's or higher education than in the part of the population that has not. A share of 34 percent of population with a bachelor's degree or higher loaned library materials. In comparison, it was 21 percent of the population with a short higher education did, 16 percent with a high school or vocational and 15 percent of those who have at most completed a primary school or preparatory education. People with short educations borrow more digitally, while highly educated people prefer the physical libraries.



The proportion of loaners is largest among the part of the population that is in employment and least among pensioners. In 2021 23 percent of the employed loaned library materials, while it was 15 percent of the pensioners. Among students, the proportion of loaners was 29 percent. Although the proportion of patrons is lower among retirees, they are the ones who borrow the most library materials, while students borrow the least. The retirees loaned an average of 22 materials during the period, while the students took out on average 11 materials. Borrowers in employment took out 20 materials, the unemployed where at 11 materials, while others and people outside the labor force would loan an average of 16 materials.

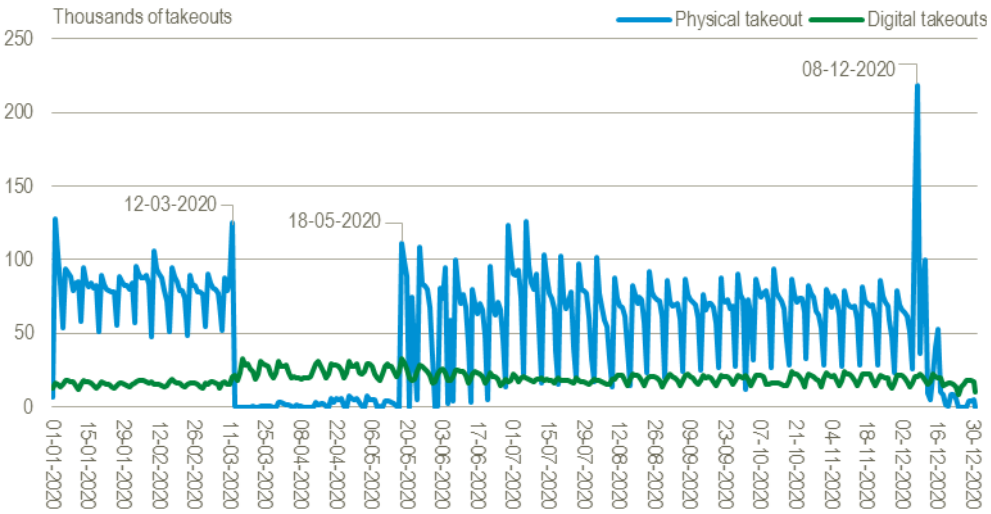
The proportion of patrons in the population generally increases with personal income. 33 percent of those with a personal income of between DKK 400,000-499,999 lend library materials, while 23 percent of the population with a personal income of less than DKK 25,000. Among persons with an income of DKK 500,000 or more, the share was 30 percent Among persons with an income of DKK 300,000-399,999, it was 25 percent, while loaners accounted for 19 percent of those with an income of DKK 200,000-299,999, 18 percent of those with an income of DKK 100,000-199,999 and 22 percent of those with a personal income of DKK 25,000-99,999.

The loaners that takeout materials from the physical libraries typically visit branches close to their place of residence. In 2021 31 percent of materials where taken out by loaners living within a radius of less than 1 km from the library. 47 percent of the material was taken out by loaners living at least 1 km, but less than 5 km from the library., while 11 percent went to loaners between 5 and 10 km away. Finally 8 percent of material went to loaner living more than 10 km away.



3.1 Library behavior during the COVID-19 crisis

As data is received every night and thus on a daily basis, it has been possible to follow the impact of the lockdowns during COVID-19 very closely and even down to the minut. There was a clear "hoarding effect" in the days leading up to the first Danish shutdown on March 12 and even greater effect at the second shutdown around December 8 2020. Danes rushed to the library to ensure that they would have books to read during the lockdown. The first lockdown came as a surprise to all, but when the second lockdown was announced it is clear to see how the number of material taken out rocket above 200.000 in a single day.



4. Expanding the statistics

The following section will cover the data needs of the world and whether data can be further developed or improved.

4.1 Access to researchers and ongoing projects

Due to the high quality of the microdata about library use in our new expanded statistic, data have been made available to researchers through Statistics Denmark Research Service in January 2022. The data has already been requested and used by several researchers, as library lending can be an indicator of learning levels at e.g. children or used to measure initiatives. A group of researchers have for example used library data to measure the effect of COVID-19 shutdowns on equality between children of affluent and less affluent families and whether the children lend digital materials from eReolen. The conclusion of this study was that families with lower education level stop using the libraries all together, while the more affluent families will substitute the physical books with digital versions. The study is listed in the references for further reading.

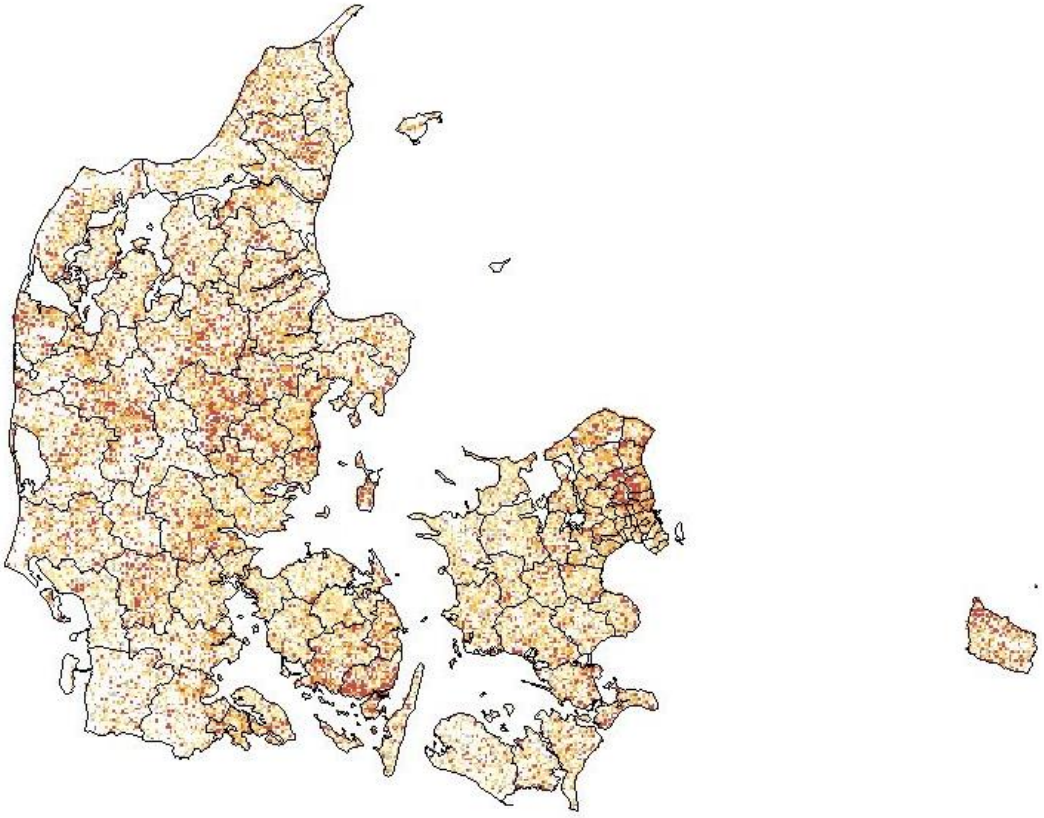
In addition, another group of researchers has used data to investigate whether children's interest in science subjects has increased after the children have seen a Christmas calendar focusing on the science subject. The future prospects for microdata involve, among other things the Statistics Denmark's science lab trying to aply machine learning algorithms to the detailed microdata about library loans.

4.2 The Danish Square network (Kvadratnet)

As library data can be linked to geographical locations, it is possible to visualize the borrowers' behavior using the Danish square network. The square network was established in 2002 and had to meet challenges with changing geographical divisions such as municipalities, regions and cities in order to be static over time. When using the square network, Denmark is divided into small squares which can be of different sizes such as 100x100 km, 10x10 km, 1x1km and all the way down to 100x100 meters.

By connecting data with the square network, you can better get a glimpse of how the lending behavior is dynamic within the individual municipality and between

municipalities. In addition, it is possible to connect the geographical location of the individual branches so that the effect of a library's presence or potential closure can be measured. The visualization below shows how large a proportion of the population in question in the square is a library user within the past year.



4.3 Can the same data model be used on other subjects?

The model developed here has is very simple, but yields powerful results. The principle that you can have a simple data structure where you in each row have a social security number and a time and place of an action performed that is then linked to demographical registers is a principle that can be applied in many different contexts. In the autumn of 2020, Statistics Denmark and the Danish Digitization Agency collaborated on developing a statistic to compile better socially illuminating statistics on digitization without putting a strain on citizens or enterprises. The setup was largely inspired by the approach and process behind the expanded library statistics. The collaboration is expected, for example, to give New Generation Digital

Post the opportunity to compile statistics on which types of users access or non-access their digital mail at certain times, for example by municipality. This type of statistic can create an opportunity to send digital mail more efficiently, communicate more purposefully to citizens and, overall, increase the likelihood of citizens reading mail, etc. It may also be possible to exhibit this type of statistics in an aggregated form to relevant senders, eg municipalities etc. It remains to be seen whether this project will take off.

5. References

Inequality in learning opportunities during Covid-19: Evidence from library takeout (2020) – Mads Meier Jæger/Ea Hoppe Blaabæk

Blaabæk, Ea. 2022. "*Stratification in Parents' Selection of Developmentally Appropriate Books for Children: Register-based Evidence from Danish Public Libraries".* European Societies (in press)