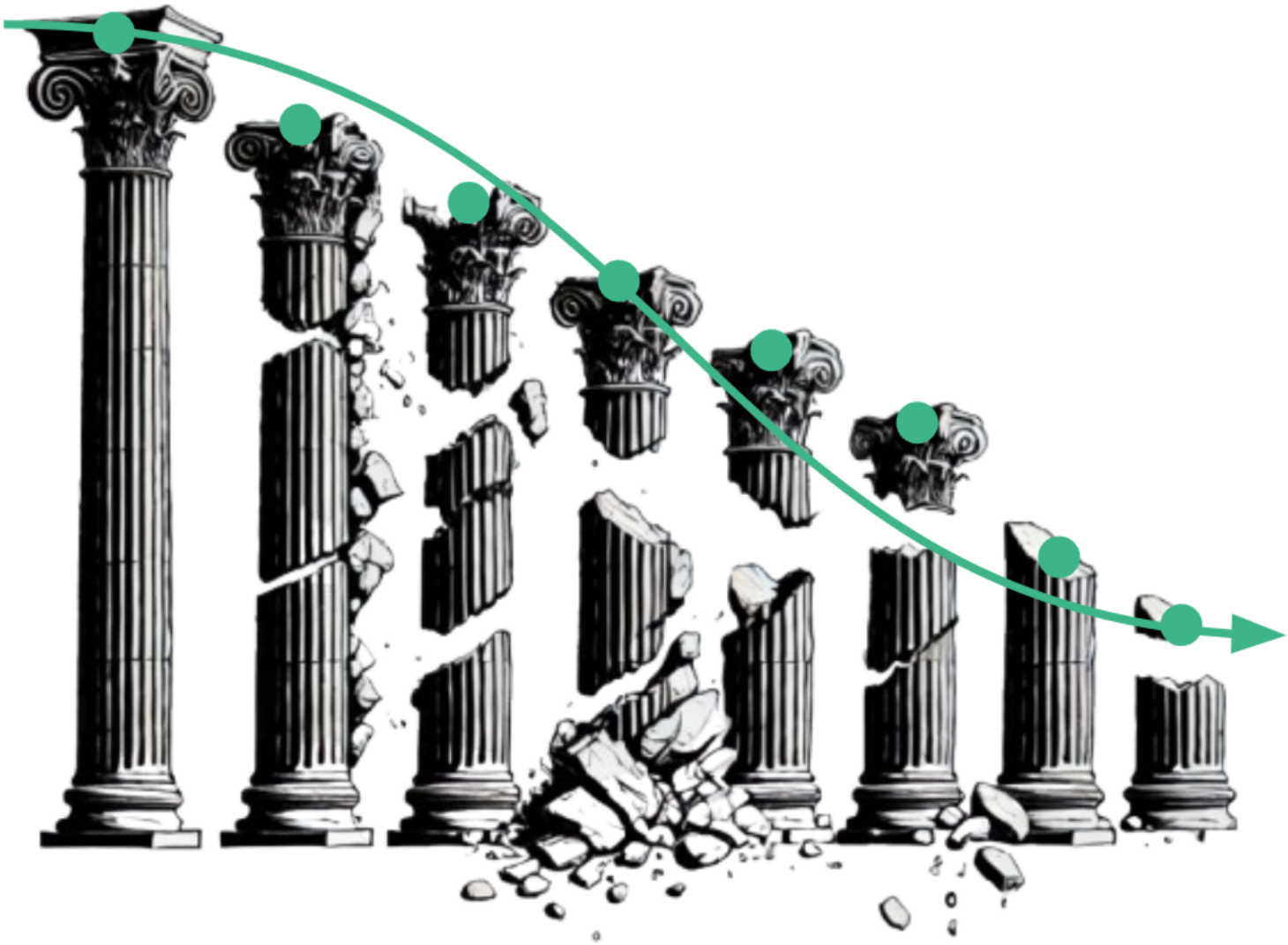
DTU Compute
Department of Applied Mathematics and Computer Science

The fit of the survival

A statistical analysis of building survival rates

Bachelor's Thesis



The fit of the survival
A statistical analysis of building survival rates

Bachelor's Thesis
December, 2024

By
Olivia Sommer Droob and Martha Falkesgaard Nybroe

Published by: DTU, Department of Applied Mathematics and Computer Science, Richard Petersens Plads, 324, 220, 2800 Kgs. Lyngby Denmark

Formalities

This bachelor's thesis was prepared at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in partial fulfillment of the requirements for acquiring a B.Sc degree in Mathematics and Technology.

Martha Falkesgaard Nybroe
s214692

Olivia Sommer Droob
s214696

Abstract

The new Life Cycle Assessment requirements in the Danish building sector call for more knowledge about the decay process and service life of Danish buildings. The analyses performed in this thesis provide the foundation for future research into data-driven service life estimation of Danish buildings.

We present a dataset of $153,775$ buildings demolished in Denmark from 2010 to 2024 and of $4,518,257$ buildings standing in 2024, with a variety of available variables for each building. We present a comprehensive overview of the most important variables, focusing on service life, construction year, building use-category, materials used and the interactions between these. We find that the average service life of a building demolished in the observational period is 71.81 years for all use-categories and 96.72 years when considering only the residential buildings.

We model the service life as probabilistic distributions and find that phase-type distributions outperform classical statistical distributions. By considering the service lives as realizations of an underlying Markov chain, we show different decay scenarios of buildings with real-life interpretations in terms of construction, reconstruction and demolition.

We demonstrate the usefulness of the Turnbull algorithm in estimating survival curves of buildings in a censored data setting. As our dataset does not contain left-censored observations, we show methods for creating left-censored data using simulation and backcasting. For residential buildings we estimate the probability of surviving to 50 years as ranging from 91.6% to 93.8%, depending on methodology.

Acknowledgements

We would like to thank our outstanding team of supervisors; Nicolai Siim Larsen, Bo Friis Nielsen, Carsten Rode and Niels-Jørgen Aagaard.

We want to thank our main supervisor Assistant Professor Nicolai Siim Larsen, who has contributed to this project in all its phases in all types of ways. The project would not have been possible without his knowledge, his thoroughness, and his belief in us.

We thank Professor Bo Friis Nielsen for his expertise on stochastic processes and his skillful sharing of it. Also, for always bringing the USB-to-HDMI converter to our meetings.

For building a bridge between our equations and the real-world outside of them, we owe a big thanks to Professor Carsten Rode and Associate Professor Niels-Jørgen Aagaard.

Thanks to Assistant Professor Rune Andersen, for granting us access to his data on Danish building demolition.

Lastly, thanks to our friends and family for their support through the project and for patiently listening to our tirades about building data.

Notation

$\mathbb{I}(\cdot)$	Indicator of event
$\mathbf{0}_p$	Row vector of p zeros
$\mathbf{1}_p$	Column vector of p ones
τ	The time until absorption in an absorbing Markov jump process
PH_p	Phase-type model with p phases
$\mathbf{T}$	Sub-intensity matrix in a phase-type distribution
$\mathbf{\Lambda}$	Intensity matrix in a phase-type distribution
$\boldsymbol{\pi}$	Initial probability vector in a phase-type distribution
$\mathcal{L}$	The age of a building when left-censoring occurs
$\mathcal{X}$	The service life of a building
$\mathcal{R}$	The age of a building when right-censoring occurs

Contents

1 Introduction

Estimates from the Danish Authority of Housing indicate that construction of buildings accounts for 10 percent of Denmark's total consumption-based CO_2-equivalent (CO_2e) emissions. From July 2025 a law imposing stricter regulations on the Danish building sector will enforce upper bounds on emissions for all new buildings. The bounds will vary based on the use of the building and will be gradually tightened in the following years. The emission of a new building is based on a mandatory Life Cycle Assessment using a standardized 50-year observation period. While this observation period aligns with international standards and ensure comparability, it may disadvantage durable building materials and long-lasting construction practices, as these often have a higher initial CO_2e imprint. Stakeholders from the building sector have argued that accounting for differences in service life between building uses and materials in the Life Cycle Assessment would encourage contractors and construction firms to prioritize durable, long-term solutions. While this argument has merit, it also exposes a new challenge: there is limited data on the subject, few long-term studies, and no standardized methodology for estimating building service lives.

The Danish Building and Housing Register (BBR) presents a unique opportunity to address this knowledge gap. As a well-documented and centralized national database maintained since 1976, it offers valuable data for service life estimations. This thesis examines the possibilities and limitations of using the BBR dataset for mathematical modeling of building service lives. It provides an analysis of the data, highlighting key insights into the Danish building stock as of 2024 and of buildings demolished between 2010 and 2024. Additionally, it identifies limitations within the BBR data and related datasets from Statistics Denmark. Multiple modeling angles are tested on the data. One approach uses phase-type distributions to model building service life as a probabilistic distribution. This leads to an interesting interpretation of building decay as an absorbing Markov jump process. Another approach examines the building survival curve, where the BBR dataset is supplemented with synthetic datasets - one derived from studying yearly demolition rates and one derived using the phase-type model. Algorithms for estimating survival curves in settings with censored data are tested on these datasets. Together, these methodologies provide a data-driven understanding of Danish building service lives.

2 Literature review

To gain an understanding of the academic field of research within service life analysis of buildings, we reviewed the existing literature to answer the questions:

1. What are the current regulations on climate impact in the Danish building sector and what are the implications of these regulations?
2. Which methods for service life estimation have been explored in academic literature?

Current service life regulations in the Danish building sector

It is required to do a Life Cycle Assessment (LCA) for every newly constructed building in Denmark, assessing the climate impact of the building over its whole life cycle [12]. From July 2025 new upper bounds on yearly climate impact will be enforced based on the LCA. The bounds vary based on the use of the buildings, but will be based on an observation period of 50 years for all buildings, no matter the building use-category and construction materials.

This uniform length of the observation period impacts which kinds of buildings are considered most sustainable. It provides little incentive to use materials designed to last beyond this timeframe, thereby failing to encourage sustainable constructions. Longer-lasting materials such as concrete have a higher initial CO_2e emission, but may outperform lighter alternatives in durability long-term. This is not incorporated into the current governmental service life standards. Furthermore, enforcing a uniform service life requirement on all buildings may lead to an inefficient resource use when constructing buildings where a 50-year service life is excessive. Accurate service life predictions can support decision-making in the building industry, informing both policy and practice towards lower environmental impacts.

Research on service life estimation

The CO_2e emissions of a building are evaluated over a designated observation period, the length of which is still a debated topic. Various sources have attempted to estimate building service lives, to use as the observation period. No definitive consensus has been reached due to the often unreliable methods and the many factors influencing the estimation.

Much of the building sector still relies on The Factor Method for service life estimation as presented in ISO 15686 in 2011. This method gained widespread usage as a simple tool for service life prediction of building materials and components [39]. The method estimates the service life of a particular component or building under specific conditions. It is based on a reference service life and a series of modifying factors that relate to the specific conditions of the case: quality of components, design level, work execution level, indoor environment, outdoor environment, in-use conditions and maintenance level. While the Factor Method remains widely used, researchers have explored multiple enhancements to make it more flexible and accurate. One simple improvement is a stochastic version, which incorporates minimum, maximum, and expected values for each factor to account for intra-building variability [28].

Others have suggested moving away from the Factor Method entirely. Inspired by the successful implementation of Markov chain modelling for highway degredation in (Tee, Ekpiwhre, and Yi, 2018), an approach estimating the transition probabilities of buildings's decay has been proposed by (Duling and Jacobus, 2007) and (Johann et al., 2008). Both studies found themselves to be restricted by the very limited amounts of historic performance data, regarding degradation of building materials.

Another approach to service life estimation has been to consider the entire Danish building stock

through an economic lense. One commonly employed method involves assuming a constant yearly demolition rate. For example, the Annual Danish Aggregate Model (ADAM), a model for the Danish economy used by the Ministry of Finance assumes a 1% demolition rate. Under a geometric distribution, this yields a mean service life of 100 years, but this figure has been subject to scrutiny. Instead, (Aagaard et al., 2013) proposes a more true-to-life method using a demolition-free "loan period" of 20 or 30 years followed by a demolition rate of 0.3%. Under a geometric distribution, this would yield a service life of 353 to 363 years, much higher than the assumption used by the Ministry of Finance. However, none of these estimates are based on data estimates. An estimate by (Andersen, 1992) uses demolition data on Danish houses 1980-1989 to estimate the yearly demolition rate and finds an even lower value at 0.15%. This rate corresponds to a geometric distribution with expected service life 667 years. In a working paper by The Knowledge Center for Housing Economics [47] an overview of different estimation methods is presented with average service life estimations ranging from 34 to 300 years and medians ranging from 55 years to 289 years. This clear inconsistency demonstrates the challenges in estimating the service life of Danish buildings.

Steps towards a data-driven approach have been taken in recent years. Estimations made by performing multiple linear regression on demolition data obtained from BBR find the mean service life for all buildings to be 78 years [34]. In (Andersen and Negendahl, 2023) it is concluded, based on BBR data from 2010 to 2017, that service life is decreasing as a function of construction year, such that newer buildings have a shorter lifespan than the average lifespan. As noted by (Haugbølle et al., 2021) and (Jensen et al., 2024) these studies have significant challenges regarding selection bias as the demolition data is limited to a specific period and conclusions are generalized without considering data censoring and representability.

Previous research has implemented censorship mechanisms [40], where a survival curve estimation algorithm is employed with 12.4% left-censored and 87.6% right-censored data. The data foundation is rather limited, consisting of visual assessments of only 2960 buildings over 3 years. In a different article only right-censored data is included in the estimate of the survival curve while left-censoring is not considered [14]. The authors obtain an unrealistically high survival rate, which isn't surprising considering the one-way approach to censoring.

In the South-Korean study (Ji, Lee, and Yi, 2021) a big-data approach is used through deep neural networks and Extreme Gradient Boosting models, finding high coefficients of determination in prediction of service lives. The authors note that the dataset, containing demolition cases from the 1950s till today (though only 1% of the data is from before 2000), is characterized by "discrepancy in the data quality and bias". Their estimated service lives are surprisingly short - ranging from 16 to 32 years by region. While this approach is interesting, these estimations are hardly transferable to a Danish setting.

Generally, the use of data-driven models is fairly underexplored in this field, due to the lack of good data - and because the topic is often studied through a civil engineering lens more so than a data science lens.

3 Methodology

We wish to estimate the service life of Danish buildings. To do so, we will obtain data on Danish buildings standing in 2024 as well as on buildings demolished between 2010 and 2024. For buildings demolished before 2010 and after 2024, we do not know the exact service lives. Data for these buildings is called censored data. Later in this thesis, we will describe the process of acquiring and cleaning data. We are able to acquire uncensored data as well as right-censored data. The problem of the missing left-censored data is discussed in this and future sections.

But first we establish a framework for understanding and handling censored data. In the following section we will introduce notation to describe different kinds of censored data and the information we can extract from it. We will use this notation to explain the problem of bias and how this is handled in our thesis. We will denote random variables with capital cursive letters such as $\mathcal{X}$ and $\mathcal{C}_i$. Realizations of such variables will be denoted with lower case letters.

3.1 Censored data

Let $\mathcal{X}_i$ be a random variable representing the service life of building i. The service lives $\mathcal{X}_i$ are assumed to be independent and identically distributed with probability density function $f(x)$. $\mathcal{X}$ only takes integer values in our data, but the values range from 1 to 563, so $\mathcal{X}$ can be treated as approximately continuous. The service life of building i is determined by

$$\mathcal{X}_i = \mathcal{D}_i - \mathcal{C}_i, \tag{3.1}$$

where $\mathcal{C}_i$ is a continuous random variable denoting the year of construction of building i with a probability density function $g(c)$. $\mathcal{D}_i$ is a random variable denoting the demolition year of building i and it holds for all buildings that $\mathcal{D}_i > \mathcal{C}_i$. We assume that the service life $\mathcal{X}_i$ is independent of the construction year. In our setting, the service life $\mathcal{X}_i$ will become known if the building is demolished within an observational period from 2010 to 2024, that is if:

$$2010 \leq \mathcal{D}_i \leq 2024. \tag{3.2}$$

If $\mathcal{D}_i > 2024$, the observation is considered right-censored and its demolition time is censored at 2024. For such buildings, it is known that the demolition happens after 2024, but the exact time is unknown. The age of the building when right-censoring occurs is called $\mathcal{R}_i$:

$$\mathcal{R}_i = 2024 - \mathcal{C}_i.$$

If $\mathcal{D}_i < 2010$, the observation is considered left-censored and its demolition time is censored at 2010. For such buildings, it is known that they were demolished before 2010, but the exact time is unknown. The age of the building when left-censoring occurs is called $\mathcal{L}_i$:

$$\mathcal{L}_i = 2010 - \mathcal{C}_i.$$

The information from one observation can be represented by a pair of random variables $(\mathcal{T}_i, \delta_i)$. Here δ_i indicates whether the service life $\mathcal{X}_i$ is uncensored ($\delta_i = 1$), is right-censored ($\delta_i = 0$) or is left-censored ($\delta_i = -1$):

$$\delta_i = \begin{cases} -1 & \mathcal{D}_i < 2010 \text{ (Left-censored)} \\ 1 & 2010 \leq \mathcal{D}_i \leq 2024 \text{ (Uncensored)} \\ 0 & \mathcal{D}_i > 2024 \text{ (Right-censored)} \end{cases}$$

$\mathcal{T}_i$ is a random variable equal to $\mathcal{X}_i$ if the lifetime is observed and equal to one of the censoring limits if not:

$$\mathcal{T}_i = \begin{cases} \mathcal{L}_i & \mathcal{D}_i < 2010 \text{ (Left-censored)} \\ \mathcal{X}_i & 2010 \leq \mathcal{D}_i \leq 2024 \text{ (Uncensored)} \\ \mathcal{R}_i & \mathcal{D}_i > 2024 \text{ (Right-censored)} \end{cases}$$

The variable pair has three scenarios - left-censored observations $(\mathcal{L}_i, -1)$, uncensored $(\mathcal{X}_i, 1)$ and right-censored $(\mathcal{R}_i, 0)$. We have thereby established a framework for describing censored data as well as a convenient notation which we will rely on in the following sections.

3.2 Bias in data

Ideally we would have a complete dataset of all service lives of all Danish buildings ever to exist. This would allow us to analyze demolition trends and service life distributions through time without any time bias.

The data that is currently available is the standing building stock and uncensored data from 2010 - 2024 providing us with exact service lives in the observation period. We note that the uncensored data only accounts for tendencies seen in demolition from 2010 to 2024 and is therefore not necessarily representative for demolition before and after this period. Furthermore, analyzing the uncensored data introduces survivorship bias which will be described in the following section.

3.2.1 Survivorship bias

The available data with an exact service life is based on a retrospective enrollment, meaning that the buildings are observed after the event of demolition has occurred. This type of data introduces the issue of survivorship bias [1], which is a specific form of selection bias. It stems from only enrolling the survivors in the experiment and thereby skewing the conclusions toward the survivors.

The buildings we can observe are only those that have survived up until 2010, automatically excluding all buildings demolished before that time. This skews the dataset in favor of newer buildings as the probability of being demolished before 2010 is higher for buildings built a longer time ago. To show this, the theoretical probability of being left-censored is derived:

$$\begin{aligned} P(\delta_i = -1) &= P(\mathcal{D}_i < 2010) \\ &= P(\mathcal{X}_i + \mathcal{C}_i < 2010) \\ &= P(\mathcal{X}_i < 2010 - \mathcal{C}_i) \\ &= F(2010 - \mathcal{C}_i). \end{aligned}$$

where F is the cumulative distribution function of $\mathcal{X}_i$. As $P(\delta_i = -1)$ is a decreasing function of construction year $\mathcal{C}_i$, it results in a higher theoretical probability of being left-censored for buildings built a long time ago. The probability of being right-censored can be derived in the exact same manner, where the conclusion is that $P(\delta_i = 0)$ is an increasing function of construction year C_i. The probability of a building being demolished in our observational period, $P(\delta_i = 1)$, is dependent on the distribution of $\mathcal{X}$:

$$\begin{aligned} P(\delta_i = 1) &= P(2010 \leq \mathcal{D}_i \leq 2024) \\ &= P(2010 \leq \mathcal{X}_i + \mathcal{C}_i \leq 2024) \\ &= P(2010 - \mathcal{C}_i \leq \mathcal{X}_i \leq 2024 - \mathcal{C}_i) \\ &= F(2024 - \mathcal{C}_i) - F(2010 - \mathcal{C}_i), \end{aligned}$$

The first-order derivative of this probability is

$$P'(\delta_i = 1) = f(2010 - \mathcal{C}_i) - f(2024 - \mathcal{C}_i),$$

where f is the probability density function of $\mathcal{X}_i$. For continuous distributions with a single maximum, the probability $P(\delta_i = 1)$ will be increasing as a function of $\mathcal{C}_i$ up until $f(2010 - \mathcal{C}_i) = f(2024 - \mathcal{C}_i)$ and then decrease. The location of this peak depends on the distribution. For our data, this means that some construction years are more likely to be present in the uncensored data, but that the earliest and latest construction years are less likely to be present.

For our observational period, the buildings from the earliest construction years, are those that have survived for a long time already. The observed service lives will then be unrealistically high of the buildings with the oldest construction years, and unrealistically low for the newest constructed buildings. This makes it impossible to use our data to analyze the service life development through time. To avoid making biased conclusions, we will operate under the assumption that the service life of a building is independent of its construction year and refrain from making any claims to the contrary. This assumption is strong, but necessary. It alleviates the concern about survivorship bias and also means that our observational period can be considered more representative.

3.2.2 Non-uniformity of construction year distribution

Our data is colored by the economic cycles of the Danish building sector, where heavy construction activity in the 1960s and 1970s means that a large number of buildings from these years are present in the data. Even when assuming stationarity in service life over time, we might find that buildings built in the 1960s and 1970s make up a bigger proportion of demolition cases, skewing conclusions towards service lives of 40-64 years. The conclusions about mean service life and the fitted models are thus descriptive more than predictive, as they have this overrepresentation.

When concluding on the uncensored data, the conclusions hold exactly for the period 2010-2024. The biases discussed here might affect how the conclusions can be extrapolated to the rest of the Danish building stock. When estimating models on uncensored data, we will keep our conclusions close to the data, while our models incorporating censored data can provide broader conclusions outside of the observational period.

4 Data

In the following section we will describe our data. We begin by describing the acquisition and cleaning process and our considerations through this process. We will then present the variables in the resulting dataset, using descriptive statistics and data visualization to gain an understanding of the variables and the interactions between them. We conclude this section with a discussion of the data quality and a comparison between this dataset and data available through Statistics Denmark.

4.1 Data acquisition and cleaning

Our dataset is based on the Danish Building and Housing Register (BBR), a registry under the authority of the Danish Ministry of Taxation. The registry contains information about Danish buildings and was established in 1976 with the purpose of providing information for property valuation, planning, and statistics. The registry is semi-openly available but requires extensive formatting before being useful for our purposes. Here, we describe our data acquisition and cleaning process.

Through the web service "Datafordeler" [26] we were able to download the dataset "Bygninger" as a JSON file, containing registrations on buildings from 2017 to 2024. The reason for registration in BBR is described in the column `Forretningsproces` [48][ENG: `Business Process`], where the code 3 corresponds to the building having been demolished. From this, we added the binary variable `Demolished` to our dataset, where 1 corresponds to the building having been demolished.

Each observation then contains information on the construction year, materials used, area, use, heating, location, demolition date if relevant, and a unique ID. As the BBR is updated each time something is changed in a building, eg. a reconstruction or change of heating method, a lot of duplicates are present. We can identify these based on their building ID. Buildings without a construction year are discarded at loading and are thus not part of the analysis.

The time of an update in the BBR is registered in the variables `virkningFra` [ENG: `Effect From`] and `registreringFra` [ENG: `registration from`]. Here `virkningFra` is the actual time of demolition, while `registreringFra` is the time of registration. For duplicated demolished buildings we kept the building with the earliest `virkningFra`, and for standing buildings we kept the one with the newest `registreringFra`. This variable `virkningFra` contains date, month and year, but `construction year` only contains the year of construction. To keep the formats uniform, we only keep the year of demolition. We compute the variable `Service life` as the difference between demolition and construction year.

We removed bureaucratic variables such as `registreringsaktør`[ENG: `Registration authority`]. For the variables `stormskadet` [ENG: `Storm Damage`], `fredet` [ENG: `Listed`], and `omtilbygningsår` [ENG:`Reconstruction Year`], most buildings do not have any information, which we assume is because the building isn't affected by storm, listed or reconstructed, not necessarily due to missing information.

Additionally we removed 57 variables that were missing more than half of their values. These included information about drainage, garbage conditions, and specific area measures of a building, but none of importance for our particular purpose. In the dataset there are two variables for building area; `bebyggetAreal` [ENG: `Covered area`] and `samletBygningsareal`[ENG: `Total Covered Area`]. As they are highly correlated we choose to delete the one with the most missing values. The Ministry of Taxation considers any construction year before 1400 to be an error [37]. Following this guideline, we remove all buildings with a construction year before 1400.

The variable `BygningensAnvendelse`[ENG:`Building use`] contains 101 subcategories of building use. These are mapped into the 7 building use-categories described by (Aagaard et al., 2013). The categories are "Housing", "Agriculture", "Production", "Transport and Commerce" (abbreviated T & C), "Institution", "Leisure" and "Others". As there is no mapping of small sheds, garages etc. these have been put in the category "Others". We keep the original `BygningensAnvendelse` and will refer to these as use-subcategories.

For each building we have a set of coordinates. The coordinates are given in UTM-WGS84 coordinates, which project the Universal Transverse Mercator(UTM) onto the World Geodetic System (WGS)[46]. A transformer using `Proj` from the python package `pyproj` [16] was used to map these coordinates into longitude and latitude coordinates. Using the open-source geocoding software Nominatim [33] (sourcing from OpenStreetMap data) we can map the latitude and longitude to an address. This mapping is carried out for every building in our dataset, resulting in the variable `Address`. The address can lead to more fine grained geographical analysis to study geographical differences in service lives across Denmark. It can also be used to investigate single cases by looking at them on eg. Google Maps.

The described data cleaning left us with the 18 variables presented below:

Variable	**Description**	**Type**
`ID`	Unique ID for each building	UUID
Demolished	Indicator for whether the building was demolished 2010 - 2024	binary
`Demolition year`	Year the building was demolished (if relevant)	integer
`Construction Year`	Year the building was constructed	integer
`Age`	Service life of demolished building or age of standing building	integer
`Area`	Building area	float (m^2)
`Use-subcategory`	Use of the building	categorical
`Use-category`	Use of the building (SBI 2013:30)	categorical
`Location coordinates`	Coordinates for the location of the building	UTM (WGS84)
`Address`	Address of the building	string
`Outer wall material`	Material used for the outer walls	categorical
`Roof material`	Material used for the roof	categorical
`Municipality`	Municipality where the building is located	categorical
`Listed building`	Whether the building is a listed building	binary
`Heating`	Type of heating	categorical
`Asbestos`	Whether the building has asbestos	binary
`Storm damage`	Whether the building has storm damage	binary
`Reconstruction year`	Year building was reconstructed or "Not reconstructed"	integer/string

We have $74,819$ observations of buildings demolished between 2017 and 2024 and $4,518,257$ buildings standing in 2024.

Comparison and merging with supplementary dataset

During our literature review, we came across another dataset of demolished buildings obtained from the BBR on $104,927$ buildings. The dataset is created by KMD and the data acquisition process is described in [6]. We found it relevant to combine the two datasets, which we were granted permission to do from Rune Andersen, Assistant Professor at the Department of Environmental and Resource Engineering at the Technical University of Denmark.

Andersen's dataset contains demolitions from 2010 up until 2020, which results in a 4-year overlap with our dataset from 2017-2024. 25,706 buildings are present in both datasets. When comparing the IDs present in both datasets 17.0% differ in demolition year. The differences in demolition year between our dataset and Andersen's dataset are presented in the table below:

Difference (years)	-2	-1	1	2	3	4	5	6	7	8	9	10	13
Buildings	3	8	2952	662	267	152	128	78	38	29	33	7	2

Table 4.1: Differences between the demolition year of a building as registered in our dataset and in Andersen's dataset

In most cases, the demolition year in our dataset is greater than for Andersen's dataset. Most of the differences are one-year differences (67.9%), and the mean absolute difference is 1.76. These differences are small in scale compared to the service lives, so we are not particularly concerned with this.

To obtain the most data for our future research we merge our dataset with Andersen' dataset, prioritizing Andersen's for buildings present in both datasets. This leaves us with a dataset containing 153,775 buildings demolished from 2010 to 2024 as well as all 4,518,257 buildings still standing in Denmark in June 2024. From now on in the thesis we will refer to this dataset as the BBR dataset.

4.2 Data description

This section presents an overview of the most important variables of the BBR dataset, considering in particular the service life, building use-categories and construction materials. We consider the buildings demolished 2010-2024 and buildings standing in 2024 separately, but will compare and contrast between the two groups.

4.2.1 Service life

We begin by examining the distribution of the service life $\mathcal{X}$ for the uncensored population and the right-censoring age $\mathcal{R}$ for the right-censored population. The following summary statistics describe the variables:

	Buildings demolished 2010-2024	**Buildings standing 2024**
Count	153,612	4,502,680
Mean	71.81	54.59
Std	44.70	41.86
$\mathbf{Q_{0.25}}$	41	24
Median	60	47
$\mathbf{Q_{0.75}}$	97	70
Max	563	624

Table 4.2: Summary statistics of $\mathcal{X}$ and $\mathcal{R}$

The complete distributions are shown in Figure 4.1. The histogram to the left shows the uncensored service life for all buildings demolished 2010-2024, that is $P(\mathcal{X} \in \text{10-year bin } |\delta = 1)$. The histogram to the right shows the right-censoring age for all buildings standing in 2024 expressed as $P(\mathcal{R} \in \text{10-year bin}|\delta = 0)$:

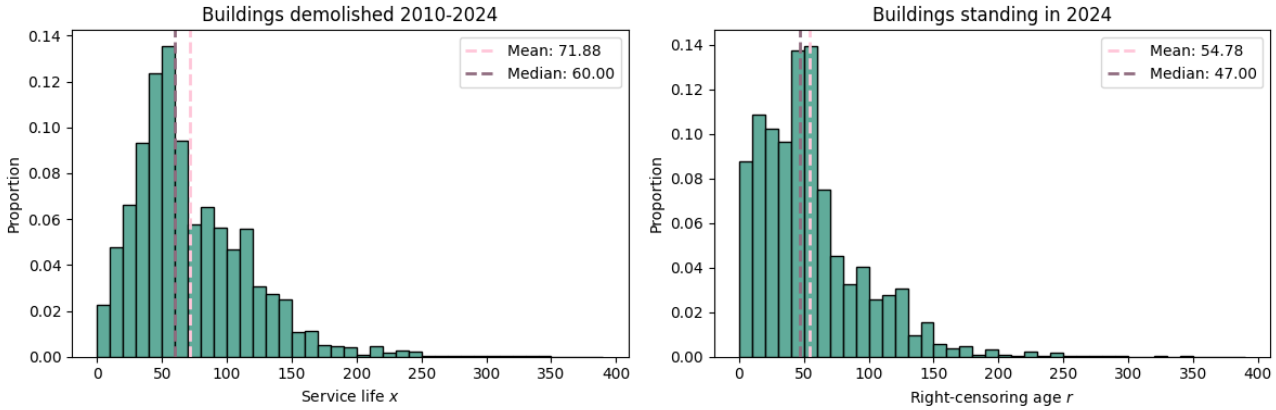


Figure 4.1: Probability distribution of $\mathcal{X}$ for buildings demolished in the observational period (shown to the left, $\delta = 1$) and $\mathcal{R}$ for standing buildings (shown to the right, $\delta = 0$) with vertical lines marking the mean and median values.

For the buildings demolished in the observational period ($\delta = 1$) the mean of $\mathcal{X}$ is $\overline{x} = 71.81$ years and the median is 60 years. The distribution of $\mathcal{X}$ is right-skewed, with a thin tail to the right. This is natural for positive, bounded data and leads to the median service life 11.8 years less than the mean, as a few very old buildings inflates the mean. The distribution has a peak for 50 to 60 years, corresponding to the building boom of the 1960s and 70s. We note that the industry standard of a 50 year service life is at the 35^{th} percentile, well below the observed mean and median service lives. The maximal service life $\max(\mathcal{X})$ is 563 years, much higher than the third quartile at 97 years. This is indicative of a long tail.

For the buildings still standing in 2024 ($\delta = 0$), they are in general younger, with mean age $\overline{r} = 54.78$ and a median of 47 years. We see a big peak in the number of buildings with ages 40 to 70, corresponding to the building boom of the 1960s and 1970s.

It makes sense that the demolished buildings are older, as older buildings are more likely to be demolished. This holds in general until 370 years, where the age itself becomes a reason for preservation. This means that the proportion of buildings aged 370 and older is bigger for the standing than the demolished buildings.

4.2.2 Building use-categories

In the following we will investigate the use-category of buildings, considering the distribution of use-categories amongst the standing and demolished populations as well as how the different use-categories affect the service life distribution.

We examine the distribution of building use-categories conditioned on δ, that is $P(\text{use-category} = j|\delta = 1)$ and $P(\text{use-category} = j|\delta = 0)$ where $j \in$ use-categories. These are presented in Figure 4.2:

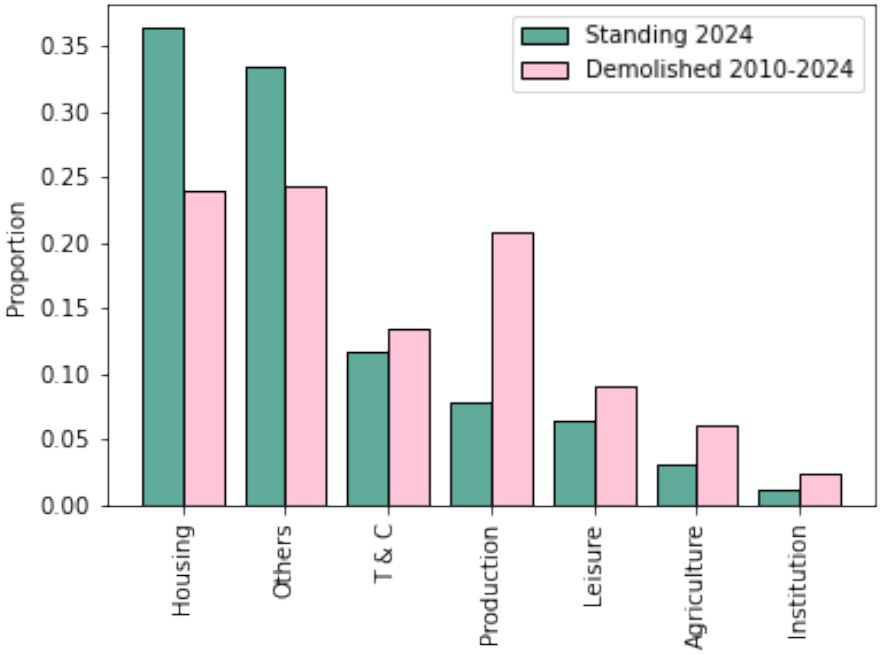


Figure 4.2: Proportions of use-categories among buildings standing in 2024 ($\delta = 0$) and buildings demolished between 2010 and 2024 ($\delta = 1$)

The distribution of use-categories for the standing and demolished buildings show different tendencies. A χ^2-test yields a p-value smaller than 10^{-16}, which rejects the null hypothesis of equal distributions between the right- and uncensored buildings. The most common use-category for the demolished buildings are "Housing" and "Others". A similar pattern is observed for standing buildings, but the proportions of these two use-categories are even higher, accounting for two thirds of all standing buildings. Conversely, the proportion of "Production" buildings is greater for the demolished buildings. This does not necessarily imply that a higher proportion of production buildings is demolished relative to other use-categories. Instead, this could reflect which building use-categories were predominantly constructed 50-70 years ago, which aligns with the average service life. To explore this, we analyze the distribution of building use-categories across construction periods.

Figure 4.3 shows the proportion of use-categories which were built in a specific construction period. To the left, the data is the buildings demolished between 2010 and 2024, corresponding to $P(\text{use-category} = j | \delta = 1, \mathcal{C} \in \text{10-year bin})$, where $j \in$ use-categories and $\mathcal{C}$ denotes the construction year. To the right, the data is buildings standing in 2024, corresponding to $P(\text{use-category} = j | \delta = 0, \mathcal{C} \in \text{10-year bin})$.

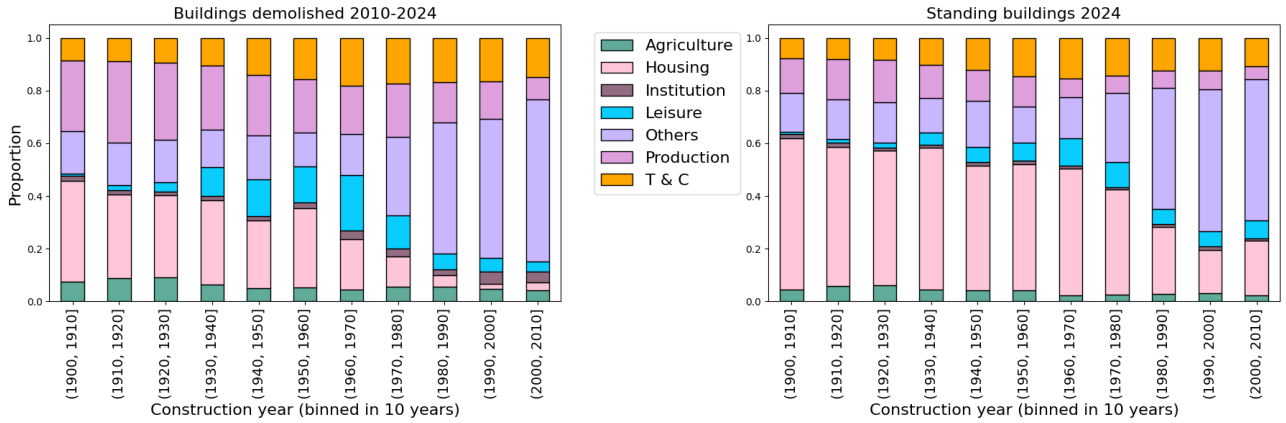


Figure 4.3: The proportion of each building use-category based on 10-year bins of construction year for buildings demolished 2010-2024 and buildings standing 2024

The proportion of "Other" is greater for the youngest buildings in both the demolished and standing population. No data exists on the building use-category distribution at the time of construction. So the large proportion of "Others" in the most recently constructed buildings could be explained by most buildings in this sub-category built before 2010 also having been demolished before 2010, which would make them absent from our dataset.

The proportion of housing is greater for standing buildings, but decreases for both standing and demolished buildings as the construction year increases. This may be due to houses accounting for a smaller proportion of buildings being built today. It can also be that the absolute number of houses built is the same as in previous decades, but that the proportion is smaller, due to the amount of buildings with use-category "Others" in the most recent years. It could also be that each house today is bigger, so fewer housing units are built. As we do not have historic data on the building use-category or area of constructed buildings, it remains for us to speculate.

As we expect the use-category of a building to significantly impact its decay process, we will now investigate the observed distributions of $\mathcal{X}$ given the different use-categories, corresponding to $P(\mathcal{X} \in \text{10-year bin} | \delta = 1, \text{use-category} = \text{j})$, for $j \in$ use-categories:

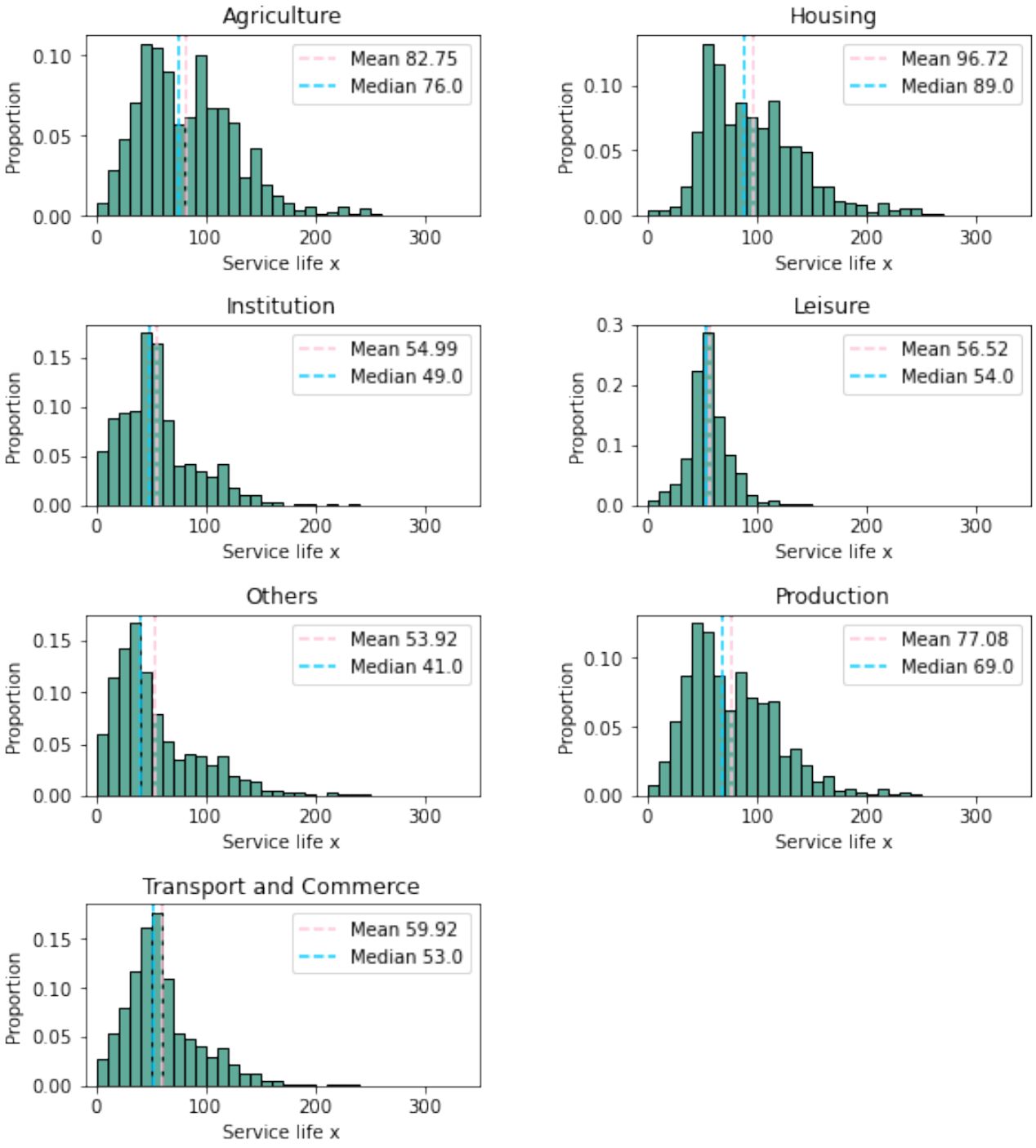


Figure 4.4: Histograms of uncensored service life $\mathcal{X}$ for each use-category of buildings demolished between 2010 and 2024. The vertical lines represent the mean and median value of $\mathcal{X}$ for each use-category.

We denote the service lives of all agricultural buildings demolished 2010-2024 as making up the distribution of $\mathcal{X}_{\text{Agriculture}}$ and correspondingly for the other use-categories.

The distributions of $\mathcal{X}_{\text{Agriculture}}$ and $\mathcal{X}_{\text{Production}}$ seem almost bimodal. These use-categories might cover a heterogeneous subset of buildings with different decay processes. As each of the 7 building

use-categories consists of a range of use-subcategories, we investigate how the service life looks within each building use-category and their use-subcategories. This investigation is found in appendix A.3. General descriptions and key findings for each of the use-category distributions of $\mathcal{X}$ are presented below:

Agriculture: $\mathcal{X}_{\text{Agriculture}}$ shows a clear bimodal distribution that peaks at 40-50 years and 90-100 years of service life. From investigating the use-subcategories we find that the bimodal distribution is consistent through the use-subcategories and thereby a general tendency of the distribution of $\mathcal{X}_{\text{Agriculture}}$. This could be a sign that a mixture distribution is needed to fit this data.

Housing: $\mathcal{X}_{\text{Housing}}$ has quite a widespread and flat right-skewed distribution. It has a big peak around 60-70 years, and then another small peak at 120 years. This could indicate a mixture distribution. From investigating the use-subcategories of housing we find that the category "farmhouse for agricultural property" is mainly present in the second peak of the distribution, while "single family houses" fills out almost the rest. We are not too concerned with this particular case, as farmhouses make up a smaller proportion of the housing use-category, and the other use-subcategories follow the same distribution as the grouped distribution.

Institution: The distribution of $\mathcal{X}_{\text{Institution}}$ peaks around 50 years of service life. It exhibits some bimodal tendencies, which is either a characteristic of $\mathcal{X}_{\text{Institution}}$ or the result of a mixture distribution of the use-subcategories. From the investigation of the use-subcategories we find that the bimodal distribution tendency is not due to the use-subcategories but rather something that is a general tendency in this use-category. The use-category encompasses a range of public buildings such as schools, kindergardens, museums, theaters and hospitals. The bimodal tendency peak at ages corresponding roughly to periods of Danish economic cycles with expansive financial politics, resulting in heightened building activity in public spaces (for example the 1960s). It could also be that each use-subcategory contains both longer lasting and shorter lasting kinds of buildings.

Leisure: The distribution of $\mathcal{X}_{\text{Leisure}}$ has a very sharp peak at 50-60 years of service life, with a quite sudden drop on both sides. The building use-category "Leisure" is mostly (86%) summer houses, a category that only became mainstream 60 to 70 years ago, explaining why only a few older buildings are found in this use-category. It is still somewhat right skewed with a slight tail, but all use-subcategories follow the same distribution.

Others: The distribution of $\mathcal{X}_{\text{Others}}$ has a peak at 30-40 years of service life. It exhibits tendencies of a mixture distribution. We find that the use-subcategory "garages" is responsible for most of the peak. The concept of a garage only became popular after the increasing prevalence of cars in the 1950s. This limits the majority of the service lives of this use-subcategory to 70 years. The use-subcategory "sheds" generally has a more flat structure, although still peaking around the same time as "garages", just more heavy-tailed. The grouping of these two use-subcategories could explain some of the general structure of this service life distribution.

Production: The distribution of $\mathcal{X}_{\text{Production}}$ has a big peak around 50 years and another slightly flatter peak around 100 years, once again resulting in a bimodal distribution. From looking at the use-subcategories we find that the buildings used for the production of agriculture account for most of the general distribution while the rest of the categories mainly add to the spike around 50 years of service life. This means that the grouping of these use-subcategories could explain some of the bimodal structures of this service life distribution.

Transport and Commerce: The distribution of $\mathcal{X}_{\text{Transport and Commerce}}$ has a distinct peak at 50-60 years of service life and then a sharp drop on the right side. The distribution is slightly heavy-tailed. Some of the tail behavior could be due to a mixture distribution with the sharp peak as the first distribution and a wide symmetric distribution as the other.

A general statistical overview of the building use-categories' observed service life $\mathcal{X}$ is presented in Table 4.3:

Building use	**Agriculture**	**Housing**	**Institution**	**Leisure**	**Others**	**Production**	**T & C**
Count	9239	36907	3622	13889	37472	31971	20675
Mean	82.75	96.72	54.99	56.52	53.92	77.08	59.92
Std	45.07	46.94	36.05	23.40	42.34	42.30	36.04
$\mathbf{Q_{0.25}}$	48	61	31	45	25	46	37
Median	76	89	49	54	41	69	53
$\mathbf{Q_{0.75}}$	111	121	68	64	73	103	74

Table 4.3: Summary statistics of service life $\mathcal{X}$ given each building use-category

The medians are in general lower than the mean, which is also seen on Figure 4.4 as right-skewed distributions. The building use-categories with the most demolitions in the period are "Housing" and "Others". The average service life for "Others", $\overline{x}_{\text{Others}} = 53.92$ years, is the lowest, due to all the sheds, garages and outhouses that are classified as "Others". This is followed by $\overline{x}_{\text{Institution}}$ and $\overline{x}_{\text{Leisure}}$. The lowest standard deviation is $std(\mathcal{X}_{\text{Leisure}}) = 23.4$ years. This might be because this use-category is the most homogenous in terms of use-subcategories.

The longest service life is that of housing buildings and agricultural buildings, with $\overline{x}_{\text{Housing}} = 96.72$ years and $\overline{x}_{\text{Agriculture}} = 82.75$ years. The third quantile $Q_{0.75}(\mathcal{X}_{\text{Housing}})$ is 121 years. We saw this tail in the distribution, where some houses are preserved for a long time before demolition. The same is true for agricultural buildings, where $Q_{0.75}(\mathcal{X}_{\text{Agriculture}}) = 111$. These two building use-categories are also some of the use-categories that have been in use for the longest time, meaning that - even by pure chance - some of them will be very old when our observational period begins. We will return to this again when discussing outer wall materials of buildings in Section 4.2.3. From this section it is clear that buildings with different use-categories have very different decay processes, which is something to pay attention to when analyzing the BBR-dataset.

4.2.3 Outer wall material

Another variable that is likely to influence the service life of a building is the outer wall material[1]. The distribution of the outer wall material conditioned on δ is presented in Figure 4.5, showing the distribution for standing buildings $P(\text{outer wall material} = w|\delta = 0)$ and buildings demolished 2010-2024 $P(\text{outer wall material} = w|\delta = 1)$ for $w \in$ outer wall materials.

[1]From "Executive Order on Updating the Building and Housing Register (BBR)" [9] it is noted that the outer wall material refers to the main material covering the building's exterior walls between the ground and roof. If multiple materials are used, the one with the highest priority is recorded based on a specified ranking system (e.g., wood-clad concrete is considered concrete, not wood).

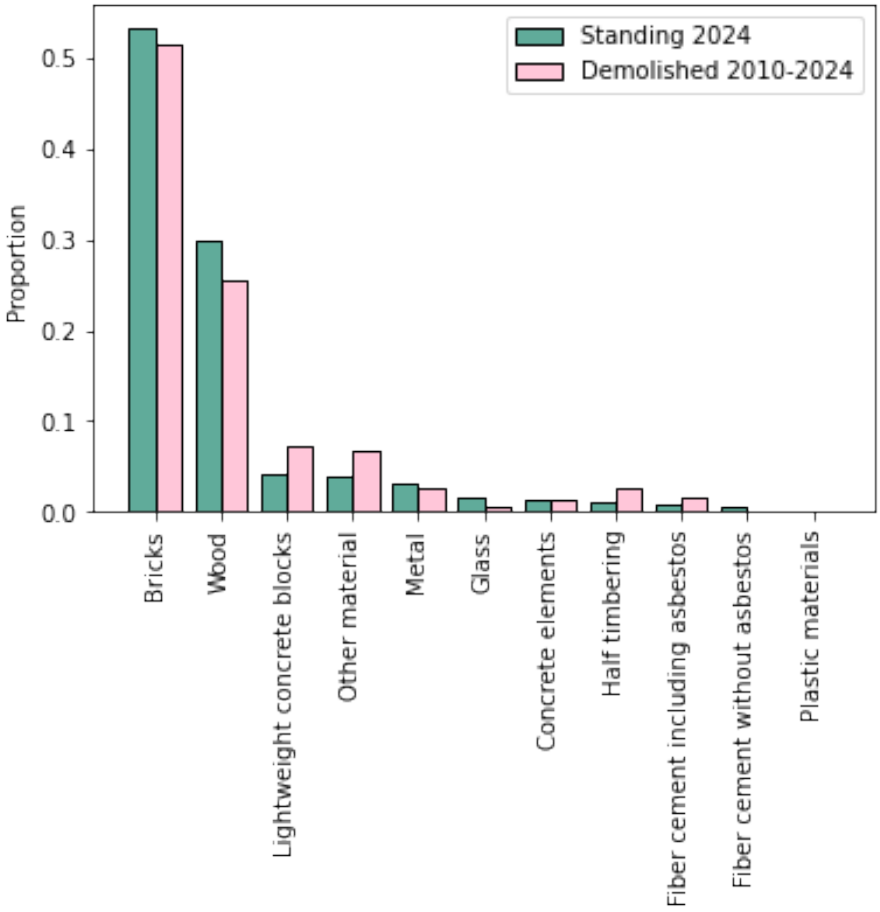


Figure 4.5: Distribution of outer wall material for buildings demolished between 2010 and 2024 ($\delta = 1$) compared with buildings not yet demolished ($\delta = 0$)

The most common materials are bricks and wood for both demolished and standing buildings. Using a χ^2-test on the two populations yields a p-value smaller than 10^{-16}. The test statistic is discussed in more detail in Section 5.2.2. This rejects the null hypothesis that the demolished population comes from the same population as the standing. The greater proportion of buildings with the outer wall material wood in the standing population could be tied to the greater proportion of buildings with the use-category "Others" in the standing building mass, as a lot of the sheds and garages have an outer wall constructed of wood. It is therefore an obvious next step to look at the interaction between the outer wall material, building use-category and service life.

4.2.4 Outer wall material by building use-category

To look at the interactions between the outer wall material, building use-category and service life we plot $P(\mathcal{X} \in \text{10-year bin} | \delta = 1, \text{outer wall material} = w)$, which is the distributions of service life conditioned on the outer wall material for buildings demolished in the observational period. The plot is stacked according to building use-category, and n corresponds to the number of buildings with the specific type of outer wall material.

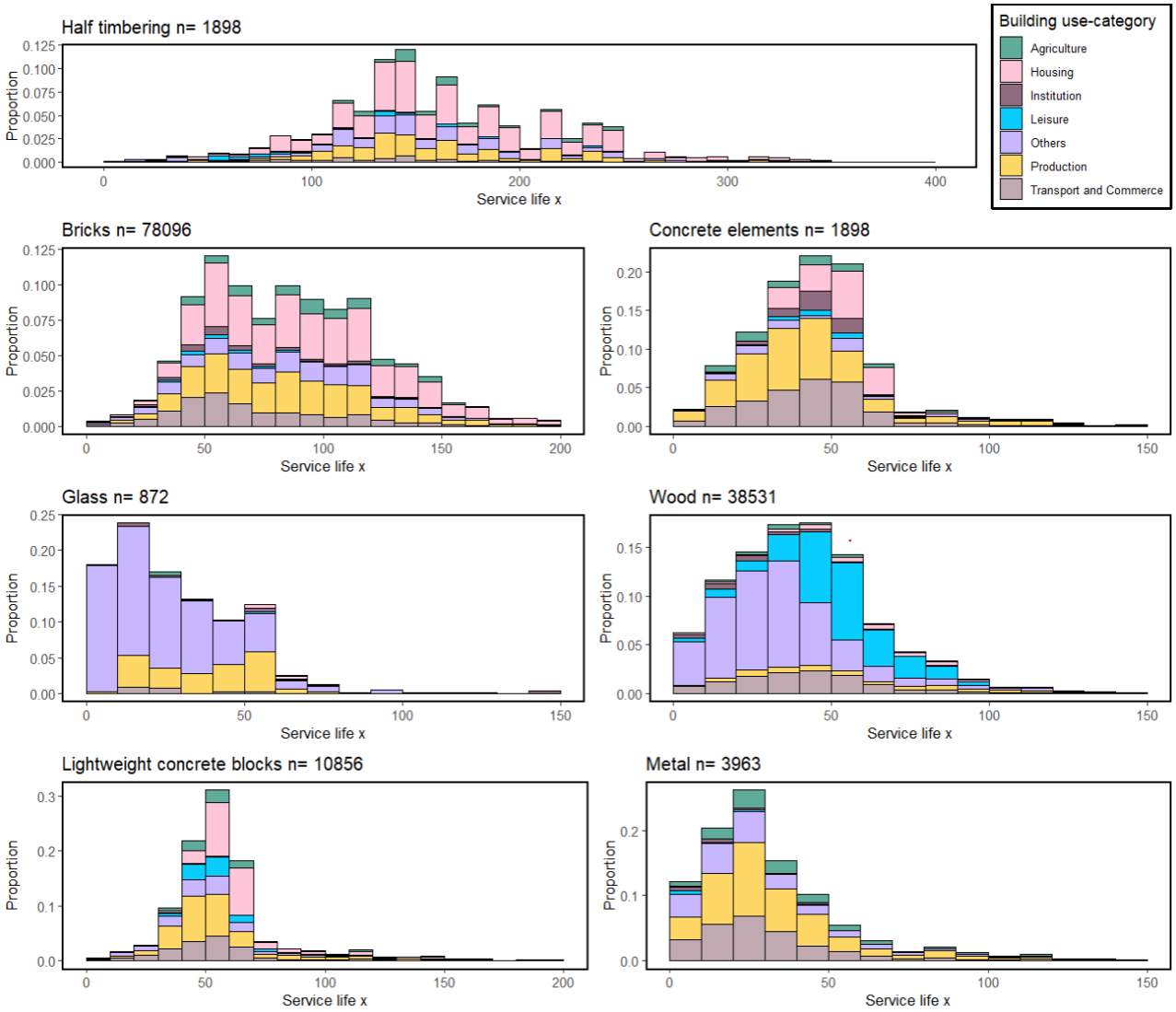


Figure 4.6: Service life distribution grouped by outer wall material and colored pr. use-category for buildings demolished between 2010 and 2024 ($\delta = 1$)

We denote the service lives of all buildings with outer wall material wood demolished 2010-2024 as the distribution of $\mathcal{X}_{\text{Wood}}$ and correspondingly for the other materials.

Most buildings have outer walls made of brick or wood. Of the buildings with outer wall material wood, 97.7% have a service life $\mathcal{X}_{wood} \leq 100$ with a median service life of 41 years. This group consist of younger buildings with use-category "Others" and older buildings with use-category "Leisure". The building use-category of "Others" is mostly sheds with a short average service life. For the leisure buildings with outer wall material wood, that have longer service lives, it is mostly summer houses. Only a small proportion of these houses have a service life exceeding 100 years, which may be attributable to the fact that the concept of the "summer house" only gained widespread popularity less than a century ago. One could suspect that other use-categories might be present in the censored groups instead. We therefore compared the proportions of building use-category for the outer wall material wood. The biggest difference found was a bigger proportion of the building use-category "Others" in the standing buildings where $P(\text{use-category} = \text{"Others"}|\text{outer wall material} = \text{wood}, \delta = 1) = 0.12$ and $P(\text{use-category} = \text{"Others"}|\text{outer wall material} = \text{wood}, \delta = 0) = 0.2$, but the general distribution is alike.

The distribution of $\mathcal{X}_{brick}$ is relatively spread out, with observed demolitions at a variety of service lives. Few are observed demolished before the 40 year mark, most likely due to brick being a robust outer wall material and because many of them are residential buildings that might have long-running mortgages. Though brick has been a common building material since the 1600s, Danish brick houses were popularized in the first half of the 20th century. Most of the demolished buildings with outer walls of brick are from this period with service lives between 50 and 120 years. The brick buildings demolished 2010-2024 are older than the brick buildings still standing today. We see that few buildings are demolished after 200 years. In the standing buildings, there are also only few buildings with brick outer walls older than 200 years, which corresponds to brick not being a popular building material a longer time ago. We find that the proportions broken down by use-categories differ with the standing population, as $P(\text{use-category} = \text{"Housing"}|\text{outer wall material} = \text{brick}, \delta = 1) = 0.19$ compared to $P(\text{use-category} = \text{"Housing"}|\text{outer wall material} = \text{brick}, \delta = 0) = 0.33$. This shows that a bigger proportion of the standing buildings are housing. In contrast, the "Production" use-category is over-represented in the demolished data: $P(\text{use-category} = \text{"Production"}|\text{outer wall material} = \text{brick}, \delta = 1) = 0.12$ compared to $P(\text{use-category} = \text{"Production"}|\text{outer wall material} = \text{brick}, \delta = 0) = 0.04$. Maybe people value the brick houses for aesthetics reasons, thus keeping them standing for a longer time, while prioritizing the functionality of new-built production facilities higher.

Buildings with outer walls of light concrete are overrepresented in the demolition data. Concrete is mostly used as an outer wall material for apartment complexes, "Transportation and Commerce" buildings and "Production" buildings, where many investors have an interest in keeping the buildings standing. It is also possible that these kinds of building uses are less susceptible to changing aesthetic tastes than single family houses. A topic of current debate is how to make the sturdy concrete buildings more climate friendly by using less polluting production methods or using modular building blocks to better allow reconstruction and reuse.

We note that the buildings with outer wall material half-timbering have long service lives. As this form of outer wall material was most popular in the 1700s and 1800s, most of their service lives will be left-censored. This leaves only the exceptionally sturdy ones and those deemed worthy of preservation to survive until the observational period. They are used mostly for "Housing", but also for "Agriculture", "Production" and "Others". A quirk for this group is the seeming 25-year cycles of $\mathcal{X}_{\text{half-timbering}}$ due to inaccurate dating of the construction year, where people opt to round to the nearest quadranscentennial (see Appendix A.1).

The demolished buildings with outer walls of glass have the shortest service life with mean $\overline{\mathcal{X}}_{\text{glass}} = 29$ years. They are mostly classified as "Others", a group which includes green houses, pavillions and garages, all typically made of glass and not necessarily build to last.

It is clear that the service life of a building is influenced by the outer wall material, which also correlates with the use-categories. When drawing conclusions based on the BBR-dataset it is therefore highly necessary to consider the interactions between the outer wall material, the building use-category and the service life.

4.2.5 Building area

We will now investigate the covered building area A among the demolished buildings, denoted $P(A \in 100m^2\text{bin}|\delta = 1)$, compared to the standing buildings, denoted $P(A \in 100m^2\text{bin}|\delta = 0)$.

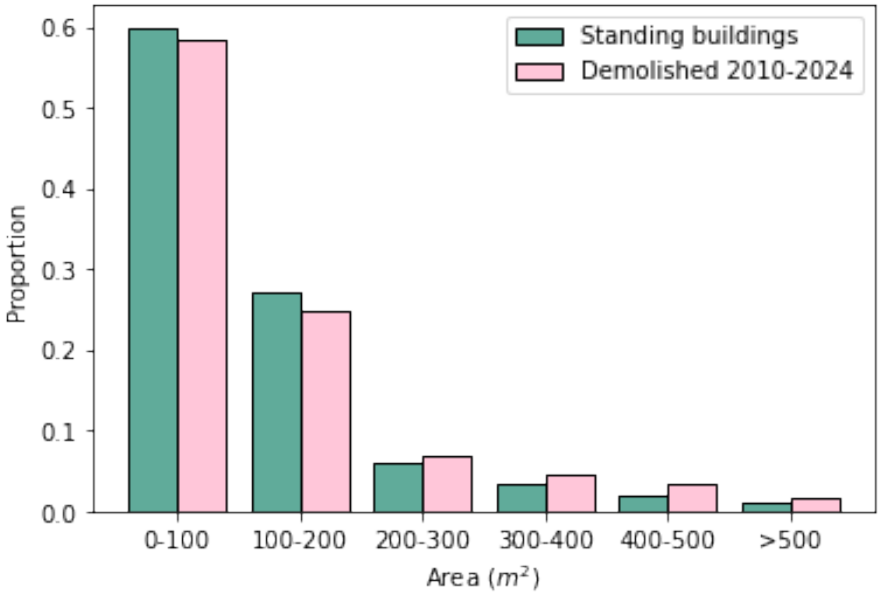


Figure 4.7: Comparison of areas between buildings demolished in 2010 to 2024 ($\delta = 1$) and buildings not yet demolished in 2024 ($\delta = 0$), with area divided into bins of 100 m^2.

The standing buildings are slightly overrepresented in the smaller areas compared to the demolished buildings. For the bigger buildings ($200m^2$and above) the proportion of demolished buildings is higher than the standing buildings. This is most likely due to the bigger proportion of buildings with the use-category "Others" in the standing building mass. As they generally have a smaller area (a mean area of $35.9m^2$ compared to the overall mean area of $138.0m^2$) this will increase the proportion of smaller buildings.

Using a χ^2 test yields a p-value smaller than 10^{-16}, which rejects the null hypothesis of the demolished population coming from the same population as the standing.

4.2.6 Heating method by building use-category

The most used heating methods vary among the different use-categories. The stacked barplot in Figure 4.8 shows the proportion of each heating method, stacked by building use-category. That is $P(\text{Heating} = h|\delta = 1)$ for $h \in$ Heating method:

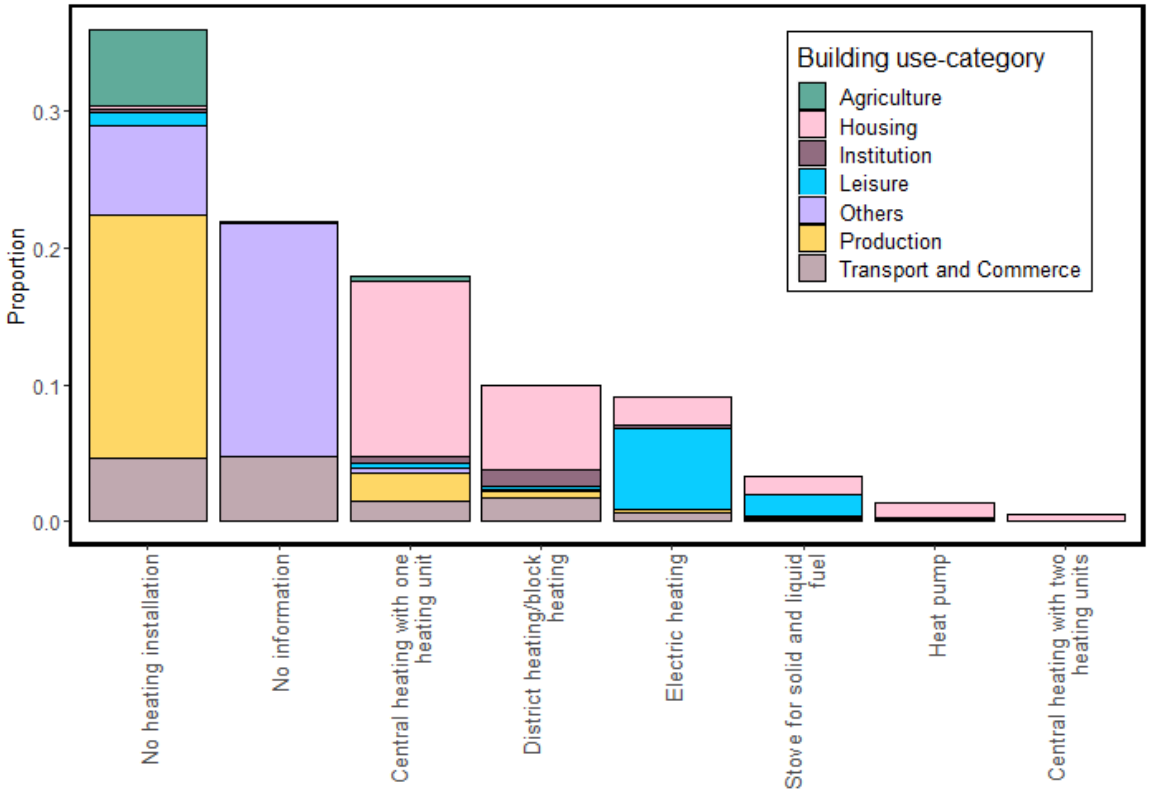


Figure 4.8: Distribution of heating method among buildings demolished between 2010 and 2024 ($\delta = 1$). The bars are colored according to building use-category.

From this plot it is clear that the type of heating is dependent on the building use-category. Among the buildings with no heating installed, the most represented use-categories are "Agriculture", "Production" and "Others". This makes sense as many agricultural and production buildings are not built for human residence and can thus save on the heating. In the category with no heating information, the "Others" category is dominant. It might be that people don't bother to add a heating method, when registering a new garage in BBR.

4.2.7 Roofing material

We have information about the roofing material which leads to the plot below, showing the distribution for the demolished buildings $P(\text{Roofing material} = r|\delta = 1)$ and the standing buildings $P(\text{Roofing material} = r|\delta = 0)$ for $r \in$ Roofing materials.

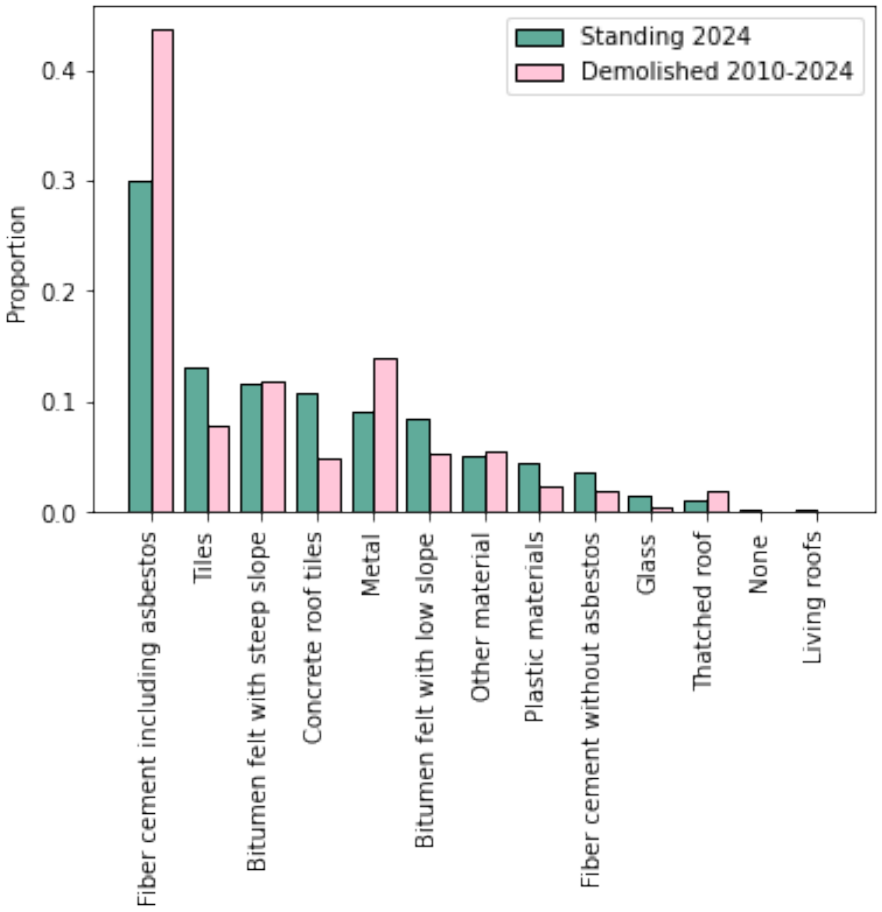


Figure 4.9: The proportions of buildings with different roofing materials among the buildings demolished 2010-2024 ($\delta = 1$) and buildings still standing in 2024 ($\delta = 0$)

The distributions of the roofing material for the standing and demolished buildings show different tendencies. Using a χ^2-test yields a p-value smaller than 10^{-16}, which rejects the null hypothesis of the demolished population coming from the same population as the standing. The most common materials for the demolished buildings are "Fiber cement incl. asbestos" and "Metal". The demolished buildings have a bigger proportion of buildings with "Fiber cement incl. asbestos" than the standing buildings. This hardly means that the houses were torn down because of the roofing, but rather that this roof type was common in houses built in the middle of the 20th century (mean building year for buildings for "Fiber cement incl. asbest" is 1943, compared to 1953 for buildings with other types of roofing.)

Among the buildings still standing ($\delta = 0$), tile roofing and cement stone (similar to tiling) constitute a much bigger proportion of the standing buildings than of the demolished. This harmonizes with our previous findings as these roofing materials are mostly used for housing, a building use-category which is underrepresented in the demolition data (See Figure 4.2).

4.2.8 Building reconstruction

For 20.0% of the demolished buildings, a reconstruction year. is indicated in the BBR. We want to examine when buildings are reconstructed and whether the reconstructed buildings differ from the rest of the buildings in service life. We compute the time before reconstruction $t_{\text{reconstruct}}$ for each reconstructed building. The table below shows the proportion of buildings that are reconstructed in each building use-category as well as some summary statistics for the reconstruction time:

	Agriculture	**Housing**	**Institution**	**Leisure**	**Others**	**Production**	**T & C**
% reconstr.	14.8%	36.7%	32.3%	33.1%	5.1%	15.0%	16.9%
$\overline{x}$	82.75	96.7	55.0	56.5	53.9	77.1	59.9
$\bar{t}_{\text{reconstruct}}$	51.56	62.35	40.73	27.01	49.24	45.38	43.20
$std(t_{\text{reconstruct}})$	40.64	45.60	32.25	22.70	42.25	36.14	38.01
$\text{Med}(t_{\text{reconstruct}})$	45	55	32	23	39	38	31

Table 4.4: Summary statistics for reconstruction time for buildings demolished 2010-2024. The statistics are shown for each building use-category along with the previously shown mean service life $\overline{x}$ for each building use-category

Buildings with use-category "Agriculture" and "Others" are rarely reconstructed. The "Others" category might rarely be reconstructed because people don't reconstruct their sheds and garages. For agricultural buildings, the need for reconstruction might be smaller, as there are rarely aesthetic concerns to motivate reconstruction and the wear and tear might be smaller than for residential buildings.

The average time before reconstruction pr. building use-category is in general lower than the mean service life $\overline{x}_j$ of a given building use-category j. The housing buildings are on average reconstructed after 62 years of life, but with a big standard deviation. Some house owners might follow shifting aesthetic trends more closely and reconstruct earlier to accomodate them. Other times residential buildings are only reconstructed in connection with house sales.

For each building use-category j we perform a χ^2-test to investigate whether for each building use-category j. For all building use-categories, we get p-values smaller than 10^{-11} and thus conclude that the reconstructed buildings have a service life distribution significantly different from that of the non-reconstructed buildings. We find that the service lives for reconstructed buildings is longer than for other buildings across all use-categories. Here it is worth noting that this comparison itself is subject to survivorship bias. The buildings being reconstructed have to be standing up until a suitable reconstruction-age. This means that realistically no buildings with a very short service life will be included in $\mathcal{X}_{\text{reconstructed}}$, which skews the distribution. The significantly different service life distributions also don't necessitate causation. It could be that reconstructed buildings are better cared for in general or that the reconstruction indicates wealthier building owners. Or it could be that reconstructing a building does indeed extend its service life.

4.3 Data quality assessment

The descriptive data analysis presented above and the modeling in the following sections are based on data from BBR. The registry is highly useful due to its centralized structure and comprehensive coverage; however, errors and inconsistencies are inevitable in citizen-reported data. In this section we will describe the extent of data inconsistencies, describe our considerations in using the data and compare the dataset from BBR with data available through Statistics Denmark.

4.3.1 Data inconsistencies

The data obtained through BBR may contain inconsistencies based on false, misleading or missing registrations. In the following we will document our findings of these.

Asbestos is registered in the variable `asbestos`. In the dataset we have 8,267 buildings with this registration even though 1,404,934 buildings have registered a roof material that includes asbestos. This discrepancy may be due to people forgetting, or choosing not to, register the use of asbestos in a building which goes against our assumption that no information recorded for this variable means no use of asbestos in the building.

For 150 of the buildings demolished in the period 2010 - 2024, the registered building year is the same as the registered demolition year. As the number is so low, it does not seem to be a systematical problem and may indicate buildings with foundational construction errors or similar extraordinary circumstances. Additionally 9 buildings were reported with a construction year after 2024. As this is clearly a mistake these were removed from the BBR-dataset.

Some buildings may have originally been built for one purpose, but found use in a different way later in its service life. This could be a summer house that later became an all-year housing option or the former Municipal Hospital in Copenhagen which now functions as a university facility.

We also note that for older buildings, approximative dating of construction year is widespread. This tendency is most pronounced in half-timbering as shown in Appendix A.1, but also happens on a more recent scale, where disproportionally big spikes in construction activity are observed in 1920, 1930 and 1950. While some of these buildings most likely have different construction years, the buildings do in fact exist and can not be removed. The deviance should be no more than 5 years if people round to nearest decade, which is not a huge deviance for such old buildings.

It can be tedious to update the BBR. For this reason, many may avoid doing so on small buildings, such as sheds and garages, or make registrations with only the bare minimum of information. This as well as the other described inaccuracies influences the credibility and affects the overall impression of the BBR-dataset, but none of the inaccuracies are of a scale or severity where it will affect large-scale modeling of the data.

4.3.2 Comparison with Statistics Denmark

Another publicly available source of Danish building data is Statistics Denmark's online statistics portal, Statistikbanken.dk. Here we acquire aggregated datasets regarding the building mass in Denmark, that are relevant in comparison to our dataset as well as for our future analyses:

- BYGB11 - (2007 - 2014) Buildings by location, owner, building use-category, and area
- BYGB12 - (2011 - 2024) Buildings by location, owner, building use-category, and area (in intervals)
- BYGV02 (1982 - 2023) Historical housing construction by construction stage and building use-category
- BYGV04 (1939 - 2023) Historical construction by construction stage and building use-category (m^2)
- BYGV05A - (1917 - 2023) Historical housing construction by construction stage and building use-category
- BYGV06 - (1916 - 2023) Average area in new buildings by building use-category

According to Statistics Denmark, this data is not comparable before and after 2011 due to the lack of demolition data and the inclusion of small buildings in the building mass after 2011 [43].

BYGB11 and BYGB12 share information for the years 2011-2014 but are made with two different definitions of "building", with BYGB12 using "building units" and including small buildings (sheds, garages). This results in a big difference in the building mass presented in the table below.

	2011	2012	2013	2014
BYGB11	2,529,579	2,531,701	2,530,838	2,532,730
BYGB12	4,327,292	4,345,679	4,362,070	4,380,826

Table 4.5: The total number of standing buildings according to the two datasets

To investigate these differences, we consider the database BYGV05A, containing historical housing construction data. The dataset covers the period from 1917 to 2024, but with a change in data collection method in 1981, leading to some inaccuracies. In 1981 they went from a handwritten survey based data collection to the register based data collection (BBR) [42]. BYGV05A covers the use-subcategories "Single family houses," "Terraced houses," "apartment complexes," and "others", meaning it is not a counting of the total number of buildings built. The categories are closer to the total number of housing buildings built, but still doesn't align with our building use-category "housing". We will still use it as a density function for construction year of housing buildings $g(c)$ when simulating data in Section 6.2. We will not use it to draw any conclusions, only to validate methods to be used on better data sources.

Computing the density of construction year for housing buildings $\mathcal{C}$, denoted $g(c)$:

$$g(c) = \frac{\text{buildings built in year } c}{\sum_{j=1917}^{2024} \text{buildings built in year } j}$$

yields the density distribution shown below:

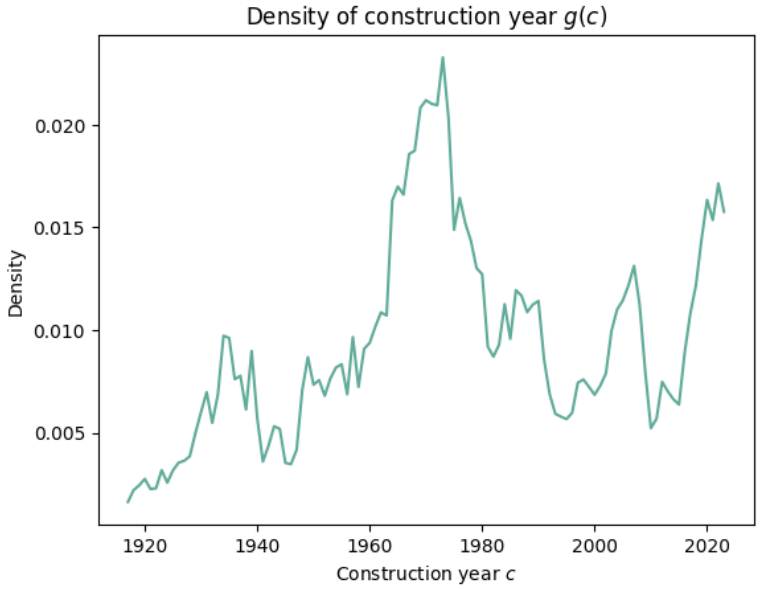


Figure 4.10: Probability density function of construction year for housing buildings based on Statistics Denmark's database BYGV05A

While we find the absolute numbers in this data to be unreliable, the graph follows societal economic trends with increased building activity in the 1960s and 1970s as well as in the years leading up to the financial crisis in 2008.

We tried to compare the number of constructed buildings pr year (BYGV05A) to the complete BBR dataset $f_{C|D>2010}(c)$. It is worth noting that BBR and Statistics Denmark use different categories

for building use, making them hard to compare. Generally we find big differences as shown on Figure 4.11

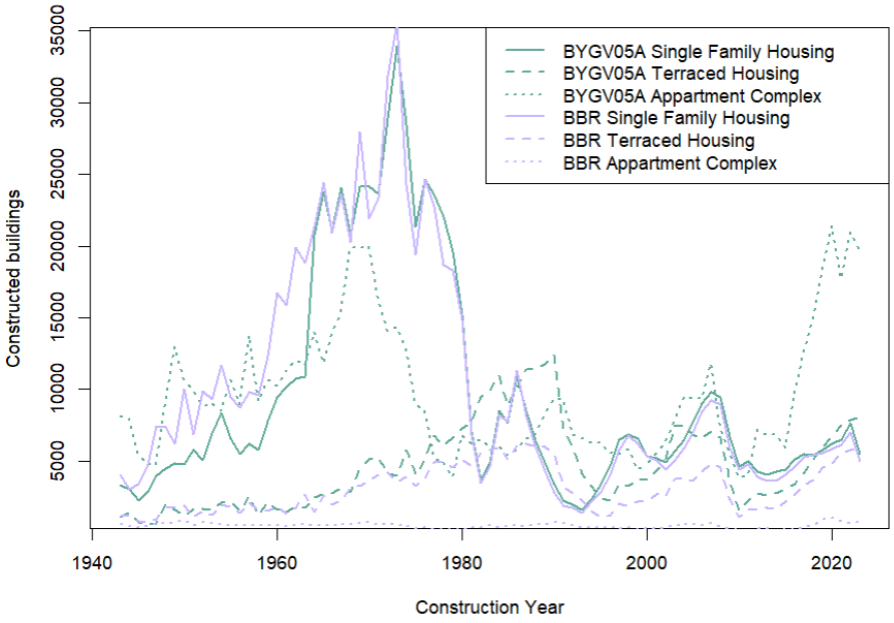


Figure 4.11: Number of constructed buildings from BYGV05A compared to BBR. The green lines represent BYGV05 data and the purple lines represent BBR data. The different linetypes correspond to the different examined use-subcategories.

For the terraced houses and apartment complexes the BBR data is generally lower (especially for the apartment complexes.) This leads us to believe that the Statistics Denmark-data might consider each "housing unit" in an apartment complex as a building. By comparing the average area of an apartment complex in the BBR data to the data in BYGV06, the apartment complexes in the BBR data have an average area of $306m^2$ while the Statistics Denmark-data has an average of $82m^2$. This further supports the idea that these datasets aren't comparable. One approach to work around this could be to compare the constructed area instead of constructed units. Another would be to limit the analysis to single-family houses, as these seem well-aligned from the 1960s. This could be done in the future, if more fine-grained analyses are of particular interest.

The oldest data available through Statistics Denmark for constructed area pr. year is BYGV04, which dates back to 1939. Comparing this to BBR yields the following illustration:

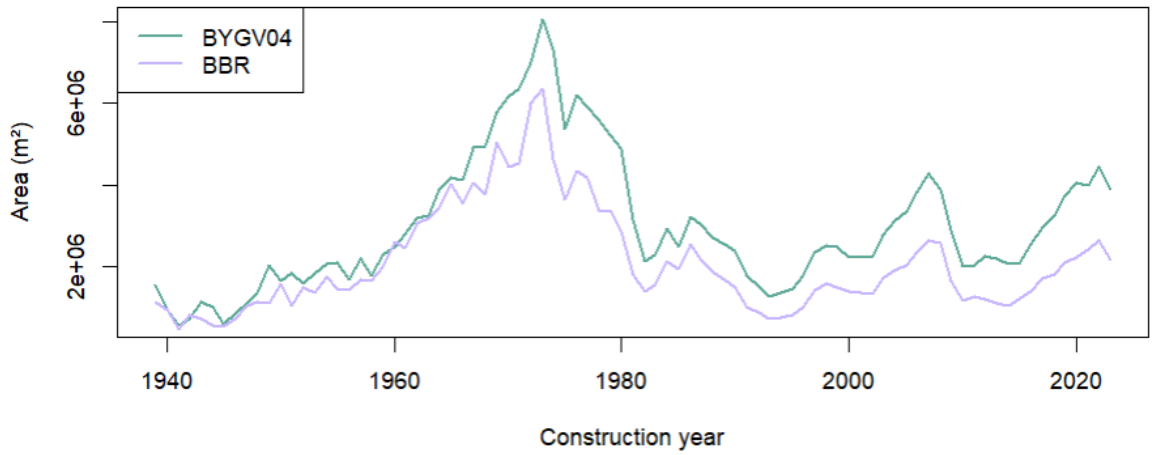


Figure 4.12: Comparison of total constructed area for each year between BBR and BYGV04.

As the BBR-data doesn't contain the buildings demolished before 2010 we would expect the purple graph to be below the green at all times, and the gap between the green and the purple graph to represent the demolished data. We would then expect the graphs to match almost perfectly from 2010. Unfortunately, this is not the case, though they do show the same general tendencies.

We tried to use the average areas for different use-categories (BYGV06) together with the historical construction data (BYGV05A) to make an estimate for the historical constructed area. The investigation can be found in appendix A.2 where we found that single family houses follow somewhat the same curve as Figure 4.11 for the number of buildings but that the BYGV05 curve for the apartment complexes is now a lot closer to the BBR curve. However, the differences remain too big for proper comparison or validation.

When comparing BYGV05A to BYGV02 the total numbers of buildings are very similar in the overlapping years, but the use-categories are completely different. BYGV02 seems to use more categories but some of them are almost completely empty. As an example it states that one shed is built pr. year on average. Statistics Denmark's use categories generally seem error-prone and therefore highly difficult to use for future analysis.

This data exploration demonstrates the range of different data sources on the Danish building stock and construction through time. Due to changes in definitions and in data collection methods they are very hard to compare, even for the most recent years. We have tried to get access to the historical BBR such that we for each year could compare the building stock and get a full dataset on demolitions pr. year. This would give us the information needed to carry out a proper service life analysis. Unfortunately this was not possible within the time frame of this project, due to very long response times at the Danish National Archives. We have also applied for access to the micro data through Statistics Denmark, another promising but yet unresolved path for data acquisition. In the following sections of the report we will either work with our dataset and be careful about the kinds of conclusions that might be drawn from it, or with simulated data in order to explore methods in a proof-of-concept manner.

5 Service life analysis

5.1 Theoretical background on service lives

To model building service life, we may consider each building's service life to be a realization of an underlying stochastic process. The following section will describe the theory behind varying stochastic models and how they may apply to building service life estimation.

We begin by considering classical probability distributions - Weibull, gamma and lognormal - which have previously been employed in academic papers for service life modeling. We then move on to describe the mathematics behind Markov jump processes, how it relates to phase-type distributions and prove some general properties of the phase-type distribution. For these sections, we will follow the notation of (Bladt and Nielsen, 2017).

5.1.1 Classical probability distributions

We will briefly describe the Weibull, gamma and lognormal distributions, as these simple distributions are often used in life time modeling. The formulas shown below are in line with the R-package `fitdistr` [17], which we will use for model estimation.

Weibull Distribution

The Weibull distribution is often used to model the time to failure in survival analysis. It is defined by a shape parameter $\kappa > 0$ and a scale parameter $\lambda > 0$. The probability density function and cumulative distribution function of the Weibull distribution is given by:

$$f(x) = \begin{cases} \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} e^{-(x/\lambda)^{\kappa}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^{\kappa}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Gamma Distribution

The gamma distribution is often used in queuing models and reliability analysis. It is defined by the shape parameter $\kappa > 0$ and the scale parameter $\lambda > 0$. The probability density function and cumulative distribution function of the gamma distribution are shown below:

$$f(x) = \begin{cases} \frac{\frac{1}{\lambda}^{\kappa}}{\Gamma(\kappa)} x^{\kappa-1} e^{-\frac{1}{\lambda}x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

$$F(x) = \frac{\gamma(\kappa, \frac{1}{\lambda}x)}{\Gamma(\kappa)}, \quad x > 0,$$

where $\Gamma(\kappa)$ is the gamma function, and $\gamma(\kappa, \frac{1}{\lambda}x)$ is the lower incomplete gamma function. For $\kappa \in \mathbb{N}$ the Gamma-distribution is a phase-type distribution, as it is then the sum of exponentially distributed random variables, which is the phase-type distributed Erlang-distribution [31].

Lognormal Distribution

The lognormal distribution is often used to model positive, skewed data. If a random variable $\mathcal{X}$ is lognormal distributed, then the random variable $\mathcal{Y} = \ln(\mathcal{X})$ is normally distributed. The lognormal distribution is then defined by the mean parameter $\mu \in \mathbb{R}$ and the standard deviation parameter

$\sigma > 0$ for $\mathcal{Y}$. The probability density function and cumulative distribution function of the lognormal distribution is:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

$$F(x) = \begin{cases} \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right], & x > 0, \\ 0, & x \leq 0, \end{cases}$$

where $\operatorname{erf}(\cdot)$ is the error function.

We will fit models of all three of these distributions to our service life data. The parameter estimates are made using maximum likelihood optimizers and the standard error for each parameter is found from the Hessian matrix at the maximum likelihood solution [17].

These models can only describe the time of demolition and do not provide insights into the process prior to demolition. A more advanced modeling framework is to model the whole building decay process as a Markov jump process.

5.1.2 Markov jump processes

We can consider the decay of a building to follow an underlying Markov jump process where the total time spent in the process is the service life of the building.

A Markov jump process $\{X_t\}_{t\geq 0}$ is a stochastic process in continuous time, taking values in a countable state space E. Because the state space is finite, the transitions between states occur in jumps, from which they get their name. Markov jump processes obey the Markov property that

$$P(X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}, \ldots, X_{t_1} = i_1, X_0 = i_0) = P(X_{t_n} = i_n | X_{t_{n-1}} = i_{n-1}) \tag{5.1}$$

for all $t_n > t_{n-1} > \cdots > t_1 > 0$ and all $i_n, i_{n-1}, \ldots, i_0 \in E$. This property is often stated as "the future depends on the past only through the present". We are working with time-homogeneous jump processes where the transition rates do not depend on the time of the transitions, but only on the time difference h. We then define the transition probability of going from state i to state j as

$$p_{ij}(h) = P(X_{t+h} = j | X_t = i). \tag{5.2}$$

Ordering these probabilities in a matrix, we get the transition matrix $\mathbf{P}$ for the Markov jump process $\{X_t\}_{t\geq 0}$ defined as

$$\mathbf{P}(h) = \{p_{ij}(h)\}_{i,j\in E}. \tag{5.3}$$

The Markov jump processes of interest have a state space $E \in \{1, 2, \ldots, p, p+1\}$ where states $1, 2, \ldots, p$ are transient and state $p+1$ is absorbing. An absorbing state is a state where the rate of leaving is 0, meaning the process cannot exit the state once the state is reached. A transient state is a state for which the probability of visiting it goes to 0 as $t \to \infty$, meaning the process cannot terminate in a transient state.

5.1.3 The phase-type distribution

The total time until the described Markov jump chain reaches state $p+1$ can be interpreted as the lifetime of the process and is denoted τ:

$$\tau = \inf\{t > 0 | X_t = p + 1\}. \tag{5.4}$$

Then τ follows a phase-type distribution:

$$\tau \sim PH_p(\boldsymbol{\pi}, \mathbf{T}),$$

where $\mathbf{T} \in \mathbb{R}^{p \times p}$ is called the sub-intensity matrix and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p) \in \mathbb{R}^p$ is the initial distribution of the process over the transient states. The representation $(\boldsymbol{\pi}, \boldsymbol{T})$ of a phase-type distribution is not unique, such that more than one parameter representation give rise to the exact same distribution.

The vector $\boldsymbol{\pi}$ has entries $\pi_i = P(X_0 = i)$. It is assumed in this report that the probability of starting in the absorbing state $P(X_0 = p+1) = 0$. The initial distribution of the process is thus $(\boldsymbol{\pi}, 0)$. As all processes must start in one state, it should hold that

$$\boldsymbol{\pi} \mathbf{1}_p = 1, \tag{5.5}$$

where $\mathbf{1}_p$ is a column vector of p 1's.

The sub-intensity matrix $\mathbf{T}$ represents the rates at which jumps occur. The rate at which the process jumps from state i to state j is given by element T_{ij}. The rates by which the process jumps from state i to the absorbing state $p+1$ is defined in the exit rate vector $\mathbf{t}$ as entry t_i. The time spent in state i before a jump is exponentially distributed with intensity $-T_{ii}$. The expected time a process sojourns in a state before jumping is denoted $\overline{s}_i$:

$$\overline{s}_i = \frac{1}{-T_{ii}} \tag{5.6}$$

Using $\mathbf{T}$ and $\mathbf{t}$ the Markov jump process has an intensity matrix $\boldsymbol{\Lambda}$ defined by:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{T} & \mathbf{t} \\ \mathbf{0}_p & 0 \end{bmatrix}, \tag{5.7}$$

where $\mathbf{0}_p$ is a row vector of p 0's. As the probability should be conserved over time, we want the inflow and outflow of each state to sum to 0, meaning that for each row $i = \{1, 2, .., p+1\}$:

$$\sum_j \Lambda_{ij} = 0 \tag{5.8}$$

$$T_{ii} + \sum_{j \neq i} T_{ij} + t_i = 0. \tag{5.9}$$

Which leads to

$$T_{ii} = -t_i - \sum_{j \neq i} T_{ij}.$$

Equation 5.8 can be expressed in matrix notation:

$$\begin{aligned} \boldsymbol{\Lambda} \mathbf{1}_{p+1} &= \mathbf{0}_{p+1}, \\ \begin{bmatrix} \boldsymbol{T} & \mathbf{t} \\ \mathbf{0}_p & 0 \end{bmatrix} \mathbf{1}_{p+1} &= \mathbf{0}_{p+1} \\ \begin{bmatrix} \boldsymbol{T}\mathbf{1}_p + \mathbf{t} \\ \mathbf{0}^\top \mathbf{1}_p + 0 \end{bmatrix} &= \mathbf{0}_p, \\ \boldsymbol{T}\mathbf{1}_p + \mathbf{t} &= \mathbf{0}_p, \\ -\boldsymbol{T}\mathbf{1}_p &= \mathbf{t}. \end{aligned} \tag{5.10}$$

From this we see that $\mathbf{t}$ is fully defined by $\mathbf{T}$, which explains the omission of $\mathbf{t}$ in the representation.

The Markov jump process can be visualized as a weighted directed graph where each state corresponds to a node. An edge exists from node i to node j if $T_{ij} > 0$, in which case the edge will have the edge weight T_{ij}. In the most general case, where jumps are possible from all transient states $i \in E^* = \{1, \ldots, p\}$ to all states $j \in E$, the chain will look like this:

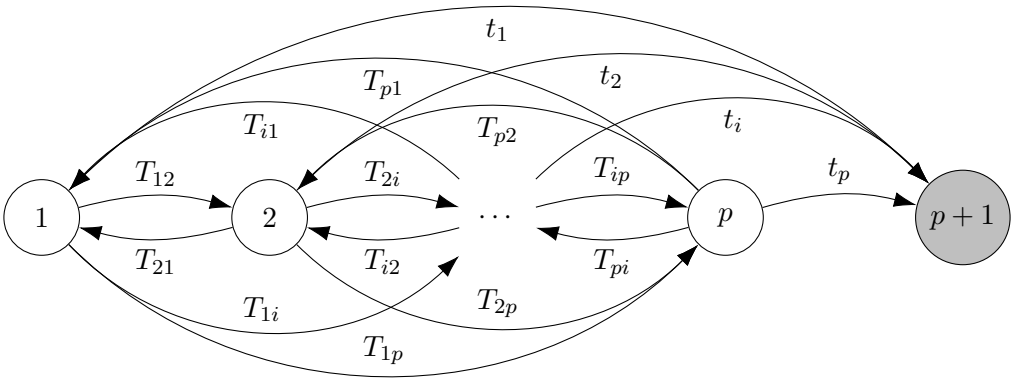


Figure 5.1: A general structure Markov chain with p transient states and one absorbing state. The edge weights T_{ij} represent the rate at which a jump occurs from state i to state j, given that the process is in state i.

Less general structures are often found in real life applications. One subclass emerges when the graph is acyclical, in which case we will also call the Markov chain acyclical. Another common structure is the Coxian structure, where a jump from state i can only happen to either state $i+1$ or the absorbing state $p+1$. The Coxian Markov chain is visualized below:

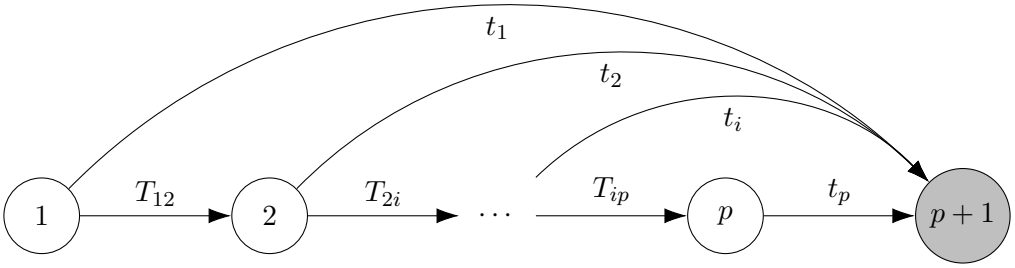


Figure 5.2: A Markov chain with Coxian structure with p transient states and one absorbing state.

The transition probabilities at time s are denoted $P(X_s = j|X_0 = i)$. For $i \in E^*$ and $j \in E$ they arae computed by:

$$P(X_s = j|X_0 = i) = \left(\mathrm{e}^{\mathbf{\Lambda} s}\right)_{ij} \tag{5.11}$$

This is proved on page 131 in Bladt and Nielsen and can be computed numerically using the Taylor series expansion:

$$e^{\mathbf{\Lambda} s} = \mathbf{I} + \sum_{n=1}^{\infty} \frac{(\mathbf{\Lambda} s)^n}{n!} \tag{5.12}$$

which is implemented in R through the `expm`-package (Maechler et al., 2024).

5.1.4 Properties of phase-type distributions

Some important properties of the phase-type distribution will be shown and proven in the following section.

Theorem 5.1.1. *Consider a continuous random variable $\tau \sim PH_p(\boldsymbol{\pi}, \boldsymbol{T})$. Its density function $f : \mathbb{R} \to \mathbb{R}^+$ is then given by*

$$f(s) = \boldsymbol{\pi} e^{\boldsymbol{T}s} \boldsymbol{t}$$

for $s \geq 0$.

Proof. Let f be the density of $\tau \sim PH_p(\boldsymbol{\pi}, \mathbf{T})$ and let $\{X_t\}_{t\geq 0}$ be the underlying Markov jump process. The infinitesimal density of τ is denoted:

$$f(s)ds = P(\tau \in (s, s + \mathrm{d}s)).$$

If $\tau \in (s, s + ds)$ then $\{X_t\}_{t\geq 0}$ must be in some transient state j at time s. From Equation 5.11 we know the transition probability to j given the initial state i and a time s:

$$P(X_s = j \mid X_0 = i) = (\mathrm{e}^{\mathbf{T}s})_{ij}. \tag{5.13}$$

And by definition we know the probability of the initial state being i is

$$P(X_0 = i) = \pi_i. \tag{5.14}$$

The probability of a jump from j to the absorbing state $p + 1$ within an infinitesimal time interval of length ds is given by the exit rate vector t:

$$P(X_{s+\mathrm{d}s} = p + 1 \mid X_s = j) = t_j \mathrm{d}s. \tag{5.15}$$

As the process can only sojourn in one state at a time, we can split the probability into sums. We will use conditional probability to condition the probability on $X_s = j$ and $X_0 = i$ and Equations 5.15 , 5.14 and 5.13 :

$$\begin{aligned}
f(s)ds &= P(\tau \in (s, s + \mathrm{d}s)), \\
&= \sum_{j=1}^{p} P(\tau \in (s, s + \mathrm{d}s) \mid X_s = j) P(X_s = j), \\
&= \sum_{j=1}^{p} P(\tau \in (s, s + \mathrm{d}s) \mid X_s = j) \left(\sum_{i=1}^{p} P(X_s = j | X_0 = i) P(X_0 = i) \right), \\
&= \sum_{j=1}^{p} t_j \mathrm{d}s \sum_{i=1}^{p} (\mathrm{e}^{\mathbf{T}s})_{ij} \pi_i, \\
&= \sum_{j=1}^{p} \sum_{i=1}^{p} t_j \mathrm{d}s (\mathrm{e}^{\mathbf{T}s})_{ij} \pi_i, \\
&= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i (\mathrm{e}^{\mathbf{T}s})_{ij} t_j \mathrm{d}s, \\
&= \boldsymbol{\pi} (\mathrm{e}^{\mathbf{T}s})\ \mathbf{t} \mathrm{d}s.
\end{aligned}$$

□

Theorem 5.1.2. *Let $\tau \sim PH_p(\boldsymbol{\pi}, \boldsymbol{T})$ and $F : \mathbb{R} \to [0, 1]$ be the distribution function of τ. Then $F(s)$ is given by*

$$F(s) = 1 - \boldsymbol{\pi} e^{\boldsymbol{T}s} \boldsymbol{1}_p, \quad s \geq 0$$

Proof. Given an underlying process $\{X_t; t \geq 0\}$ the distribution function denotes

$$F(s) = P(\tau \leq s).$$

We will consider the complementary probability that the process has *not* yet reached the absorbing state at time s:

$$1 - F(s) = P(\tau > s).$$

It follows from 5.4 that

$$P(\tau > s) = P(X_s \in E^*),$$

where $E^* = \{1, \ldots, p\}$ denotes the state space of the process excluding the absorbing state. As the process can only sojourn in one state at a time, these events are disjoint and this probability can be broken into the set of transient states:

$$\begin{aligned} P(\tau > s) &= P\left(\bigcup_{j=1}^{p} \{X_s = j\}\right), \\ &= \sum_{j=1}^{p} P(X_s = j). \end{aligned}$$

This sum can be expanded using the law of total probability, as we know that the set of transient states $\{1, \ldots, p\}$ spans the complete state space for initialization:

$$P(\tau > s) = \sum_{j=1}^{p} \sum_{i=1}^{p} P(X_s = j \mid X_0 = i) P(X_0 = i)$$

Inserting Equation 5.13 and Equation 5.14 yields:

$$\begin{aligned} &= \sum_{j=1}^{p} \sum_{i=1}^{p} (\mathrm{e}^{\mathbf{T}s})_{ij} \pi_i, \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i (\mathrm{e}^{\mathbf{T}s})_{ij}, \\ &= \boldsymbol{\pi} \mathrm{e}^{\mathbf{T}s} \mathbf{1}_p. \end{aligned}$$

We then have that

$$\begin{aligned} 1 - F(s) &= \boldsymbol{\pi} \mathrm{e}^{\mathbf{T}s} \mathbf{1}_p \\ F(s) &= 1 - \boldsymbol{\pi} \mathrm{e}^{\mathbf{T}s} \mathbf{1}_p \end{aligned}$$

As stated in the theorem □

Corollary 5.1.2.1. *Let $\tau \sim PH_p(\boldsymbol{\pi}, \boldsymbol{T})$ and $S : \mathbb{R} \to [0, 1]$ be the survival function of τ, defined as $S(\tau \geq s)$. Then $S(s)$ is given by*

$$S(s) = \boldsymbol{\pi} e^{\boldsymbol{T}s} \boldsymbol{1}_p, \quad s \geq 0$$

Proof. Shown directly in Theorem 5.1.2: □

Often, one is interested in finding the expected time that the underlying Markov process will spend in state j prior to absorption given an initial state i.

Theorem 5.1.3. *Let $\tau \sim PH_p(\boldsymbol{\pi}, \boldsymbol{T})$. Then the entries of the matrix $\boldsymbol{U} = \{u_{ij}\} = (-\boldsymbol{T})^{-1}$ is the expected time that the underlying Markov process will spend in state j prior to absorption given an initial state i. This matrix is called the Green matrix.*

Proof. We will compute this by defining Z_j as the time spent in a transient state j prior to absorption. Given an initial state i we can write Z_j as the integral of the indicator of the process being in state j:

$$\mathbb{E}(Z_j|X_0 = i) = \mathbb{E}\left(\int_0^\tau \mathbb{I}(X_t = j \mid X_0 = i)dt\right).$$

Expanding the integral to the whole positive axis and using that the expectation of an indicator is the probability of the indicated event, we get

$$\mathbb{E}(Z_j \mid X_0 = i) = \int_0^\infty P\left(X_t = j, \tau \geq t \mid X_0 = i\right) dt$$

A result from Bladt and Nielsen 2017 (page 134) is that $P(X_t = j, \tau \geq t \mid X_0 = i) = (e^{\mathbf{T}t})_{ij}$:

$$\mathbb{E}(Z_j \mid X_0 = i) = \int_0^\infty \left(e^{\mathbf{T}t}\right)_{ij} \, dt.$$

The computed value $\mathbb{E}(Z_j \mid X_0 = i)$ is the ij'th entry in $\mathbf{U}$. This integral is evaluated for every entry:

$$\begin{aligned}\mathbf{U} &= \int_0^\infty \left(e^{\mathbf{T}t}dt\right) \\ &= \left[\mathbf{T}^{-1}e^{\mathbf{T}t}\right]_0^\infty\end{aligned}$$

By definition of the transient states $1, \ldots, p$ $P_i\left(X_t = j, \tau \geq t\right) \to 0$ so $\lim_{t\to\infty} e^{\mathbf{T}t} \to \mathbf{0}$:

$$\mathbf{U} = -\mathbf{T}^{-1}$$

As stated in the theorem. We note that the subintensity $\mathbf{T}$ is invertible if and only if the states 1,2,..., p are transient, which always holds for a phasetype distribution [11]. □

We now show how to find the expectation of a continuous phase-type distributed random variable $\mathbb{E}(\tau)$.

Theorem 5.1.4. *Let $\tau \sim PH_p(\boldsymbol{\pi}, \boldsymbol{T})$. Then the first moment of τ is given by $\mathbb{E}(\tau) = \boldsymbol{\pi}(-\boldsymbol{T}^{-1})\boldsymbol{1}_p$*

Proof. Using Theorem 5.1.4 from above we know that $\mathbb{E}_i(\tau) = \sum_{j=1}^p u_{ij}$. We can get the total time spent before absorption by summing over all initial states states

$$\begin{aligned}\mathbb{E}(\tau) &= \sum_{i=1}^p \boldsymbol{\pi}_i \mathbb{E}_i(\tau) \\ &= \sum_{i=1}^p \boldsymbol{\pi}_i \sum_{j=1}^p u_{ij} \\ &= \boldsymbol{\pi}\mathbf{U}\mathbf{1} \\ &= \boldsymbol{\pi}(-\mathbf{T}^{-1})\mathbf{1}_p\end{aligned}$$

Where the last line uses Theorem 5.1.3.

□

The theorems from this section are used for the further analysis of service lives of buildings. To use them on real life data, we need to establish a method for estimating the parameters $(\boldsymbol{\pi}, \mathbf{T})$.

5.2 Methodology for service life analysis

This section outlines our methodology for estimating parameters of a given phase-type model to our data. This is done through an Expectation-Maximization-algorithm (abbreviated EM-algorithm) where the most likely $\boldsymbol{\pi}$ and $\mathbf{T}$ are found. We then describe goodness-of-fit statistics and diagnostics plots to assess the strengths and weaknesses of each model in order to choose the best fitting model to our data.

5.2.1 Model fitting

Our models will be fitted using the EM-algorithm. The EM-algorithm implementation described in this section is based on the procedure presented in (Asmussen, Nerman, and Olsson 1996) .

The data available on the Markov process is the values of observed service life τ. This data is incomplete as it only tells how much time it takes from initialization to absorption. The complete Markov process would be described by the states visited $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_M$ before absorption and the sojourn times in each state visited $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\mathcal{M}$, with $\mathcal{M}$ denoting the number of jumps until the process is absorbed. Then the total time in the process before absorption is $\tau = \sum_{i=1}^{m} \int_i$ and the observation is then completely described by $(i_1, \ldots, i_m, s_1, \ldots, s_m)$. Now suppose we have n independent samples of the Markov process. The complete dataset is then given by

$$\mathbf{x} = (i_1^{[1]}, \ldots, i_{m_1}^{[1]}, s_1^{[1]}, \ldots, s_{m_1}^{[1]}, \ldots, i_1^{[n]}, \ldots, i_{m_n}^{[n]}, s_1^{[n]}, \ldots, s_{m_n}^{[n]})$$

where the superscript denotes the sample number from $1, \ldots, n$. The vector of service lives is then

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n) = \left(s_1^{[1]} + \cdots + s_{m_1}^{[1]}, \ldots, s_1^{[n]} + \cdots + s_{m_n}^{[n]} \right).$$

We let $i \in E^* = \{1, 2, \ldots, p\}$ and $j \in E \setminus \{i\}$. Using all n data points in $\mathbf{x}$ we count the following quantities:

- $\mathcal{B}_i$: The number of processes starting in state i:

$$\mathcal{B}_i = \sum_{v=1}^{n} \mathbb{I}(\mathcal{I}_1^{[v]} = i)$$

- $\mathcal{N}_{ij}$: The number of jumps from state i to state j counted for all processes:

$$\mathcal{N}_{ij} = \sum_{v=1}^{n} \sum_{k=1}^{\mathcal{M}_v} \mathbb{I}(\mathcal{I}_k^{[v]} = i, \mathcal{I}_{k+1}^{[v]} = j)$$

- $\mathcal{Z}_i$: The total time spent in state i prior to absorption counted for all processes:

$$\mathcal{Z}_i = \sum_{v=1}^{n} \prod_{k=1}^{\mathcal{M}_v} \mathbb{I}(\mathcal{I}_k^{[v]} = i) \mathcal{S}_k^{[v]}$$

Using these, we can write the likelihood [8]:

$$L(\boldsymbol{\theta}; x) = \prod_{i \in E^*} \pi_i^{\mathcal{B}_i} \prod_{i \in E^*} e^{-T_{ii} \mathcal{Z}_i} \prod_{i \in E^*} \prod_{j \in E \setminus \{i\}} T_{ij}^{\mathcal{N}_{ij}} \tag{5.16}$$

The algorithm consists of an initialization and an E- and M-step:

Initialization: The initial parameter values $(\boldsymbol{\pi}, \mathbf{T})$ are sampled at random so they fulfill the conditions stated in Equation 5.5 and Equation 5.8.

E-step: Each iteration begins with the E-step, where the conditional expectations of $\mathcal{B}_i$, $\mathcal{N}_{ij}$ and $\mathcal{Z}_i$ are computed given the observed sample $\boldsymbol{\tau}$ and the current estimates of $(\boldsymbol{\pi}, \mathbf{T})$. Formulas for each conditional expectation are derived in (Asmussen, Nerman, and Olsson, 1996) and shown below:

$$\begin{aligned}
\mathbb{E}(\mathcal{B}_i^{[v]} \mid \tau = \tau_v, \boldsymbol{\pi}, \mathbf{T}) &= \frac{\pi_i \exp(\boldsymbol{T}\tau_v)\mathbf{t}}{\boldsymbol{\pi}\exp(\boldsymbol{T}\tau_v)\mathbf{t}}, \\
\mathbb{E}(\mathcal{Z}_i^{[v]} \mid \tau = \tau_v, \boldsymbol{\pi}, \mathbf{T}) &= \frac{\int_0^{\tau_v} \boldsymbol{\pi}\exp(\boldsymbol{T}u)\mathbf{1}_i\mathbf{1}_i^\top \exp(\boldsymbol{T}(\tau_v - u))\mathbf{t}\, du}{\boldsymbol{\pi}\exp(\boldsymbol{T}\tau_v)\mathbf{t}}, \qquad (5.17) \\
\mathbb{E}(\mathcal{N}_{ij}^{[v]} \mid \tau = \tau_v, \boldsymbol{\pi}, \mathbf{T}) &= \frac{T_{ij}\int_0^{\tau_v} \boldsymbol{\pi}\exp(\boldsymbol{T}u)\mathbf{1}_i\mathbf{1}_j^\top \exp(\boldsymbol{T}(\tau_v - u))\mathbf{t}\, du}{\boldsymbol{\pi}\exp(\boldsymbol{T}\tau_v)\mathbf{t}} \qquad (5.18) \\
\mathbb{E}(\mathcal{N}_{i0}^{[v]} \mid \tau = \tau_v, \boldsymbol{\pi}, \mathbf{T}) &= \frac{\boldsymbol{\pi}\exp(\mathbf{T}\tau_v)t_i}{\boldsymbol{\pi}\exp(\boldsymbol{T}\tau_v)\mathbf{t}}.
\end{aligned}$$

For $i \in E^*$ and $j \in E \setminus \{i\}$.

M-step: In the M-step, the likelihood from 5.16 is maximized using the conditional expectations found in the E-step, which yields new estimates of $(\boldsymbol{\pi}, \mathbf{T})$. The new estimates are computed using the formulas:

$$\widehat{\pi}_i = \frac{\mathcal{B}_i}{n} \qquad \widehat{T_{ij}} = \frac{\mathcal{N}_{ij}}{\mathcal{Z}_i} \qquad \widehat{t}_i = \frac{\mathcal{N}_{i0}}{\mathcal{Z}_i} \qquad \widehat{T}_{ii} = -(\widehat{t}_i + \sum_{j \in E_j} \widehat{T}_{ij}) \qquad i, j = 1, \ldots, p \qquad (5.19)$$

The derivations for these can be found in (Asmussen, Nerman, and Olsson, 1996). We use the algorithm as implemented by (Bladt, Yslas, and Müller, 2023) in the R package `matrixdist`.

Convergence

The likelihood of a step in the EM-algorithm is always equal to or greater than the likelihood of the previous step, but often it will stall for long periods on flat areas with little improvement. We note that the EM-algorithm is not guaranteed to find the global maximum and can find local maxima or even saddle points instead [8]. To address this we adopt the following strategy for robust parameter estimation:

The algorithm is initialized with random parameters $(\boldsymbol{\pi}, \mathbf{T})$, and the algorithm runs for $100,000$ iterations. The estimated parameters $(\boldsymbol{\pi}, \mathbf{T})^{[100000]}$ are used as the initialization for the next $100,000$ iterations. This is repeated until there has been no improvement in the log-likelihood for $100,000$ iterations.

To prevent stagnation, we added another convergence criteria. The model will be reinitialized when there is a very slow increase in log-likelihood (less than 0.00001%) for 5 consecutive loops of $100,000$ iterations, which provides the model with a chance to converge from another starting point. Among the models, we then keep the one with the best log-likelihood.

The number of parameters in a phase-type distribution

When model selecting, models with more parameters often exhibit higher accuracy, but also higher risk of overfitting. Thus, it is important to consider the number of parameters k in the estimated models. The phase-type representation $(\boldsymbol{\pi}, \mathbf{T})$ of dimension p offers $p^2 + p - 1$ values to be estimated. However the minimal representation of any phase-type distribution is just $2p - 1$ parameters, as shown in Lemma 4.2.24 in (Bladt and Nielsen, 2017). The EM-algorithm doesn't guarantee to find the minimum representation of a given model and the problem of finding minimal representations of phase-types is still an open area of research [35]. The algorithm does seem to have a predisposition for finding simple model, for example Coxian structures. One possible explanation for this is that the EM-algorithm preserves estimated zeros [8]. This preservation takes place in the E-step through Equations 5.17, where the conditional probability of an impossible event is preserved to be zero. Thus,

we never get models with $p^2 + p - 1$ parameters and the actual number of parameters is relatively closer to $2p - 1$ parameters.

The number of parameters affects our computation of goodness-of-fit measures. We will use $k = 2p-1$ for the statistical tests conducted in this report, even if the estimated model appears to have more parameters. We note that for any $p \in \mathbb{N}$ it holds that $2p - 1 < p^2 + p$ meaning we choose the least penalizing of the two bounds.

5.2.2 Model selection

Later in this thesis, we estimate instances of the described statistical models on the service life data. In the following section, we outline our procedure for comparing the goodness-of-fit of our models.

As phase-type distributions are dense in the space of distributions defined on $[0, \infty)$, they can approximate any distribution arbitrarily well. This means that for any non-negative random variable X, one can create a sequence of random variables $X_n \sim PH_n(\boldsymbol{\pi}_n, \mathbf{T}_n)$ such that $X_n \xrightarrow{d} X$. Analogously for any distribution function F with finite k-th moment $\mu_F^{(k)}$, there exist a phase-type distribution function F_p with p phases such that

$$F_p(x) \to F(x) \quad \text{as} \quad p \to \infty, \tag{5.20}$$

for all points x where F is continuous [7]. This means that our phase-type models should theoretically be able to fit any given distribution perfectly, making evaluation of goodness-of-fit tricky. We want to strike the balance between computational viability and complexity to see how each added phase adds to the explanatory power of the model. To evaluate our models, we will use goodness-of-fit statistics as well as diagnostics plots.

Visual inspection

We will plot the empirical probability density function (PDF) and cumulative distribution function (CDF) and compare these with our fitted models's PDF and CDF. Futhermore we will use the QQ-plot to investigate the tails of our models. The phase-type distribution has exponential tails meaning that $P(\mid X \mid \geq x) = \mathcal{O}(\mathrm{e}^{-Kx})$ for large x and some constant $K > 0$. Because of this, the phase-type models do not follow the exact behavior of distributions with heavier tails and underestimate the probability of extreme events, but can approximate them through step-functions. If this is the case, it will be apparent from the QQ-plot where the predicted quantiles of the phase-type model will be lower than the observed quantiles. We will keep an eye out for this behavior in the QQ-plots.

Goodness-of-fit criteria

To test whether our service life data might be accurately described by a given model, we will use the χ^2-test. The χ^2-test compares the observed frequency of data points within specified bins to the expected frequency under a given models theoretical distribution, to evaluate if the data and the model are significantly different in distribution. The χ^2-statistic is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency and E_i is the expected frequency in the i-th bin, and n is the number of bins. A lower χ^2 value indicates a better fit. The χ^2-value can be used to test the null-hypothesis:

$$H_0 : \text{The data stems from the specified distribution.}$$

By using the $\chi^2(\mathrm{df})$-distribution with df degrees of freedom we can compute the p-value under the null hypothesis. The degrees of freedom are computed as $\mathrm{df} = n - 1$, where n is the number of bins. Subtracting 1 accounts for the constraint that the sum of the observed frequencies equals the sum of the expected frequencies.

The χ^2-test is very sensitive to the choices of binning. Some general recommendations that we will follow is that at least 10 observations should be expected in each bin and that the bins follow a natural ordering, such as years or decades. These considerations contribute to the validity of the test. However, the choice of bins still introduces subjective bias and care must be taken to ensure that the binning reflects the underlying structure of the data.

The χ^2-test does not account for model complexity, which could promote the selection of more complex distributions and lead to overfitting. As our phase-type models are described by a maximum of $2p-1$ parameters, the linear scaling of parameters suggests the need to introduce parameter-penalized criteria. We will use the Akaike Information Criteria and the Bayesian Information Criteria to account for this.

Comparing models

To compare the model performance between models with different numbers of parameters, we will use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both criteria offer a balance between model fit and complexity, penalizing models with excessive parameters. The AIC is defined as:

$$AIC = 2k - 2\ell, \tag{5.21}$$

where k is the number of parameters in the model and ℓ is the log-likelihood. A lower AIC value indicates a better trade-off between model complexity and fit. Similarly, the BIC is formulated as:

$$BIC = k\ln(n) - 2\ell, \tag{5.22}$$

where n is the number of observations. The BIC applies a stronger penalty for the number of parameters in larger datasets [3].

We furthermore use a likelihood ratio test to pairwise compare the nested models to see if the increased complexity significantly improves the model fit. As long as we compare between phase-type distributions, they are nested models. Letting ℓ_0 denote the log-likelihood of the simpler model and ℓ_1 denote the log-likelihood of the more complex model, the likelihood ratio test statistic (LR) is defined as:

$$LR = -2\left(\ell_0 - \ell_1\right). \tag{5.23}$$

Under the null hypothesis that they are equally good, the LR-statistic follows a chi-squared distribution $LR \sim \chi^2_{\Delta k}$, where Δk is the difference in the number of parameters between the two models. The p-value is computed based on the χ^2-distribution.

Standardized representation of phase-types

As the representation of a phase-type distribution is not unique [32], we standardize the representations for better comparability. We ensure that the initial probability vector $\boldsymbol{\pi}$ is sorted in a decreasing order. Additionally we will organize the representation such that the jump from state i to state $i+1$ will be the most probable jump, when the initial probabilities allow it. An example of this procedure is shown for the following initial probability vector and sub-identity matrix:

$$\boldsymbol{\pi} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}$$

The standardized representation would require us to swap rows 1 and 2 as well as columns 1 and 2:

$$\boldsymbol{\pi} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{T} = \begin{bmatrix} T_{22} & T_{21} & T_{23} \\ T_{12} & T_{11} & T_{13} \\ T_{32} & T_{31} & T_{33} \end{bmatrix}$$

If $T_{23} > T_{21}$ we would repeat the procedure with states 2 and 3 as long as $\pi_2 = \pi_3$. If their initial probabilities are above zero we will prioritize the decreasing order of $\boldsymbol{\pi}$ in the representation. All $\mathbf{T}$-matrices and $\boldsymbol{\pi}$-vectors presented will be standardized using this procedure.

5.3 Service life analysis

In this section we will use the described methodology to estimate service life models of Danish buildings based on the uncensored data in the BBR dataset. The first subsection will model the service life distribution using classical statistical models. The second and third subsection will employ phase-type models and the underlying Markov Chains for the service life modeling. Here we will consider different numbers of phases and different subsets of the data.

We are especially interested in the "Housing"-use-category as it is the most frequent use-category in our dataset. This means it is of clear national-economic interest, which also manifests in it being the most researched use-category in the academic literature. From a data quality point-of-view, "Housing" is also favorable as the use-category is well-defined in the registry and is relatively homogeneous. Furthermore, the use-category has certainly been constructed consistently through time, such that housing buildings of many ages exist in the dataset. For these reasons our modeling will use the "Housing"-use-category as its data.

The EM-algorithm can incorporate uncensored and right-censored data, but not left-censored data. Using right-censored data without left-censored data can lead to biased results, so we find it better to use only the uncensored data. This means that the estimates are made on housing buildings demolished between 2010 and 2024. Conclusions from these models can therefore only be drawn on demolitions in the observation period.

5.3.1 Estimating classical distribution models

We will estimate models for service life of Danish housing buildings demolished 2010-2024 under three non-negative, continuous distributions - a gamma, a lognormal and a Weibull distribution. Due to their limited flexibility, these are likely to be too simple, but will provide initial insights into the modeling and will guide us in the direction of areas of concern. Using a maximum likelihood estimation (MLE) procedure from the R-package `fitdistrplus` [17], we find the following parameter estimates for the three distributions:

	Lognormal	**Gamma**	**Weibull**
μ (Meanlog)	4.46 ± 0.0026	-	-
σ (Sdlog)	0.512 ± 0.0019	-	-
κ (Shape)	-	4.46 ± 0.031	2.17 ± 0.00821
λ (Scale)	-	21.74 ± 0.16	109.51 ± 0.27761

Table 5.1: MLE parameter estimates for the lognormal, gamma, and Weibull distributions with standard errors (computed from the estimate of the Hessian matrix at the maximum likelihood solution)

Dividing the service lives by decade, we find that all bins from $[30, 40[$ to $[270, 280[$ have at least 10 expected observations for all three models. We group together the first three decades to a $[1, 30[$-bin and the last decades to a $[280, \infty)$ -bin, such that all bins have at least 10 expected observations. The goodness-of-fit statistics and criteria for each model is summarized below:

	Lognormal	Gamma	Weibull
χ^2**-statistic**	3876.48	3530.15	6812.40
χ^2 **p-val**	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$
Log-likelihood	-191898	-190484	-191710
AIC	$383,800$	$380,972$	$383,424$
BIC	$383,817$	$380,989$	$383,444$

Table 5.2: Goodness-of-fit statistics and criteria for the lognormal, gamma, and Weibull distributions.

We can reject the null-hypothesis of the χ^2-test on all significance levels above 10^{-16}, indicating that the models are not well-fitted to the data. The gamma model has the smallest χ^2-statistic. The log-likelihood is significantly lower for the gamma model than the other two models, resulting in the lowest AIC and BIC as well. In Figure 5.3 diagnostics plots for these models are presented:

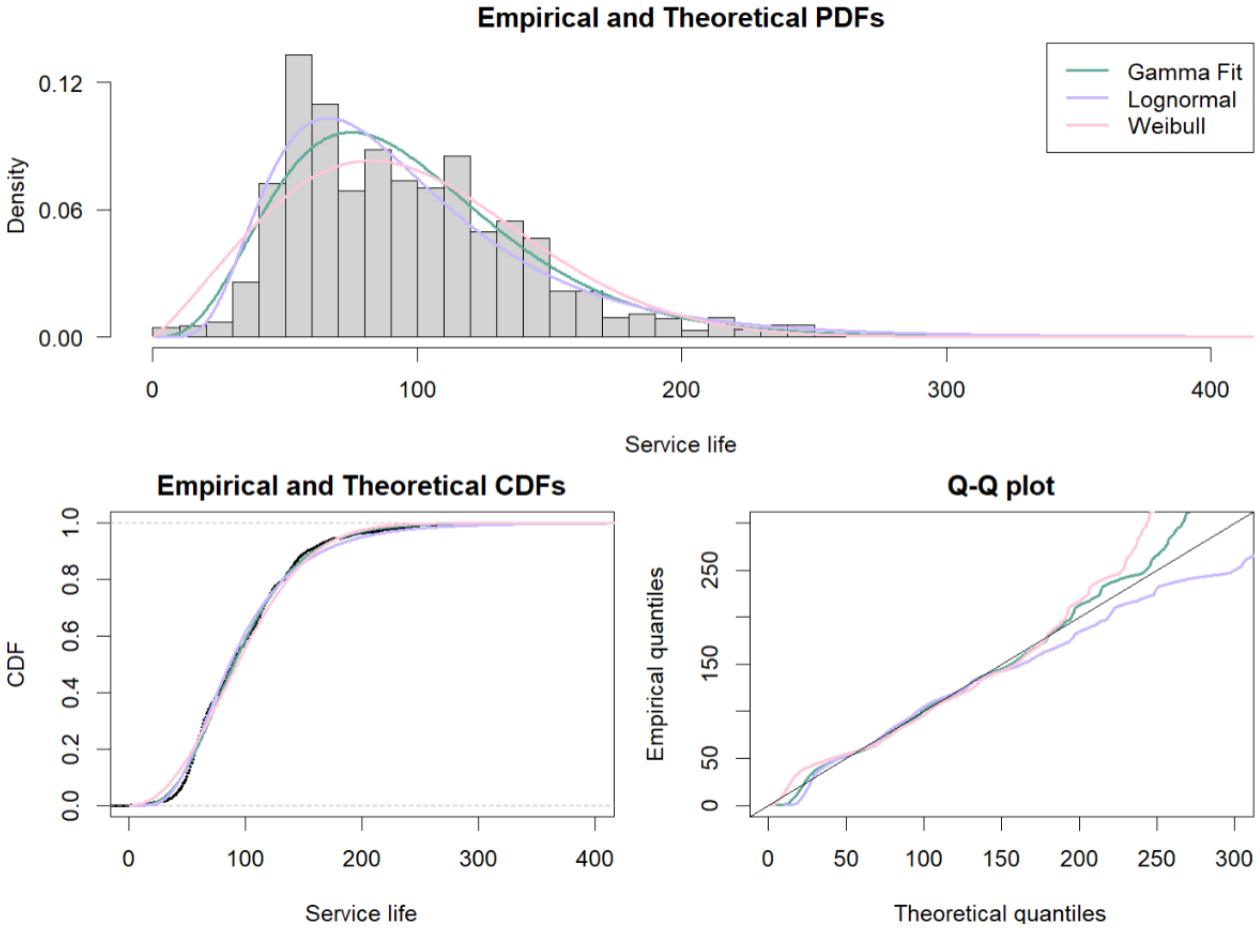


Figure 5.3: Diagnostics plots for the gamma, lognormal and Weibull models

All the models overestimate the density for 20 to 40 years and underestimate the probability density at the steep peak around year 50 to 60. Especially the Weibull model is not well-suited for this part of the data. After the peak however, all the models fare better, with the long and slow decline of density mass being captured well by the models. From the empirical CDF plot, the same issues are evident, especially the overestimation early on. But the CDFs are in general pretty well fitted by the models. The QQ-plots show that all the models follow the data almost exactly between years 50 and

150, indicating again that the fitting issues lie in the short and long service lives. When considering the behavior for the houses above age 200, the models again begin to deviate from the data, with the Weibull model underestimating the densities and the lognormal model overestimating them, with the gamma model only underestimating the densities slightly. Based on the diagnostics plots as well as the goodness-of-fit statistics above, the gamma distribution seems to be the best fit to the data.

As noted, gamma distributions are related to phase-type distributions as gamma distributions with $\kappa \in \mathbb{N}$ can be represented as phase-type distributions. Thus, the use of phase-type distributions emerges as a natural next step.

5.3.2 Estimating phase-type models

Using the uncensored service lives of Danish housing buildings, we employ the EM-algorithm from the R-package `matrixdist` implemented by (Bladt, Yslas, and Müller, 2023) to obtain a model for phase-type distributions with general structure for $p = 4, 6, 8, 10$ phases. We follow the methodology described in Section 5.2 for the parameter initialization, convergence criteria and parameter standardization. The standardized parameters are shown in Appendix A.4. The models are assessed visually using the same diagnostics plots as for the simple distributions in figure 5.3.

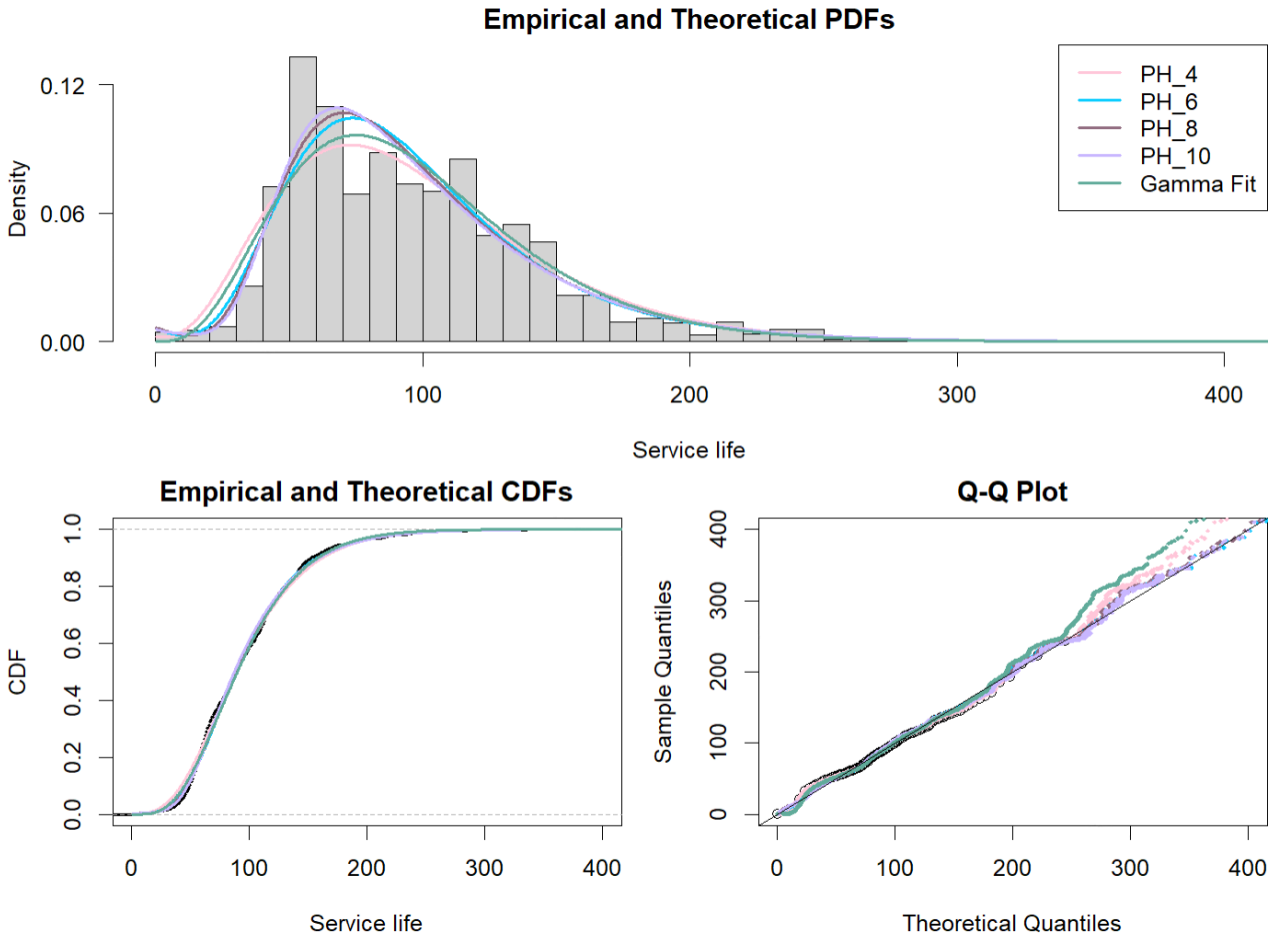


Figure 5.4: Diagnostics plots for phase-type models for $p = 4, 6, 8, 10$

The PH_4 model overestimates the density between 10 and 40 years and underestimates the density for 50 to 80 years. The distribution lacks steepness. In the tail, it underestimates the probabilities, though not as much as the gamma-distribution.

The PH_6-model has a higher probability density for the service lives of 50 to 80 years and thus reflects the data better in this interval. It also captures the very low probability density for the first 20 years better than the PH_4 model. The QQ-plot follows the empirical quantiles almost exactly.

As the number of phases increases to 8 and 10, the peak of the PDF moves to the left, getting closer to matching the actual peak of the distribution. For PH_8 and PH_{10} a more complicated distribution is starting to emerge, with an extra infliction point around the $[10-20[$-year bin. This is a bit surprising, as it suggests some chance of very short service lives.

One concern is that the sharp decline from 70 to 100 years is still not captured in the models. This is especially clear in the PDF plot, but can also be seen in the slight deviation in the CDF plot. This calls for even more phases in the model to properly fit the finer nuances of the data. However, it is important that the fits do not continue to increase the estimated probability density for the bin $[0-10[$-year bin.

The models are clearly improvements from the gamma-model. The improvement is clearest in the QQ-plot, where the PH_6, PH_8 and PH_{10}-models all continue to stay in line with the sample quantiles for the whole time period and especially matching much better than the Gamma-model from 300 to 400 years.

The goodness-of-fit statistics and criteria for the fits are summarized below. We found again that binning the data by decade meant that all bins until 280 years had more than 10 expected observations.

	$\mathbf{PH_4}$	$\mathbf{PH_6}$	$\mathbf{PH_8}$	$\mathbf{PH_{10}}$
χ^2**-statistic**	9883.0	3612.6	2879.2	2501.6
χ^2 **p-val**	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$
Log-likelihood	$-190,341$	$-189,702$	$-189,546$	$-189,435$
AIC	$380,696$	$379,426$	$379,122$	$378,908$
BIC	$380,756$	$379,520$	$379,250$	$379,069$

Table 5.3: Goodness-of-fit statistics and criteria for the phase-type fits distributions.

There is a steady decrease in the AIC and BIC test statistics as we increase the number of phases, but all distributions can still be rejected in the χ^2-test using decades as bins. The BIC and AIC are presented for each model below:

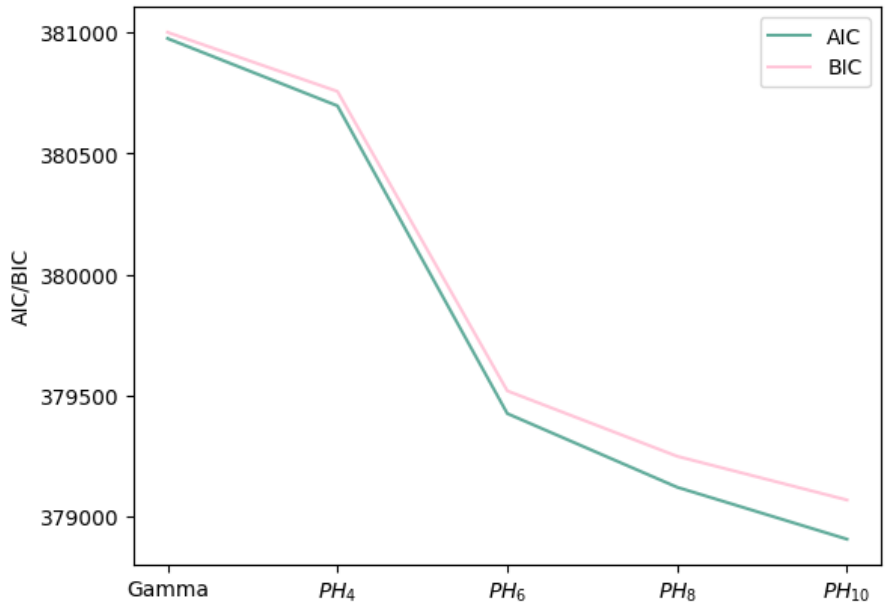


Figure 5.5: The computed goodness of fit-criteria AIC and BIC for each estimated phase-type model and for the gamma-model (utmost to the left)

The criteria decrease from the gamma model to the PH_4-model, but most steeply from the PH_4-model to the PH_6-model. The AIC and BIC is declining for each model, but the rate of decline slows down after PH_6.

For our nested models we use the likelihood ratio test to see if each added phase improves the fit significantly. We compute the LR-statistic and, using the $\chi^2(4)$-distribution, find p-values lower than 10^{-16} for all two-by-two comparisons, showing that the model does improve with each added phase:

			LR	**p-value**
$\mathbf{PH}_4$	vs	$\mathbf{PH}_6$	1278	$< 10^{-16}$
$\mathbf{PH}_6$	vs	$\mathbf{PH}_8$	312	$< 10^{-16}$
$\mathbf{PH}_8$	vs	$\mathbf{PH}_{10}$	222	$< 10^{-16}$

Table 5.4: Likelihood ratio testing phase-type models

5.3.3 Model-based Markov jump processes

Given the estimated sub-intensity matrices $\widehat{\boldsymbol{T_4}}$, $\widehat{\boldsymbol{T_6}}$ and $\widehat{\boldsymbol{T_8}}$ and $\widehat{\boldsymbol{T_{10}}}$ and the estimated initial distribution vectors $\widehat{\boldsymbol{\pi_4}}$, $\widehat{\boldsymbol{\pi_6}}$ and $\widehat{\boldsymbol{\pi_8}}$ and $\widehat{\boldsymbol{\pi_{10}}}$ we can uncover underlying Markov jump processes. As stated, the phase-type representation is not unique, which means that the presented Markov chains are also not unique. Other patterns might emerge for other representations of the same distribution. However, in the following we will interpret the Markov chains as representations of possible decay processes. Even though other representations might exist, we find that the representations presented provide interesting insights into the paths buildings might take before demolition.

The intensity matrix Λ gives us the rates with which the Markov Chain moves from state i to state j given that it is in state i. Equation 5.6 yields the average time a process sojourns in a state before jumping, which is shown for each state. For our PH_4 model, the underlying Markov chain is presented in Figure 5.6

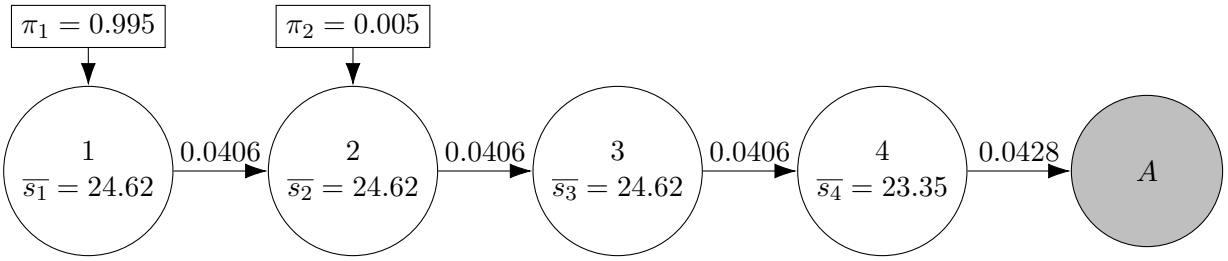


Figure 5.6: A Markov chain for our PH_4 model with "A" denoting the absorbing state

The expected service life of the chain is found using Theorem 5.1.4 and we find that it preserves the sample mean of 96.83 years with a standard deviation of 48.8 years. The standard deviation of the process is slightly higher than for the sample (which has a standard deviation of 46.85). The chain starts in state 1 with a probability of 0.995 and goes through the acyclic process $\{1, 2, 3, 4, A\}$ with varying sojourn times in each state. There is a probability of 0.005 for the chain to start in state 2. From this state and onwards it will follow the same process as the other chains. The service life for these buildings starting in state 2 will then on average be 24.62 years shorter. These buildings could be subject to poor construction or potentially be less aesthetically pleasing, leading to a service life consisting of fewer stages. Starting directly in stage 2 could also be interpreted as "skipping" the demolition-free period due to mortgage. For this Markov chain the period would on average be 24.62 years, but only a very small proportion "skips" it.

For our PH_6-model we obtain the following Markov chain:

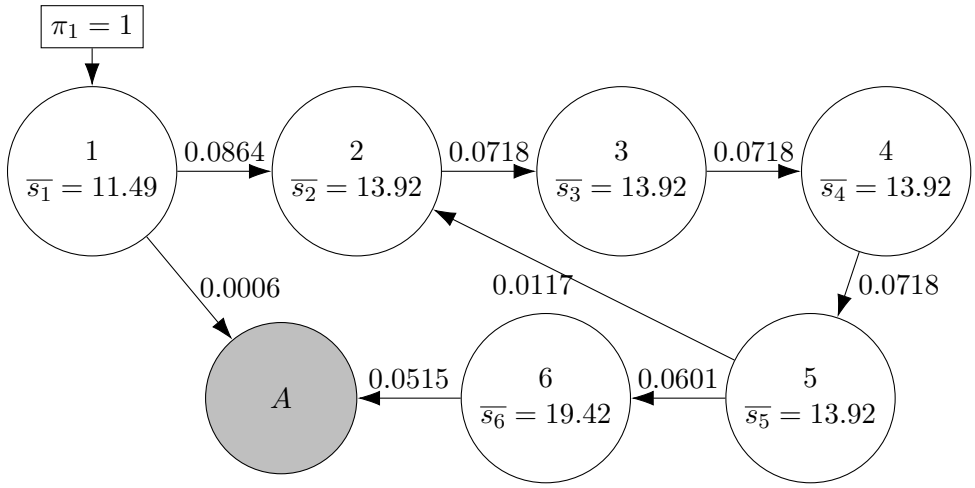


Figure 5.7: A Markov chain for our PH_6 model with "A" denoting the absorbing state

The chain starts in state 1 with a probability of 1. The chain has an expected duration of 96.83 years with a standard deviation of 46.84. The mean service life is unchanged from the PH_4-model, but the variance is slightly smaller. A few buildings will jump directly from state 1 to absorption with a rate of 0.0006. The vast majority will enter into the cycle $\{2, 3, 4, 5\}$. From 5, buildings can jump to either 6 or back to 2. 1 in 7 buildings will jump back to 2 and take another cycle $\{2, 3, 4, 5\}$ while most buildings will jump to state 6. State 6 is the state with the longest expected duration before a certain jump to absorption.

The flow of this Markov chain highlights different paths buildings can take. The buildings jumping directly from state 1 to absorption could be interpreted as construction errors and account for 0.68% of buildings.

One possible interpretation of the estimated underlying Markov chain is that it reflects the act of

reconstructing a house. From Table 4.4 we find that 36.7% of all housing buildings are reconstructed, and the average time before reconstruction is 62.35.

For this Markov chain the chance of entering the "reconstruction cycle" is 16.3%, which is quite a bit lower than the proportion of houses being reconstructed. This could be a sign that not all reconstructions prolong the service life of the buildings, and are thereby not present in the Markov chain. Adding together the mean sojourn time of each state visited before the option of reconstruction results in 67.17 years, which is relatively close to the average 62.35 years before reconstructing a house.

For our PH_8-model we obtain the following Markov chain:

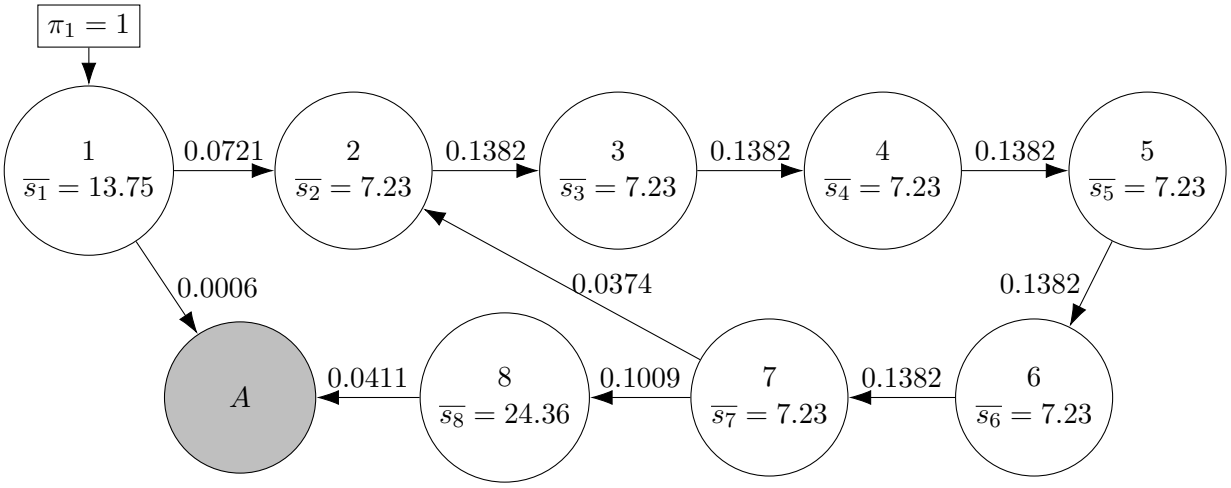


Figure 5.8: A Markov Chain for our PH_8 model with "A" denoting the absorbing state

For the PH_8 model, the expected service life is still 96.83 years. The standard deviation is 47.05 years, which is slightly higher than for the PH_6-model.

The process initializes in state 1 with probability 1. A few of the buildings will jump directly to absorption after an average of 13.75 years, with the same rate as in the PH_6-chain. The rest enters the cycle $\{2, 3, 4, 5, 6, 7\}$ where each state has a mean sojourn time before absorption of 7.23 years. From state 7 buildings can either jump to state 8 which leads to absorption or take another cycle. The probability of jumping to state 8 is 73% while the probability of taking another cycle $\{2, 3, 4, 5, 6, 7\}$ is 27%.

The cycle could be interpreted as reconstructing a building, and thereby prolonging its service life. The probability of entering the cycle in the Markov chain is 27%, which is lower than the 36.7% chance of being reconstructed seen in the data. The average time before the option of reconstruction is possible is 57.13 years, which is relatively close to the average time before reconstruction of 62.35 years. This further supports our hypothesis that the cycle resembles a reconstruction

For the model PH_{10}, we get the following chain:

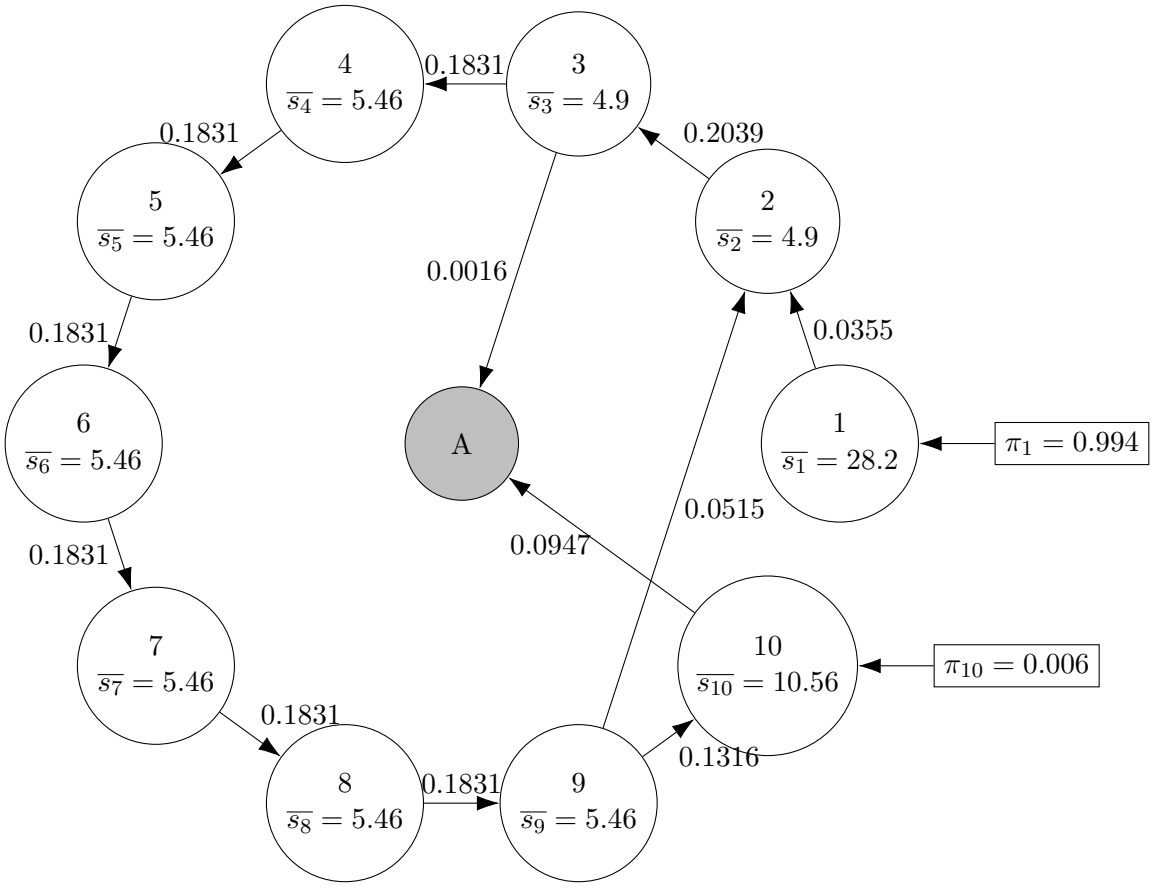


Figure 5.9: A Markov Chain for our PH_{10} model with "A" denoting the absorbing state

The standard deviation for this process is 47.39 years. The PH_{10} model initializes in state 1 with probability of 0.994. From here it begins the cycle $\{2, 3, \ldots, 8, 9\}$ with a very slight chance (0.86%) of jumping directly into absorption from state 3. From state 9 the building can either enter another cycle with a 28% chance or go to state 10 with a 72% chance. From state 10, the building is sure to be absorbed after spending an average 10.56 years in this state.

A building can also initialize in state 10 with a probability of 0.006. This provides the option of a building being demolished almost immediately after construction. This could potentially be due to construction error or it could be that some buildings are simply constructed for very short-term use.

Once again interpreting the cycle as a "reconstruction cycle" yields a 28% chance of reconstruction which is lower than the 36.7% from Table 4.4. On average this chain spends 70.76 years before the possibility of reconstruction, which is slightly higher than the actual reconstruction age of 62.35 years. This could be because the reconstructions that actually prolong the service lives happen later than reconstructions done for other reasons. The higher number of phases allows for more flexibility in the "path" a building can take through its service life, resulting in a better fit as showed on Figure 5.4. The figure showed how the more advanced models started to exhibit a slight "uptick" in density for the very youngest buildings. This is reflected in the Markov chains, where buildings can jump from the initial state directly to the absorption state with a small rate. For the PH_{10} model this accounts for 0.6% of the buildings initializing in state 10 and then jumping to absorption after an average of 10.56 years. For the PH_6 and PH_8-models it accounts for 0.69% and 0.8% respectively after similar durations.

We found that the model for PH_{10} fits our data the best. To see how the process could look for a given

building, following the underlying Markov chain for the PH_{10} model, we simulated three realizations, one acyclic path, one cyclic path and one path initializing in state 10.

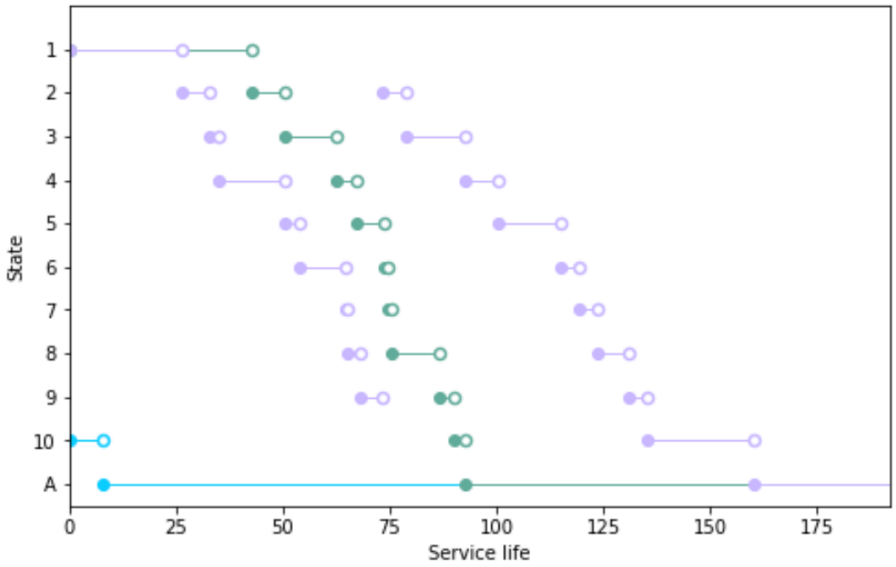


Figure 5.10: Three realizations of the underlying Markov chain for PH_{10} showing the jumps between different states over time.

The green realization on Figure 5.10 follows an acyclic process taking the straight path from state 1 to state 10 before being absorbed. Its longest sojourn time is in state 1 where it stays almost 50 years before shortly visiting the rest of the states, and then being absorbed. The service life of this building is approximately 90 years. This corresponds well with the average service life of a house, at 96 years. The process can be interpreted as a house being well taken care of for 50 years after which the decay process sets in.

The purple realization follows a cyclic process going through stages 1-9 before jumping back to state 2. The process then follows the same cycle of stages 2-9 all the way to state 10 before being absorbed. It spends the longest time in stage 1 and 10, and has a service life of approximately 160 years. The jump from state 9 to state 2 can be interpreted as a reconstruction, which "restarts" the decay process.

The blue realization follows a short acyclic process, starting in state 10. Here it spends approximately 8 years before being absorbed, resulting in a service life of 8 years. This realization is relatively unlikely as starting in state 10 only happens to 0.6% of the buildings. It is shown nonetheless to highlight the many different paths through the jump process. This path can be interpreted as being a very short-lived building.

5.3.4 Phase-type models based on building use-category

Up until now, the modeling has been focused on the "Housing" use-category. As demonstrated on Figure 4.4, the decay process differs vastly over the use-categories - thus it makes sense to model service life models for each building use-category separately and not extend conclusions from "Housing" onto other use-categories. Phase-type models are fitted for each building use-category and presented in Figure 5.11 :

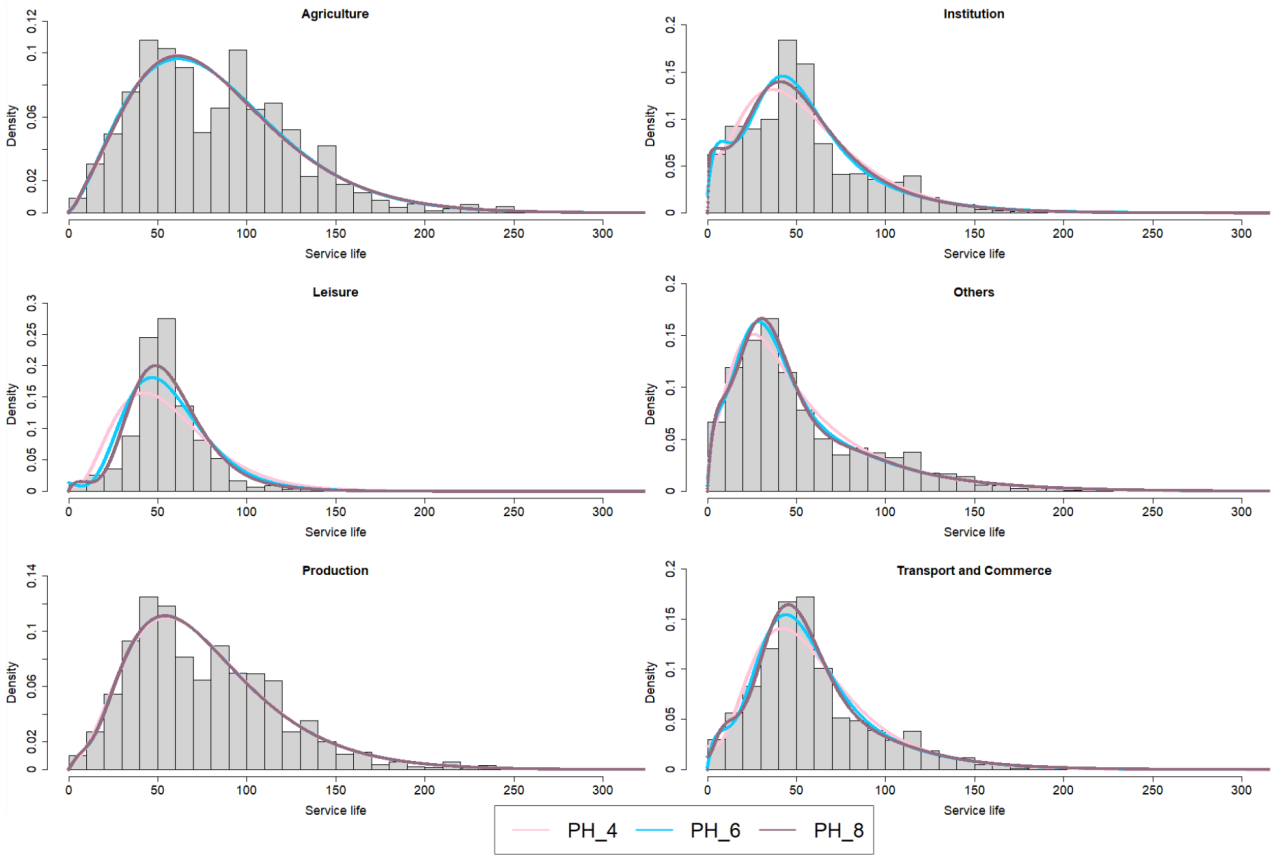


Figure 5.11: Service life histograms for each building use-category with the probability density functions of the estimated phase-type models overlaid for $p = 4, 6, 8$.

In relation to Figure 4.4 we noted that some use-categories varied a lot within use-subcategories and some exhibited bimodality. When modeling on these use-categories, we do see that the phase-type models (with a relatively small number of phases) struggle to catch these more complex distributions.

For "Agriculture" and "Production" the models fail to capture the observed multi-modality of the distributions, indicating that we would need more phases to accommodate these more heterogeneous use-categories.

For the use-category "Institution" the more complex PH_8 model is able to captue some of the plateau-like structure in the initial years. None of the models succeed in capturing the peek of the distribution completely.

It is clear for the distribution of "Leisure", that the phase-type fit isn't able to capture the peak of the distribution. As noted in relation to Figure 4.4, this building use-category has not been popular for long enough to have a consistent service life distribution. The sharp peak with a light tail is not well-captured and might also not be the actual distribution for the "Leisure"-use-category.

The distributions of the use-categories "Others" and "Transport and Commerce" are modeled fairly well by the phase-type distribution with 8 phases. While the model misses the second peak in the distribution for "Transport and Commerce", the first peak is less pronounced than in other use-categories, allowing a less steep phase-type curve to approximate the second peak anyways.

These models show that phase-type models have potential use-cases across other building use-categories than "Housing". Using a very big number of phases could be a way of improving these models and

potentially even incorporating multiple use-categories into the same model.

This concludes our service life modeling. The work demonstrates that phase-type distributions are useful in modeling the service lives of Danish houses, but also point to the need for more complex models. The estimated Markov chains showed interesting decay processes, giving insights into the decay processes of Danish houses. The modeling took place only on demolished houses 2010-2024. While right-censored data is integrable within the EM-algorithm, left-censored data is not yet. Introducing only right-censored data would induce unnecessary bias into the modeling. We will move on to studying the survival curves for the Danish houses, as there exists algorithms for estimating survival curves with both right- and left-censored data.

6 Survival analysis

6.1 Theoretical background on survival analysis

Survival analysis refers to a collection of methods designed to estimate survival probabilities. In the following section we will outline how both censored and uncensored data can be used within survival analysis and present algorithms for survival curve estimation. The theoretical foundation for this section is found in (Klein and Moeschberger, 2003).

Keeping with notation from Chapter 5, we consider each data point i to be a realization of a random variable of service life data $\mathcal{X}_i$. The $\mathcal{X}_i$'s are assumed to be independent and follow the same distribution with a density function $f(x)$ and a survival function $S(x)$. The survival function denotes the probability of a building still standing after an age $x \in \mathbb{N}$.

$$S(x) = P(\mathcal{X}_i > x). \tag{6.1}$$

$S(x)$ is a continuous, strictly decreasing function. When considering survival function estimation, we consider discrete time points when events can occur. Then $S(x)$ can be expressed as a recursive series of conditional probabilities of living past age x:

$$\begin{aligned} S(x) = P(\mathcal{X}_i > x) &= P(\mathcal{X}_i > x \mid \mathcal{X}_i > x-1) \cdot P(\mathcal{X}_i > x-1) \\ &= P(\mathcal{X}_i > x \mid \mathcal{X}_i > x-1) \cdot ... \cdot P(\mathcal{X}_i > 1 | \mathcal{X}_i > 0) \cdot P(\mathcal{X}_i > 0). \end{aligned} \tag{6.2}$$

Where the assumption $P(\mathcal{X}_i > 0) = 1$ leads to:

$$S(x) = \prod_{k=1}^{x} P(\mathcal{X}_i > k \mid \mathcal{X}_i > k-1) \tag{6.3}$$

It is this function we wish to estimate, but in a scenario where we don't have complete information about every $\mathcal{X}_i$. Specifically, we recall from Section 3.1 that:

$$\mathcal{T}_i = \begin{cases} \mathcal{L}_i & \mathcal{D}_i \leq 2010 \text{ (Left-censored)} \\ \mathcal{X}_i & 2010 \leq \mathcal{D}_i \leq 2024 \text{ (Uncensored)} \\ \mathcal{R}_i & \mathcal{D}_i > 2024 \text{ (Right-censored)} \end{cases}$$

We have access to right-censored and uncensored observations through our BBR-dataset. The Kaplan-Meier survival function is a survival curve estimator using only right- and uncensored observations.

6.1.1 Kaplan-Meier survival function estimator

The simplest estimator of the survival curve $S(x)$ is the Kaplan-Meier estimate. As it only accounts for uncensored and right-censored observations, it will not give representative results for our data, but it can provide an initial idea of how survival curves work and are interpreted. Furthermore, an initial survival curve estimate is needed for the Turnbull algorithm, which is introduced afterwards.

The Kaplan-Meier estimator is a step function with mass at the ages $x \in \mathbb{N}$. We denote the number of demolitions at age x as d_x and the number of buildings still standing at age x as y_x. We can estimate $P(\mathcal{X}_i > x | \mathcal{X}_i > x-1)$ as the fraction of buildings who are standing at age x but are not demolished at this age:

$$P(\mathcal{X}_i > x | \mathcal{X}_i > x-1) = \frac{y_x - d_x}{y_x} = 1 - \frac{d_x}{y_x} \tag{6.4}$$

Inserting this estimate into Equation 6.3 yields the Kaplan-Meier estimator:

$$\widehat{S}(x) = \prod_{x_k \leq x} \left(1 - \frac{d_{x_k}}{y_{x_k}}\right), \tag{6.5}$$

where $x_k \in \mathbb{N}, k \in \mathbb{N}$, with $x_1 < x_2 < \cdots < x_n$.

The confidence interval for Kaplan-Meier estimates are derived in (Sawyer, 2003):

$$\hat{S}(x) \pm z_{\alpha/2}\sqrt{\text{Var}\left(\hat{S}(x)\right)}$$

where

$$\text{Var}\left(\hat{S}(x)\right) = \hat{S}(x)^2 \sum_{x_k \leq x} \frac{d_{x_k}}{y_{x_k}(y_{x_k} - d_{x_k})}.$$

Right-censored data can be incorporated into the Kaplan-Meier estimator using the assumption of noninformative censoring. This means that we include the right-censored observations in y_x if they have not yet been censored at age x but we do not include them in the number of demolitions d_x. This increases the denominator, while not changing the numerator. Thus, right-censored observations will inevitably increase $\widehat{S}(x)$.

The Kaplan-Meier estimate doesn't accommodate left-censored observations. If the actual observational set-up also includes left-censoring, the inclusion of right-censored observations leads to a biased survival function. Therefore, methods for double-censored data can provide a more accurate estimator of the survival function.

Currently, we do not have left-censored observations in our dataset, but methods for constructing such data are described and implemented in Section 6.2. This will provide us with double-censored data, which can be used in Turnbull's algorithm.

6.1.2 Turnbull's algorithm for double-censored data

To estimate the survival curve for double-censored data we will use Turnbull's Algorithm. The algorithm is presented in (Turnbull, 1976) and extends the Kaplan-Meier estimator to incorporate left-censored data.

The algorithm assumes a number of discrete ages $x = 1, 2, \ldots, n$ at which each building is observed. At each age x, we observe three types of data: d_x, the number of demolitions occurring at age x; r_x, the number of observations right-censored at age x; and l_x, the number of observations left-censored at age x.

To estimate the survival function, we iteratively update the estimate based on an initial survival function and compute the expected number of demolitions at each age before x. This iterative procedure continues until the survival function stabilizes. The steps of the algorithm are as follows:

- **Step 0:** Begin with an initial estimate of the survival function, $S^{(0)}(x)$, at each time point x. Any valid estimate can be used, but we will use the Kaplan-Meier estimate obtained by ignoring left-censored data.
- **Step 1:** Estimate the probability of demolition happening within a time-interval $[x - j - 1; x]$ given that the demolition has not occurred at age x, meaning $p_{xj} = P(x-j-1 \leq \mathcal{X} \leq x \mid \mathcal{X} \leq x)$. For this, the current estimate of the survival curve is used:

  $$\widehat{p}_{xj} = \frac{S^{(K)}(x - j - 1) - S^{(K)}(x - j)}{1 - S^{(K)}(x)} \quad \text{for } 0 \leq j \leq x.$$

  With $S^{(K)}$ denoting the k'th estimate of the survival curve.

- **Step 2:** Using the probabilities $\hat{p}_{xj}$ from the previous step, estimate the number of demolitions at each time x by

$$\widehat{d}_x = d_x + \sum_{j=1}^{n} l_j \widehat{p}_{xj}.$$

- **Step 3:** Recompute the Kaplan-Meier estimate using the updated demolition counts $\widehat{d}_x$ and the right-censored observations r_x, ignoring left-censored data.
- **Step 4:** Check for convergence of $S(x)$. If converged, stop the procedure; otherwise, repeat the steps **1-3**.

We implement the algorithm in Python:

Algorithm 1 Turnbull's algorithm for double-censored data

1: **Input:** d, r, l
2: **Initialize:** $S^{(0)}(x)$,$K = 0$
3: **while** not converged **do**
4: **Step 1:** Update probability $p_{ij} = \frac{S^{(K)}(x-j-1)-S^{(K)}(x-j)}{1-S^{(K)}(x)}$, for $j \leq x$.
5: **Step 2:** Estimate the number of demolitions: $\hat{d}_x = d_x + \sum_{j=1}^{n} l_j \widehat{p}_{xj}$.
6: **Step 3:** Recompute Kaplan-Meier survival estimate $S_K = Kaplan\_Meier(\hat{d}_x, r_x)$
7: **Step 4:** $k = k + 1$
8: **end while**
9: **Output:** $S(x)$.

6.2 Construction of left-censored data

The BBR-dataset is lacking information about left-censored observations, which are needed to use Turnbull's algorithm. To construct a dataset with left-censored observations, we will use three different approaches:

1. **Stochastic simulation:**A simulated model assuming independence of service life and construction year and using the estimated phase-type PDF
2. **Backcasting by demolition rate:** A model assuming a constant yearly demolition rate γ and a geometric distribution of service life
3. **Backcasting by phase-type assumption**: A model using the phase-type survival curve estimate from our uncensored data

All models focus on housing buildings constructed between 1917 and 2009.

6.2.1 Stochastic simulation

We begin by fabricating a simulated dataset which will be used to test the correctness of Turnbull's algorithm as we will know the true survival curve of it. We use the dataset BYGV05A from Statistics Denmark on housing buildings constructed from 1917 to 2024. The density distribution is shown in Figure 4.10 and will be shown again in Figure 6.1. The density, denoted $g(c)$, is shown again on Figure 6.1. As $\mathcal{X}$ is assumed to be phase-type distributed, $f(x)$ is the density function for a phase-type variable. We assume that $\mathcal{X}$ is independent of $\mathcal{C}$, which allows us to use the same service life distribution on all buildings. We use the estimates for $(\boldsymbol{\pi}, \mathbf{T})$ from the PH_{10}-model from Section 5.3.2. Then we can simulate realizations of $\mathcal{D}$ by drawing independent values of $\mathcal{X}$ and $\mathcal{C}$ using Equation 3.1:

$$\mathcal{D} = \mathcal{C} + \mathcal{X}.$$

To create a dataset that accurately reflects the true population, we need to determine the number of buildings to simulate. The BBR dataset includes only buildings demolished after 2010. The theoretical probability of this event is computed as follows:

$$P(\mathcal{D} \geq 2010) = \sum_{c=1917}^{2024} P(\mathcal{D} \geq 2010 \mid \mathcal{C} = c) g(c), \tag{6.6}$$

where $P(\mathcal{D} \geq 2010 \mid \mathcal{C} = c) = 1 - F(2010 - c)$ is the survival probability for buildings constructed in year c and $F(2010 - c) = 1$ if $c > 2010$. Using this formula, we find that $P(\mathcal{D} \geq 2010) = 0.9046$.

This implies that the BBR dataset represents 90.46% of housing buildings constructed since 1917, corresponding to $1,679,629$ houses. The whole population of Danish houses constructed between 1917 and 2010 is then estimated to consist of:

$$\frac{1,679,629 \text{ buildings}}{0.9046} \approx 1,856,764 \text{ buildings}. \tag{6.7}$$

Therefore, we simulate $1,856,764$ realizations of construction year and service life and compute the corresponding demolition year. We check that the simulated distribution of construction years matches the distribution $g(c)$ and that the simulated service lives matches the distribution $f(x)$:

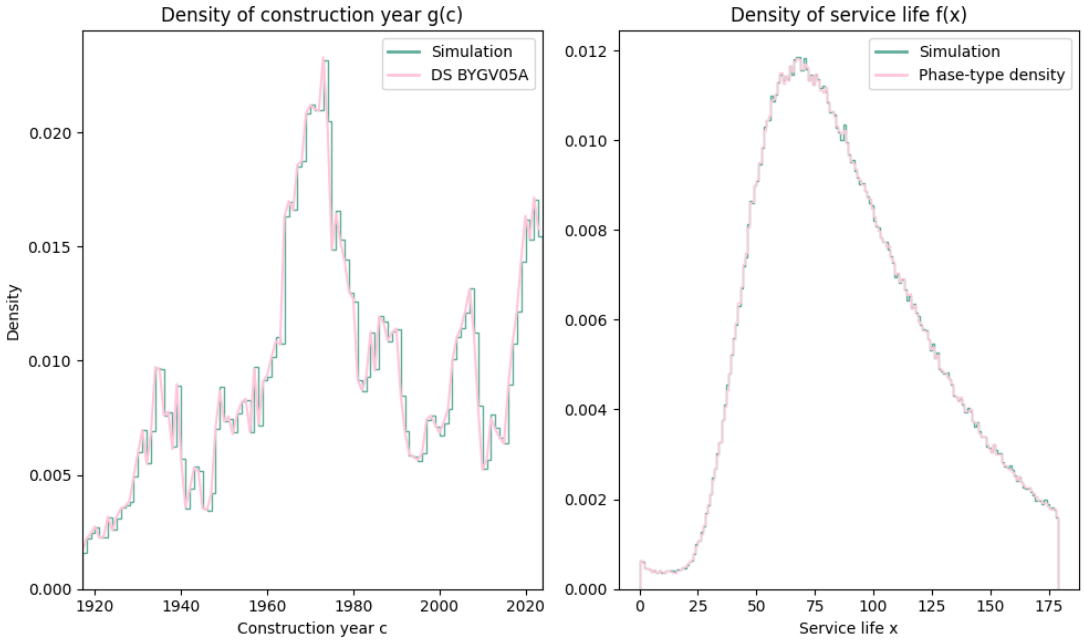


Figure 6.1: Comparison of the overlap between simulation and the data sources. To the left, construction year density in the simulation and construction year density in the BYGV05A database. To the right, the service life density in the simulation and for the PH_{10}-model.

This confirms that construction year and service life are sampled correctly from the specified distributions. Using the simulated dataset, we compute the Monte Carlo estimate of being demolished before and after 2010:

$$P(\mathcal{D} < 2010) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i + c_i < 2010) = 0.0953 [95\% \text{ CI: } 0.0948; 0.0957]$$

$$P(\mathcal{D} \geq 2010) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i + c_i \geq 2010) = 0.9047 [95\% \text{ CI: } 0.9043; 0.9052].$$

With the theoretical probability 0.9046 falling inside of the confidence interval. We compute the conditional probability $P(\mathcal{D} < 2010 \mid \mathcal{C} = c)$ and compare it with the theoretical survival curve for the phase-type distribution $F(2010 - c)$:

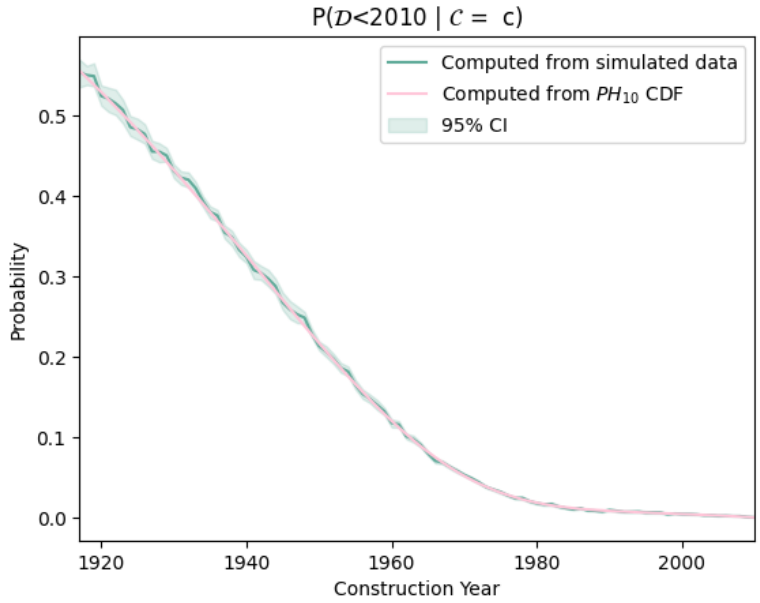


Figure 6.2: Simulated probability of the event "being left-censored" given construction year. Confidence intervals are computed using 1000 bootstrapped resamplings.

The figure shows that half of the houses built in 1917 are still standing in 2010, compared to almost all houses built in 2009 still standing in 2010. Very little demolition takes place among houses from 1980 and later, corresponding to the low density of the PH_{10}-model for $\mathcal{X} < 30$. This provides an initial sanity-check. Another way of seeing if our simulation holds, is to see if it reproduces the density function $g_{\mathcal{C}|\mathcal{D}>2010}(c)$ from our data:

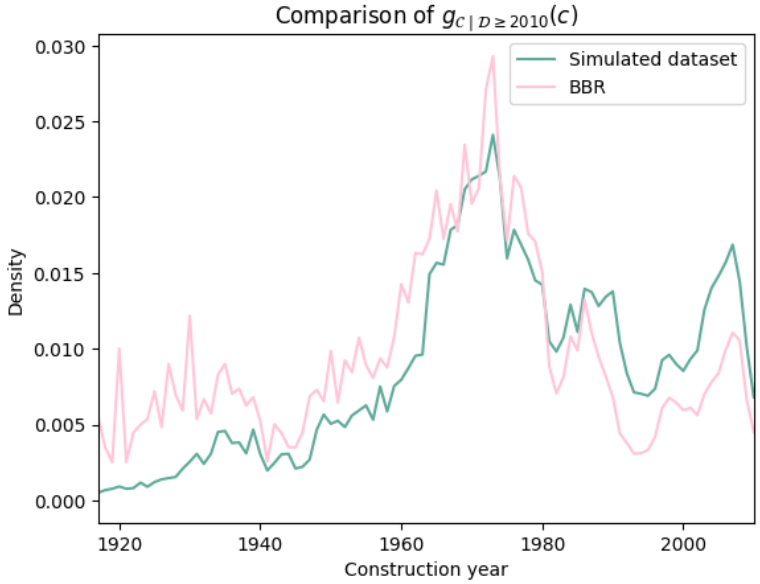


Figure 6.3: Comparison of dataset and simulation. Confidence intervals are computed using 1000 bootstrapped resamplings.

The simulation assigns too little probability mass to the years 1917 - 1940 and too much to 1990 - 2010. This discrepancy likely arises from the drawing of $\mathcal{C}$ from Statistics Denmark's BYGV05A dataset. The inaccuracies in this dataset are documented in Section 4.3.2, where the main conclusion was that BYGV05A underestimates the building construction for 1917-1940 and overestimates the building construction for 1990-2020.

The underestimation for earlier years is furthered by limitations of the phase-type model. The estimated PH_{10} model overestimates the probability $P(80 \leq \mathcal{X} \leq 100)$, as shown in Figure 4.1. For demolitions in 2010, this corresponds to buildings constructed between 1910 and 1930 - the period where the simulation underestimates the standing population.

6.2.2 Backcasting using a constant demolition rate

A different way of fabricating a dataset with left-censored observations is to assume a constant yearly rate of demolition γ. In a Danish context a constant rate of $\gamma = 1\%$ is often used [47] but this figure is disputed within the literature. In (Andersen, 1992) a much lower rate of $\gamma = 0.0015$ is estimated while (Aagaard et al., 2013) suggests $\gamma = 0.003$.

We will estimate the demolition rate γ directly from our data as the proportion of houses standing at the beginning of year j which are demolished in year j. This is

$$\gamma_j = \frac{\text{no. of buildings demolished within year j}}{\text{no. of buildings standing in the beginning of year j}}$$

The demolition is calculated for the period 2011-2023 as we are certain to have full data for these years. The yearly demolition rate for each year is shown in Table 6.1

j	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	2017	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
$\boldsymbol{\gamma_j}$ $\boldsymbol{(10^{-3})}$	15	16	15	16	17	17	16	15	16	15	18	20	16

Table 6.1: Computed yearly demolition rates from BBR data for the years 2011-2023

A linear regression model was fitted to test if γ could be considered constant in the years from 2011 to 2023. The slope coefficient of the model was not statistically significant at the 5% level ($p = 0.112$) with the 95% confidence interval for the slope including zero. Thus, we cannot reject a constant demolition rate γ in the period. The mean demolition rate was computed as $\gamma = 0.00163[\text{CI: } 0.00154; 0.00171]$, matching well with the estimate from (Andersen, 1992).

We backcast the number of buildings constructed in the year c, denoted N_c, using the estimated demolition rate γ and the number of buildings constructed in year c still standing in 2010, denoted $N_{ss,c}$. The service life is geometrically distributed with the constant rate γ.

$$N_c = \frac{N_{ss,c}}{P(D \geq 2010 | \mathcal{C} = c)} = \frac{N_{ss,c}}{(1-\gamma)^{2010-(c)}}.$$

The difference between N_c and $N_{ss,c}$ corresponds to the number of left-censored buildings for a construction year c, with the left-censoring age $\mathcal{L} = 2010 - c$. This yields a dataset with left-censored observations.

6.2.3 Backcasting using phase-type survival curve

A constant yearly demolition rate is a strong assumption. One of the main counterarguments for this assumption is that we do not expect a building to be demolished in its early years. A simple solution, suggested in (Aagaard et al., 2013), is to implement a demolition-free loan period of 20 to 30 years and then use a constant demolition rate for the rest of period.

Another more complex approach, which also takes this into account, is to assume that the service life distribution follows a phase-type distribution. We use our estimated PH_{10}-model. Using the assumption that the service life is independent of construction year, $N_{ss,c}$ is computed:

$$\begin{aligned} N_{ss,c} &= N_c \cdot P(\mathcal{D} > 2010 | \mathcal{C} = c) \\ &= N_c \cdot P(\mathcal{X} + c \geq 2010) \\ &= N_c \cdot P(\mathcal{X} \geq (2010 - c)) \\ &= N_c \cdot S(2010 - c). \end{aligned}$$

By rearranging the equation, the number of buildings constructed in year c can be estimated as:

$$N_c = \frac{N_{ss,c}}{S(2010 - c)}.$$

We can once again compare N_c and $N_{ss,c}$ to get the number of left-censored buildings for a construction year c, and compute the corresponding $\mathcal{L} = 2010 - c$. We thereby obtain a dataset consisting of both left- and right-censored data by assuming a phase-type distributed service life.

6.2.4 Comparison of backcasted data to BYGV05A

The values of N_c obtained from the backcasted datasets are plotted against the constructions pr. year from Statistics Denmark's dataset BYGV05A in the period 1917 to 2010:

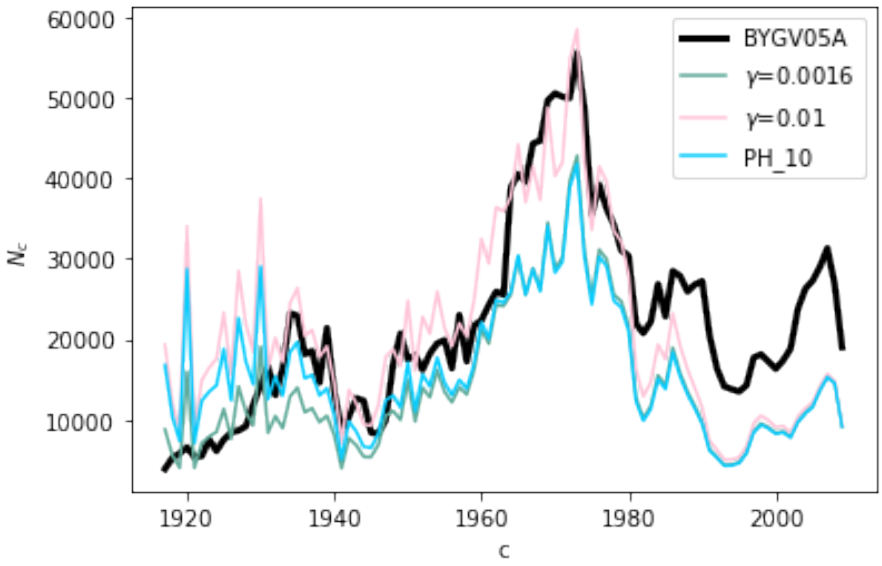


Figure 6.4: Backcasted N_c for different values of γ and based on a phase-type assumption (PH_10) compared to Statistics Denmark's dataset BYGV05A (shown in bold black)

The model with demolition rate $\gamma = 0.16\%$ generally undershoots the estimate from Statistics Denmark (the black graph). The model based on $\gamma = 1\%$ overshoots in the early years and then follows the estimate quite nicely from 1930 till 1980. After that it undershoots like the other models. All models undershoot in the most recent years, where we would expect the data to fit relatively well as almost no demolitions should have occurred and the assumptions of the demolition process therefore have very little effect. This once again emphasizes the finding that the data obtained from BYGV05A is not comparable to the BBR-data. Based on the available data, it is not possible to conclude which dataset is most reasonable to use, so we will continue to work with all three to see how they affect the estimated survival curves.

6.3 Survival analysis

We now use the Kaplan-Meier and Turnbull algorithm on these constructed datasets. The Kaplan-Meier estimator requires only uncensored and right-censored data and thus works with only our BBR-data. The Turnbull estimator requires the datasets with left-censored observations as constructed in Section 6.2 above.

Kaplan-Meier estimation on BBR dataset

When applying the Turnbull estimate, we use the Kaplan-Meier as an initial survival curve. The BBR-data contains $1,679,629$ houses, from which 2.2% are demolished in the observational period. We estimate the Kaplan-Meier survival function and get the green curve below:

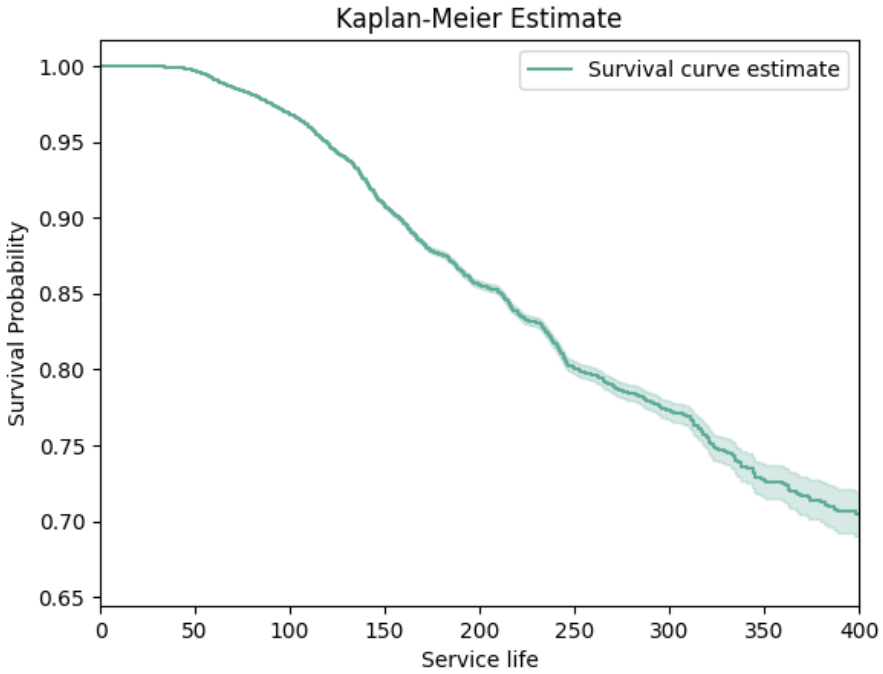


Figure 6.5: Kaplan Meier estimation of the survival curve with the BBR-data (2.2% observed data and 97.8% right-censored data), shown with confidence intervals

We find this survival function to be very unrealistic, as it estimates that 70% of all housing buildings will live more than 400 years. As described earlier, this is typical for a Kaplan-Meier estimator, as it is estimated without left-censored data.

Turnbull estimation on simulated data

We want to improve upon the Kaplan-Meier estimate by also including left-censored buildings. For these buildings we don't know their exact service life, but we know that their service life $\mathcal{X}_i < \mathcal{L}_i$. To test the correctness of Turnbull's algorithm, we use the simulated datasets constructed in Section 6.2.1 with $n = 1,856,764$ buildings.

We aggregate the data by counting the number of observations in each censoring group with an age a:

$$N_{\text{left-censored},a} = \sum_{i=1}^{n} \mathbb{I}(\mathcal{D}_i < 2010) \cdot \mathbb{I}(2010 - \mathcal{C}_i = a)$$
$$N_{\text{uncensored},a} = \sum_{i=1}^{n} \mathbb{I}(2010 \leq \mathcal{D}_i \leq 2024) \cdot \mathbb{I}(\mathcal{D}_i - \mathcal{C}_i = a)$$
$$N_{\text{right-censored},a} = \sum_{i=1}^{n} \mathbb{I}(\mathcal{D}_i > 2024) \cdot \mathbb{I}(2024 - \mathcal{C}_i = a)$$

We get the following simulated data with 1,856,764 simulated observations:

Age	**Left-censored observations**	**Uncensored observations**	**Right-censored observations**
1	11	169	29162
2	23	133	31845
⋮	⋮	⋮	⋮
92	2261	907	4505
93	1671	786	5613
⋮	⋮	⋮	⋮
106	0	38	1418
107	0	24	965
sum	177,492	150,209	1,529,063

Table 6.2: The first and last rows of our simulations as well as two intermediate rows. Logically many "old" observations are left-censored and many "young" observations are right-censored. Due to our sampling of construction year in the interval [1917,2024], the oldest left-censored buildings are 93 years old.

Using Turnbull's Algorithm as outlined in Algorithm 1 with convergence criteria $\mid S^{(K-1)} - S^{(K)} \mid < 0.0001$, we obtain a survival function estimate. For this dataset we have the theoretical survival curve, as it is the survival curve for our estimated PH_{10} for housing. The estimated and theoretical survival curve are shown in the figure below:

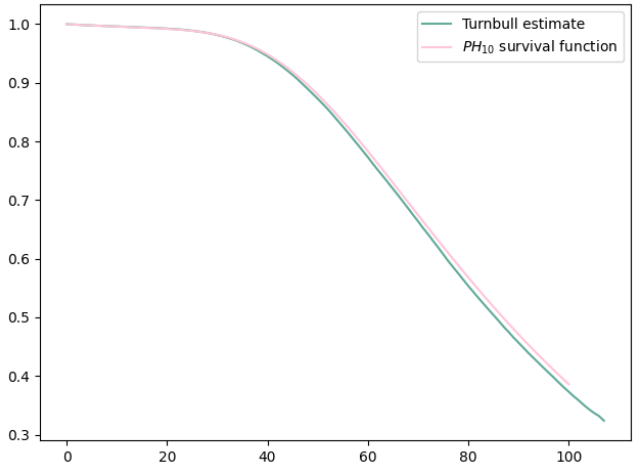


Figure 6.6: Survival curve for PH_{10} and the Turnbull estimate on simulated double-censored data

There is a close overlap between the Turnbull estimate on the double-censored data and the theoretical survival curve. This indicates that the Turnbull-estimator is efficient in reconstructing the survival function. In (Turnbull, 1976) it is shown that the estimator is self-consistent. As we here show its correctness in our experimental setting, we can apply it to our backcasted datasets.

Turnbull estimation on backcasted data

We use the Turnbull algorithm on the two backcasted datasets obtained in Section 6.2.2. We also test the value found in literature of $\gamma = 1\%$.

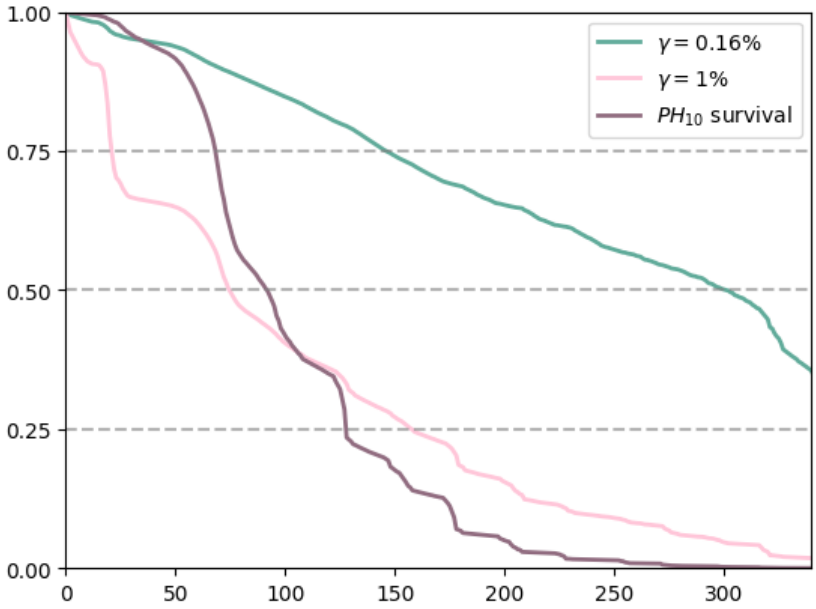


Figure 6.7: Turnbull estimated survival curves for the three different censored datasets. Our estimated demolition rate yields quite high survival probabilities compared to the demolition rate from both the literature and the PH_{10}-model.

The demolition rate 0.16% yields a dataset with relatively little demolition and thus relatively little left-censored data. This results in estimation of relatively large survival probabilities. The two other datasets are more pessimistic in their estimations. When assessing survival probabilities of more than 100 years, the PH_{10}-estimated dataset is the most pessimistic, while the literature demolition rate of 1% is most pessimistic on short-term predictions. The survival curve based on a PH_{10}-demolition rate approximates the S-shape seen in the actual phase-type survival curves, which is expected.

In Table 6.3 the quartiles for each survival curve are presented:

Method	$\mathbf{Q}_{0.25}$	$\mathbf{Q}_{0.5}$	$\mathbf{Q}_{0.75}$
Regression based demolition rate $\gamma = 0.16\%$	146	301	-
Literature value of $\gamma = 1\%$	21	75	158
Phase-type survival curve	68	92	128

Table 6.3: Estimates of the first, second and third quartile for housing buildings made using the backcasted datasets and Turnbull's algorithm

For the demolition rate $\gamma = 0.16\%$ we find that 75% of housing lasts at least 146 years. The median service life is 301 years. For the literature value of 1% demolition rate we find that 75% lasts longer than 21 years, as it begins with a quite big drop in probability. The median service life for this survival curve

is 75 years, after which it has a much slower decline as it approaches the third quartile at 158 years. The PH_{10}-estimated dataset follows the quartiles from the demolished BBR-data closely, reaching the quartiles at 68, 92 and 128 years, compared to the observed quartiles for demolished houses of 61, 89 and 121 years.

We find that the backcasted data using a demolition rate of 0.16% is methodologically strongest, using the fewest assumptions. By using the dataset more closely, it avoids the issue of extrapolating the service lives for buildings 2010-2024 to the rest of the period. However, we find the estimated service lives to be quite high, with a median of 301 years. We imagine that more uncensored data might soften up this estimate, but for now that is not available. We still find that this estimate might reveal trends in coming years, as it is based on real data and on a real demolition rate.

To determine if our estimates are in line with the scientific literature, we will compare the figures presented above with other estimates on building service life.

6.4 Comparing results to literature

In Table 6.4, different studies on Danish demolition rates are presented along with the estimates found in this thesis. The demolition rate estimates vary from 0.091% in (Jensen et al., 2022) to 1% in (Statistics Denmark, 1988):

Study	Estimated γ	Method/Source
Our thesis	0.16%	Regression on demolished housing 2011-2023
(Statistics Denmark, 1988)	1%	Methodology unclear, standard in financial models
(Andersen, 1992)	0.15%	Regression on demolished housing in the 1980s
(Aagaard et al., 2013)	0.3%	Literature review
(Jensen et al., 2022)	0.091%	Demolition data on detached single-family houses 2011-2017

Table 6.4: Comparison of estimated demolition rates of Danish houses

The relatively low estimate from (Jensen et al., 2022) is made only on detached-family houses, which have a longer service life and are believed to have a lower demolition rate.

The demolition rate estimated in this thesis fall within the previous estimates found in other papers and matches especially well with (Andersen, 1992). As Andersen estimates the demolition rate for Danish houses in the 1980s and acquires an almost identical value this shows that our assumption of a constant demolition rate through time is true for the 1980s. Andersen argues that a demolition rate of 1%, as found by (Statistics Denmark, 1988) and used in economic models, is unrealistically high and that the actual demolition rate is likely lower. She demonstrated that only 0.15% of the housing stock was demolished each year in the 1980s. This presented a logical gridlock: comparing the proportion of demolished buildings to the entire housing stock seems intuitively reasonable for estimation of the demolition rate, but using a 0.15% annual demolition rate in a traditional model results in an unrealistically long expected service life (classically, a geometric distribution has been used yielding a median service life of $\frac{-1}{\log_2(0.0015)} = 462$ years). In this context, our method presents a compromise; using backcasting and Turnbull's algorithm, we apply the same demolition rate in a much more real-life, data-driven scenario. This means that we can use the found demolition rate of $\gamma = 0.16\%$ and still keep our median service life estimates at more realistic levels.

Table 6.5 shows various median service life estimates for Danish houses. The median estimates range from 55 years, found in (Østergaard et al., 2018) to 301 years (found in this thesis).

Study	Method	Service life estimate(years)
This thesis	BBR data on demolished houses 2010 - 2024	Median: 89, Mean: 96
This thesis	Turnbull estimate with $\gamma = 0.16\%$	Median: 301
This thesis	Turnbull estimate using PH_{10} survival curve	Median: 92
(Aagaard et al., 2013)	$\gamma = 0.5\%$ with 20 year demolition free loan-period	Mean: 220
(Aagaard et al., 2013)	$\gamma = 1\%$ with 20 year demolition free loan-period	Mean: 120
(Østergaard et al., 2018)	Regression on Danish building stock (2009-2015)	Median: 55, Mean: 67
(Videnscenter, 2023)	Needlemans formula and BYGB12	Median: 196
(Jensen et al., 2022)	BBR data on demolished buildings 2010-2021	Mean: 85

Table 6.5: Comparison of service life estimates on Danish houses

The methodology behind the estimate in (Jensen et al., 2022) corresponds to the methodology of our estimate from the demolished buildings of 96 years. The difference lies in our slightly longer observational period and in (Jensen et al, 2022) only considering houses demolished with the stated purpose of building a new house.

The estimate by (Østergaard et al., 2018) is generally deemed too low. The critiques are based on the dominance of commercially owned buildings in their data, a lack of differentiation between housing types and the regression methodology not adjusting for the construction materials or the era in which the buildings were erected [47].

The estimates by (Aagaard et al., 2013) is a simple geometric model using different common demolition rates. The estimate by (Videnscenter, 2023) seems the most methodologically sound and also aligns with international estimates from comparable countries. Our estimate of a median of 301 years is the closest to these, but still quite a bit higher than the estimates from these papers.

Table 6.6 presents international service life estimates:

Study	Country	Service life estimate (years)	Method
(Rincón, Perez, and Cabeza, 2013)	Spain	Mean: 80	Comparison of dwelling stock census
(Liu et al., 2014)	China	Mean: 34	Hedonic modelling
(Kornmann and Queisser, 2012)	Switzerland	Mean: 180	Comparison of dwelling stock census
(Kornmann and Queisser, 2012)	Switzerland	Median: 289	Needlemans formula
(Bradley and Kohler, 2007)	Germany	Median: 300+	Historical and regional data analysis

Table 6.6: Overview of international building service life estimates

These international estimates highlight the significant geographic variation in service life expectations. The construction practices in Denmark are much more similar to those of Germany and Switzerland

than China and can be used to assess the reliability and accuracy of Danish estimates. The different median and mean estimates in both Danish and international literature are presented in Figure 6.8 below:

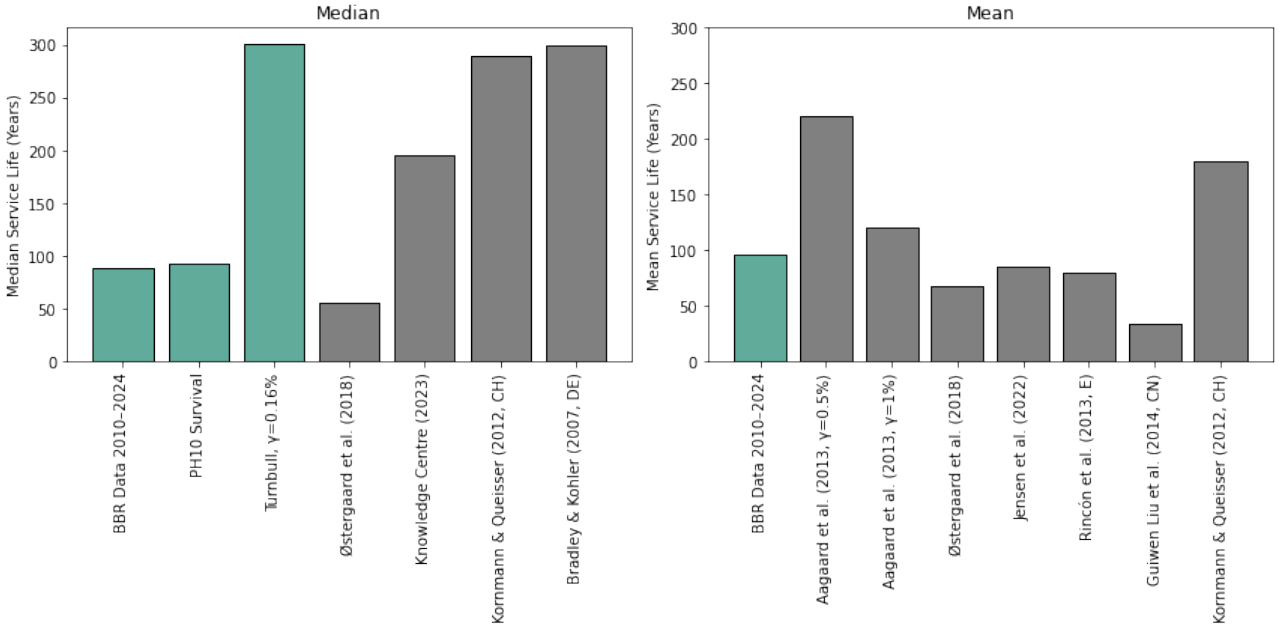


Figure 6.8: Median and mean service life for houses as estimated in our thesis and 8 research papers (4 Danish, 4 international) with our estimates highlighted in green.

Our Turnbull median estimate based on a phase-type assumption aligns with the mean service lives estimated by (Jensen et al, 2022) and (Rincón et al, 2013). These estimates are descriptive of the demolished building mass as it stands today.

We find that our Turnbull median estimate using $\gamma = 0.16\%$ is among the longest estimated service lives, aligning with the German and Swiss median estimates. While we still consider this estimate to be quite high, it strengthens our conclusions that other studies, using different methodologies, estimate similar figures.

Several countries are moving towards longer service life estimates than the traditional 50-60 years used for regulatory LCA. For example, Germany considers lifespans of 120 years for residential buildings to better reflect sustainability goals. The German estimations from (Bradley & Kohler, 2007) support this shift. As our estimates as well as estimates by (Aagaard et al., 2013) and (Videnscenter, 2023) all far exceed the current observation period of 50 years, we find that a similar approach would be beneficial in a Danish context.

7 Discussion

This section discusses the assumptions made in our analysis, critiques their validity, and evaluates their impact on the results. Additionally, we explore how data limitations and different causes of demolition affect our analysis.

7.1 Causes of demolition

Various underlying factors contribute to whether a building is demolished or not. When investigating service life through a data-driven perspective we assume that we can, at least to a certain degree, uncover the important factors to why a building is being demolished through the available data. Even though the building data available in Denmark is well registered and relatively comprehensive compared to international standards, there are still many reasons for demolition that this data cannot cover.

Dynamic processes like changes in housing demand, evolving aesthetic tastes, and technological advancements all affect the service life of buildings - as does political sensibilities, for example when an apartment complex is demolished to address socioeconomic segregation or eliminate areas associated with crime.

Demolitions can be costly. Some buildings may well qualify for demolition based on criteria of functionality and damage, but if no one has an economic interest in the demolition, the building may decay on its own with no demolition. For certain kinds of buildings, economic factors can also influence the decision between demolition and reconstruction. For example, polychlorinated biphenyls (PCB) was used in buildings up until the 1970s. As they have been proven to cause health problems, PCB's should be handled with care, making reconstructions very costly and thus encouraging demolition of the building. Futhermore, in times of national economic growth and rising land prices, the incentive to tear down fully-functional but less profitable buildings increases [20]. Twenty years ago, the land value accounted for $36 - 40\%$ of the real estate value of a single-family house. Today, the land value accounts for an average 60% of the real estate value, strongly increasing the incentive of so-called "teardown-sales"[5].

This trend is counterbalanced by the growing emphasis on climate-consciousness in the building sector. Both activists and policymakers have underscored the need to shift focus from new construction toward a circular building sector, with mottos such as "Preserve or explain" (DA: "Bevar eller Forsvar"). Jens Bomann Christensen, director of the lobby organization *Brancheforeningen Cirkulær*, goes so far as to suggest that all newly constructed buildings should be automatically listed for preservation during the first 100 years of their service life [15]. These considerations highlight the complex dynamics leading to a building demolition and how it can be influenced by economic trends, societal needs, political considerations, and environmental awareness. While data-driven approaches can help identify current patterns, they cannot capture the wide range of dynamic influences.

7.2 Data shortcomings

A thorough discussion of data errors in the BBR-dataset was undertaken in previous sections, highlighting areas of uncertainty. Still, compared to the large number of observations, these errors account for a miniscule part of the data. As we have the whole population of buildings demolished from 2010-2024, the figures and statistics presented for these uncensored observations can be considered true findings from the BBR data.

Through the BBR, only data on the outer wall material is available, not the load-carrying structure of

the building. This makes it especially hard to do service life estimation for different types of buildings as we expect the load-carrying structure to have a huge influence. From a legislative perspective, the availability of this data is also of significant importance, as a substantial portion of a building's climate impact comes from its load-carrying structure. If we were to implement varying observational periods (instead of the current uniform 50-year observational period) for different types of buildings, such data would be essential to support the development of these policies.

As outlined in Section 4.3.2 there are clear misalignments between the BBR and the aggregated datasets from Statistics Denmark. There is a possibility that some of the discrepancy could be solved if the micro-data was made available, such that the exact use-subcategories used for each dataset could be compared. The datasets from Statistics Denmark also exhibit internal inconsistencies. These include, but are not limited to, whether the data tracks the number of buildings, houses or building units, and if sheds, garages etc. are included in the data. In spite of these misalignments we used the numbers of constructed buildings pr. year obtained through Statistic Denmarks BYGV05A for our simulation. The lack of reliability of this data naturally affected this analysis, but as we only used this data to test the Turnbull algorithm and not for any of the final results, we do not find it too worrying.

In Section 4.2 we present a range of figures and tables on the data, exploring how use-category, building material and construction year are connected. Here, we do not assume that construction materials and use-categories are independent of construction year. We cannot do that, as the data is so deeply coloured by these dependencies and all explanation about the data needs to consider this. However, the lack of data to substantiate the claims made often mean that only weak conclusions or speculations can be made. We hope that more data on this would give rise to stronger conclusions about the data. Getting left-censored data from the 80s, 90s and 00s would already strengthen the possible conclusions and show these dynamics on a longer scale.

7.3 Key assumptions and biases

1. **Assumption of constant demolition rate:** We find that the demolition rate is constant in the period 2011-2023 and then assume that this demolition rate holds for the rest of the period 1917-2024. This hypothesis is sustained by (Andersen, 1992), where the same rate was found for the 1980s. However, the demolition rate may vary across use-categories and geography, so even a time-dependent demolition rate could be an oversimplification.

   The demolition rate γ is modeled such that it does not increase with the age of the building. But as it follows a geometric distribution, the probability of *having been* demolished does increase as the building ages. This is intuitive up to some high age, but at some point this assumption may not hold, as buildings can be listed or deemed preservation-worthy. This mechanism effectively adds a decreasing demolition rate for very old buildings (found to be from 370 years in our data).

   A constant demolition rate is the simplest assumption. But it could be that the demolition rate correlates with the construction rate, either positively or negatively. If all resources in the building sector are generally utilized, in periods with much building activity, there are fewer resources for demolitions and in periods with much demolition, there are fewer resources for construction. This would impose a varying demolition rate. We find no conclusive evidence for this hypothesis in our data or in the literature, meaning that our assumption of stationarity is the safest choice.

2. **Assumption that service life is independent of construction year:** We assume that the service life distribution is independent of construction year, and thereby that the buildings demolished in 2010-2024 are representative of the buildings before and after. This assumption can break for a number of reasons. One that is of particular interest is if building quality itself is not constant in the period. Some papers have fallen into the survivorship bias trap to say that

building quality has decreased over time ("Look at this half-timbering house from the 1600s! If we still built like that all houses would stand for 400 years") while building sector stakeholders will generally argue that technological advancements has increased building quality significantly. As long as no long-term data exists, we find the assumption of stationarity of service life to be convenient and simple. When we compute average service lives and model phase-types on this data, it is realistic that the conclusions hold for the coming years - but not necessarily decades into the future or the past, as the yearly building activity is dependent on the population size, the population growth and the national economic situation. The buildings demolished today show the aftermath of the building trends of the past. This means that the construction boom in the 1960s and 1970s affects the data to this day, with the building trends of this time being heavily represented. As these buildings age, we might see shifts in demolition trends. For example, this period was the first time summer houses were mass built, which may show up more heavily in demolition data in future decades.

3. **Bias induced by non-uniformity of construction year distribution:** We know that particularly many buildings were built in the 60s and 70s, meaning that - by sheer volume of buildings - many buildings from this period will be demolished in our data. These buildings are between 40 and 64 years old and can affect the service life estimation by overestimating the probability of these service lives. While building booms will happen occasionally, it is an open question whether they should be considered a natural part of the data or something to be skeptical of, but either way, it is important to consider how it might affect our conclusions.

7.4 The impact of assumptions and biases

We will describe how these assumptions affect the estimates and figures presented in this thesis.

7.4.1 Phase-type model estimation

When estimating phase-type distributions on houses demolished 2010-2024, the estimation can be affected by the non-uniformity of construction year. As the building boom of the 1960s and 1970s result in an overrepresentation of these construction years, we might have an overrepresentation of service lives of ages 40 to 64. We find that this is among the earliest service lives where demolition is plausible, meaning that the service life is likely underestimated because of this. This also means that the age of buildings is likely to rise in the coming years, as this massive bulk of buildings age.

We also use the phase-type model to backcast a dataset. Here, the assumption that service lives are independent of construction year is used. As described above, this assumption is strong, but we cannot support any other hypothesis in a data-driven manner. The use of the phase-type survival curve in the Turnbull algorithm, means that we compute survival probabilities very similar to those from the data. This extends the demolition trends of recent years onto conclusions about the future. As we suspect the non-uniformity of construction year weights down our service life estimates, the survival probabilities computed from this assumption might also be too low.

One could have chosen to do the service life estimations on the entire population of Danish buildings demolished 2010-2024. An argument in favor of this approach is that the current Danish legislation does not distinguish the service life between different use-categories of buildings, and if they are to keep this approach, a common service life distribution is needed. However, we do not find this approach reasonable, considering the very disparate service life distributions for the different use-categories.

Instead, we chose to focus on modeling of houses, based on both societal and data considerations. This use-category is still not a homogenous mass, incorporating both old farmhouses in rural

areas and big apartment complexes in urban centres. We found that the distribution of service life was mostly the same across the categories, but not identical. A more fine-grained analysis might reflect specific categories more closely, but this would also limit the amount of data available.

7.4.2 Markov chains

We computed the underlying Markov jump process from our phase-type models. We interpreted the various states as different states of decay and found several interesting parallels to the data - for example the $\approx 0.6\%$ chance of jumping directly to absorption after around 13 years is confirmed in the dataset. We also found that the time before reconstruction from our data was captured by the Markov chain - a rather surprising catch, as the Markov jump processes were not modeled with this data. These findings can confirm some of the findings of the models.

Still, our interpretations of construction errors and reconstruction cycles remain hypothetical models as the states do not correspond directly to real-life decay processes. Furthermore, as the representation of the phase-type distribution is not unique, there might be different Markov chains representing the same distribution. Under an acyclic representation, such as a Coxian representation, our interpretations of reconstruction-cycles would not be meaningful. But as the parameters are the way they are, the process *can* be described by the depicted Markov chains and find that they provide interesting interpretations.

As stated before, a new focus on longevity, reconstruction and preservation of buildings might also change the dynamics presented here - maybe more buildings will jump into reconstruction cycles in the future or maybe fewer buildings will be built for ultra short-term use.

7.4.3 Regression-based demolition rate

The demolition rate of 0.16% stems from linear regression on the demolition rates of 2011-2023. Our regression shows that this assumption holds in the period 2011-2023, but if the demolition rate actually varies, this would affect the backcasting. Instead, we could assume that the demolition rate is inversely proportional to the construction rate, which changes over time. We find that the period 2010-2024 has a higher average construction rate than the period 1917-2009. From this it would follow that the average demolition rate is higher than our estimate. This would lead to more left-censored data, thus decreasing the survival probabilities and meaning that our estimate would be too high.

The considerations presented in this section function to underline the assumptions of our models and their implications. We find that our results are highly sensitive to these assumptions. Acquiring further data will enable us to inspect our assumptions on a longer timeframe as well as weaken the influence of assumptions.

8 Future research

This report demonstrates initial steps towards a data-driven building service life estimation and in doing so, it opens up a range of new questions and points towards several next steps.

One of the key challenges discussed in this report is the issue of left-censored data. A first step is to get access to additional years of BBR data, as this would increase the proportion of uncensored data. However, left-censored observations will always exist. An interesting area of research would be to implement an EM-algorithm for phase-type distributions which can accommodate left-censored observations. Futhermore, the problem of a stalling EM-algorithm raises additional research questions regarding how variable step-length or other heuristic elements could be implemented in the EM-algorithm without losing the special property that the log-likelihood always increases.

We note that some empirical distributions seem bimodal. This was not the case for our primarily investigated use-category of housing, but for example the distribution of "Agriculture" and "Production" demolished 2010-2024 exhibits signs of this. For more modeling into these use-categories, bimodal initial guesses in the EM-algorithm might be interesting to explore, particularly in combination with a large number of phases.

Accessing more long-term BBR data would also help identify trends over a more extended period. For example, it could shed light on how economic cycles influence demolition rates, providing data for modeling a variable demolition rate.

We have a lot of variables available in the data about each building that could be integrated into service life estimation. When dividing into use-categories we still have some heterogeneity which in future work could be incorporated into the regression modeling by including the use-subcategories, such as apartment complexes and single-family houses, or by including the geographic area, divided into rural areas, suburbs and urban areas. Phase-type regression modeling is an evolving field and building service lives would be a novel application area. More differentiated service life estimations would improve the stochastic simulation of left-censored data and lead to more reliable Markov chain models. This is of particular interest to the industry, where many stakeholders advocate for differentiated observational periods in LCA.

Our report also highlights issues within the BBR dataset. With the reinstatement of government property valuations from the Danish Property Assessment Agency, attention has been drawn to the importance of accurate and consistent data collection. We strongly second this notion, as further work is still required by Statistics Denmark and BBR in quantifying and addressing these issues. For instance, problems like the misalignment between building units and buildings, as well as the inconsistent use of building use-categories, pose significant challenges to the data's reliability.

While the dataset already offers a wealth of variables, external variables could also be included in the modeling. We have considered the mean wages pr. municipality with no proven significance. We have also examined if geographic location have any influence. Other factors that might be of interest are ownership information as well as the building and sales activity of the geographical area. Aesthetic value is also an important factor in demolition decisions. However no parameters for aesthetic value is included in the dataset and quantifying aesthetic value in a comprehensive and comparable way for all Danish buildings is practically infeasible. An interesting proxy is The Agency for Castles and Culture's registry of listed buildings and buildings worthy of preservation

[41]. Currently, this registry is not mergeable with the BBR dataset. Nonetheless, incorporating this data into future research could provide valuable insights.

9 Conclusion

We commenced this thesis by investigating the most important variables in the BBR. We found that the mean service life of buildings demolished between 2010 and 2024 is 71.81 years, with a mean service life of 96 years for housing.

We found that reconstructed buildings have a longer service life than those not reconstructed. This finding contributes to the debate of demolition versus reconstruction, but due to the risk of survivorship bias, the fact may be used only with care.

We modeled phase-type distributions to our service life distributions and found that the model accuracy improves with each added phase. We did not reach a plateau where the BIC or AIC-scores were constant or rising with increasing phases, indicating that the modeling has not yet reached a satisfying description of the data.

From the phase-type models, we computed the underlying Markov jump process, showing how building decay could be described through various phases. Three paths emerged; a very small group jumped directly to absorption, which we interpret as buildings with errors or very short-lived buildings. The majority jumped from phase i to phase $i+1$ for all $i = 1, \ldots, p$, corresponding to an uninterrupted decay process. About 1 in 4 buildings enter what was interpreted as a "reconstruction cycle", showing how intervention can prolong the service life by turning back the building to an earlier state of decay.

A yearly demolition rate for housing in the years 2011 - 2023 was estimated to be 0.016. This harmonizes with demolition rates estimated in the 1980's but differs from the currently used 1% demolition rate in Danish financial models and by Statistics Denmark. Here, we find the data-driven approach taken in this thesis to be more reliable.

Based on our demolition rate, we estimated a survival curve using Turnbull's algorithm resulting in a median service life of 301 years and a probability of surviving to age 50 of 93.8%. Assuming a non-constant demolition rate based on a phase-type model, and using Turnbull's algorithm to estimate a survival curve, we found a median service life of 92 and a probability of surviving to age 50 of 91.6%. While the two methods find similar (high) probabilities of housing buildings surviving 50 years, they quickly diverge from there as seen by the discrepant median values.

Comparing these estimated medians to existing literature, we conclude that our estimate using $\gamma = 0.0016$ is generally in the high end of estimates, but similar to estimates made in comparable countries. The estimate made using a phase-type assumption is in line with other estimates made using uncensored data.

Compared to current Danish legislation all our estimates of service life is significantly higher than 50 years, creating an incentive to reconsider the legislation. Additionally we showed that the service life distributions differ greatly between use-categories, further supporting the argument that service life estimates used for LCA should be differentiated across use-categories.

Bibliography

[1] Videnskabsetiske Medicinske Komiteer (VMK). *Retrospective Enrollment*. Accessed: 2024-11-05. 2024. URL: https://researchethics.dk/research-ethics-themes/retrospective-enrollment.

[2] Niels-Jørgen Aagaard et al. *Levetider af bygningsdele ved vurdering af bæredygtighed og totaløkonomi*. 2013.

[3] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[4] Ellen Andersen. "En bedre boligmodel". Dansk. In: *Nationaloekonomisk Tidsskrift* 130.1-2 (1992), pp. 181–188.

[5] Marc Lund Andersen. *Grundpriser for enfamiliehuse 1996-2019 - med fokus på huse købt til nedrivning*. Boligøkonomisk Videncenter, 2022.

[6] Rune Andersen and Kristoffer Negendahl. *Lifespan prediction of existing building typologies*. 2023.

[7] Søren Asmussen. *Applied Probability and Queues*. 2nd ed. Vol. 51. Stochastic Modelling and Applied Probability. New York, NY: Springer, 2003.

[8] Søren Asmussen, O. Nerman, and Marita Olsson. "Fitting Phase-Type Distributions via the EM Algorithm". In: *Scandinavian Journal of Statistics* 23 (1996).

[9] *Bekendtgørelse om ajourføring af Bygnings- og Boligregistret (BBR)*. Danish Ministry of Taxation. 2012. URL: https://www.retsinformation.dk/eli/lta/2012/1010.

[10] Martin Bladt, Jorge Yslas, and Alaric Müller. *matrixdist: Statistics for Matrix Distributions*. R package version 1.1.9. Comprehensive R Archive Network (CRAN), 2023.

[11] Mogens Bladt and Bo Friis Nielsen. *Matrix-Exponential Distributions in Applied Probability*. Springer, 2017.

[12] Social- og Boligstyrelsen. *Energiforbrug og klimapåvirkning (Paragraph 250 - 298)*. 2024. URL: https://www.bygningsreglementet.dk/Tekniske-bestemmelser/11/Krav/297_298.

[13] Patrick Erik Bradley and Niklaus Kohler. "Methodology for the survival analysis of urban building stocks". In: *Building Research & Information* 35.5 (2007), pp. 529–542.

[14] Patrick Erik Bradley et al. *Survival functions of building stocks and components*. 2005.

[15] Jens Bomann Christensen. "I Danmark er vi vilde med at rive bygninger ned, men det tærer på klimaet". In: *Altinget* (Nov. 2023).

[16] PROJ contributors. *pyproj 3.6.1 documentation*. Accessed: 2024-09-30. 2017. URL: https://pyproj4.github.io/pyproj/stable/api/proj.html.

[17] Marie-Laure Delignette-Muller et al. *fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*. Version 1.2-1. R package. 2024.

[18] Mc Duling and Johannes Jacobus. *Towards the development of transition probability matrices in the Markovian model for the predicted service life of buildings*. 2007.

[19] Kim Haugbølle et al. *BUILD Rapport 2021:32*. 2021.

[20] G. C. Hufbauer and B. W. Severn. "The Economic Demolition of Old Buildings". In: *Urban Studies* 11.3 (1974), pp. 349–351.

[21] Andreas Kryger Jensen et al. *I gamle dage byggede man udødelige bygninger eller hvad?* Accessed: 2024-11-25. 2024. URL: https://videnskab.dk/kultur-samfund/i-gamle-dage-byggede-man-udoedelige-bygninger-eller-hvad/.

[22] Jesper Ole Jensen et al. *Nedrivning af enfamiliehuse: Omfang og årsager*. 2022.

[23] Sukwon Ji, Bumho Lee, and Mun Yong Yi. *Building life-span prediction for life cycle assessment and life cycle cost using machine learning: A big data approach*. 2021.

[24] Johann et al. *Service Life Prediction Beyond the 'Factor Method'*. 2008.
[25] John P. Klein and Melvin L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Springer, 2003.
[26] Klimadatastyrelsen. *"Datafordeler"*. https://datafordeler.dk/dataoversigt/. Accessed:20/08/2024.
[27] Michel Kornmann and Andreas Queisser. "Service life of the building stock of Switzerland". In: *Mauerwerk* 16 (2012), p. 210.
[28] L.I.Aarseth and P.J.Hovde. *A stochastic approach to the factor method for estimating service life*. 1999.
[29] Guiwen Liu et al. "Factors influencing the service lifespan of buildings: An improved hedonic model". In: *Habitat International* 43 (2014), pp. 274–282.
[30] Martin Maechler et al. *expm R Package*. Version 1.0-0. 2024. URL: https://R-Forge.R-project.org/projects/expm/.
[31] David Meisch. *"Multivariate phase type distributions - Applications and parameter estimation"*. 2014.
[32] Bo Friis Nielsen. *02443 Stochastic Simulation Course Slides*. Accessed: 2024-09-30. 2024. URL: https://www2.imm.dtu.dk/courses/02443/.
[33] Nominatim. *Open-source geocoding with OpenStreetMap data*. Accessed: 2024-09-30. URL: https://nominatim.org/.
[34] Natasha Østergaard et al. *Data Driven Quantification of the Temporal Scope of Building LCAs*. 2018.
[35] Reza Pulungan. "The Order of Phase-type distributions". In: *The 6th SEAMS-GMU 2011 International Conference on Mathematics and Its Application* (2011).
[36] Lídia Rincón, Gabriel Perez, and Luisa F. Cabeza. "Service life of the dwelling stock in Spain". In: *The International Journal of Life Cycle Assessment* 18 (June 2013).
[37] SAS. *Datakvaliteten af ejerboliger i BBR*. https://bbr.dk/file/665988/bbr-datakvalitet-af-ejerboliger-i-bbr.pdf. 2015.
[38] S. Sawyer. "The Greenwood and Exponential Greenwood. Confidence Intervals in Survival Analysis". In: *Department of Mathematics - Washington University in St. Louis* (2003).
[39] PJ Hovde Norwegian University of Science, Department of Building Technology, and Norway Construction Engineering. *The Factor Method For Service Life Prediction From Theoretical Evaluation To Practical Implementation*. 2002.
[40] Carles Serrat. *The use of survival analysis in building maintenance*. 2009.
[41] Slots- og Kulturstyrelsen. *Fredede og Bevaringsværdige Bygninger (FBB)*. Accessed: 2024-09-24. 2024. URL: https://www.slks.dk/fredede-og-bevaringsvaerdige-bygninger.
[42] Danmarks Statistik. *Sammenlignelighed for byggevirksomhed*. Accessed: 2024-10-16. URL: https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/byggevirksomheden.
[43] Danmarks Statistik. *Sammenlignelighed for bygningsopgørelsen*. Accessed: 2024-10-16. URL: https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/bygningsopgoerelsen.
[44] Kong Fah Tee, Ejiroghene Onome Ekpiwhre, and Zhang Yi. "Degradation modelling and life expectancy using Markov chain model for carriageway". In: *International Journal of Quality & Reliability Management* (2018).
[45] Bruce W. Turnbull. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 38.3 (1976), pp. 290–295.
[46] UP42. *Coordinate reference systems*. Accessed: 2024-09-30. URL: https://docs.up42.com/data/reference/utm.
[47] Boligøkonomisk Videnscenter. *Karakteristika for huse der rives ned med henblik på nybyggeri*. 2023.

[48] Vurderingsstyrelsen. *"Kodelister - Forretningsproces"*. https://teknik.bbr.dk/kodelister/0/1/0/ForretningsProcess. Accessed:20/08/2024.

A Appendix

A.1 Approximative dating patterns

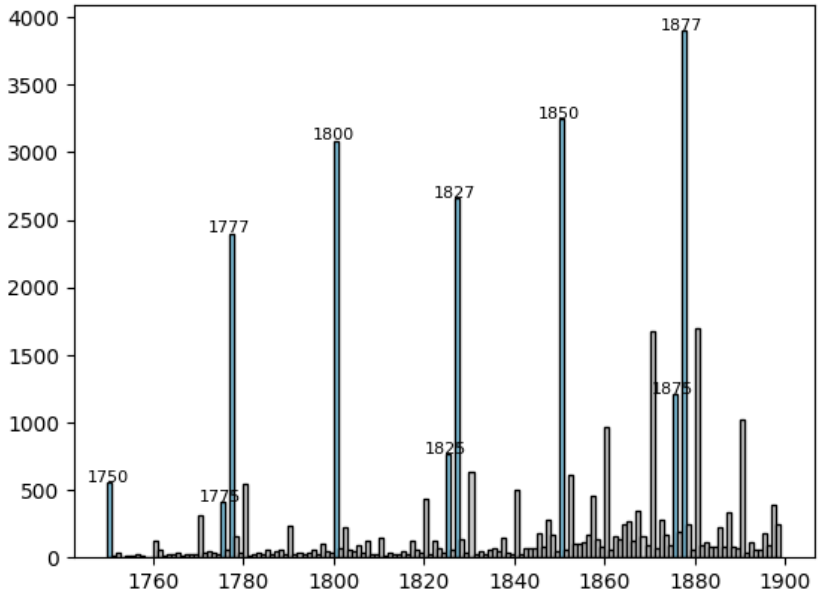


Figure A.1: People seem to round the construction year of half-timbering houses to closest quarter century. And for some reason also to the years 1777, 1827 and 1877

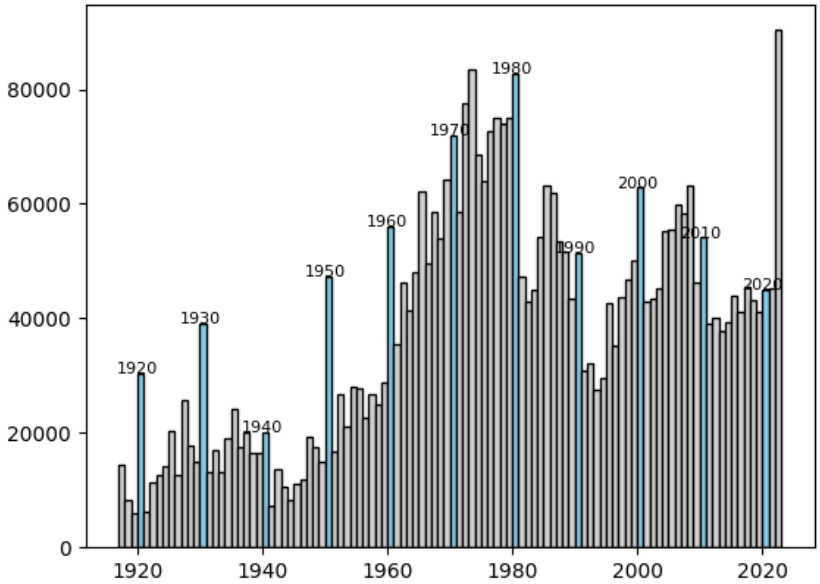


Figure A.2: Construction year for all buildings standing 2024 or demolished 2010 to 2024. The approximative dating is evident in the spikes in 1920, 1930, 1950 and 1960.

A.2 Constructed area comparison of BBR and BYGV05 weighted by BYGV06

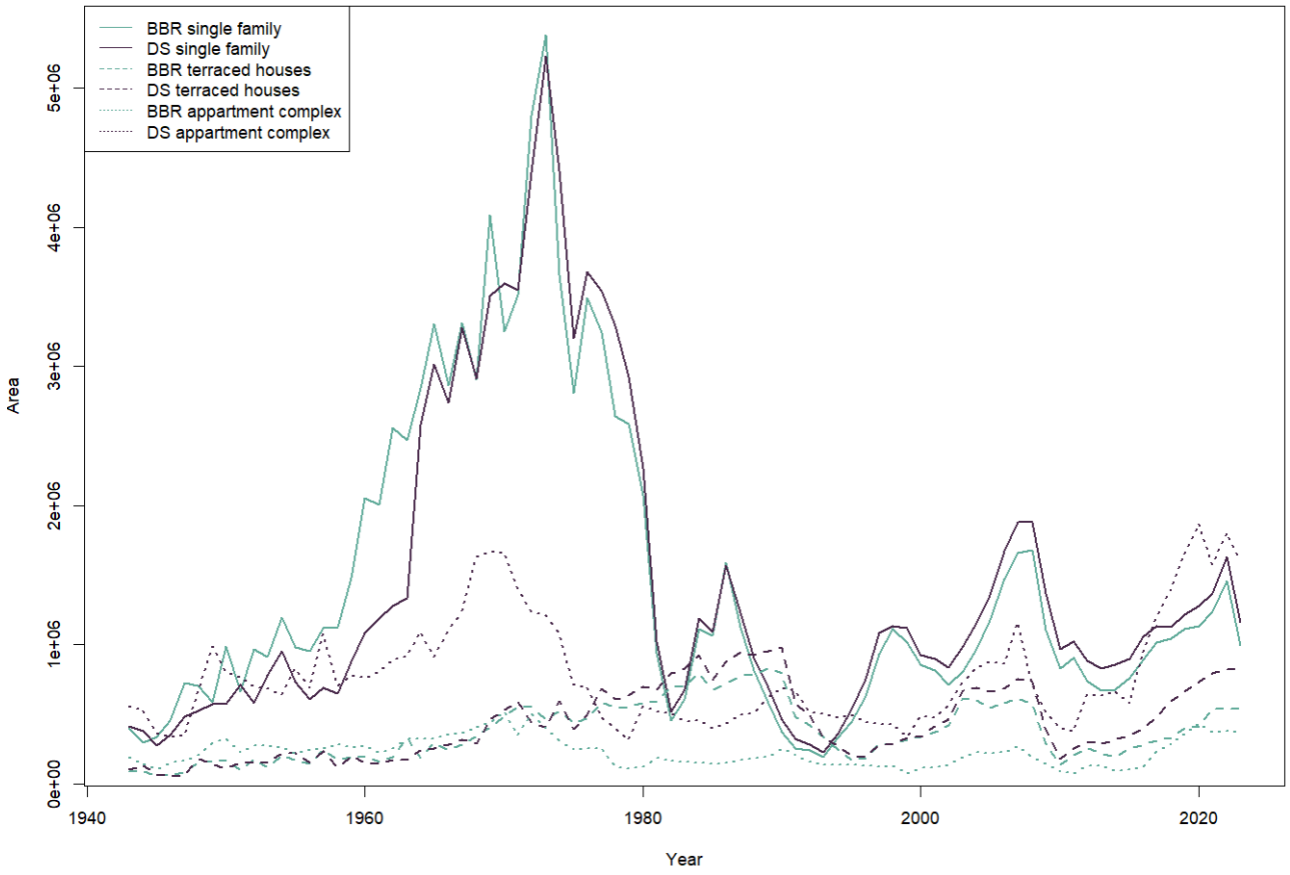


Figure A.3: Comparison of total constructed area from BBR and estimated total constructed area from BYGV06 and BYGV05A.

A.3 Investigation of old use-categories

In this appendix we wish to investigate the original distribution of the uses for each use-category. This might reveal if the service life follows a mixture of distributions or can be modelled as one joint distribution.

Housing

For the use-category "housing" we see a quite widespread and flat right-skewed distribution. It has a big peak around 60-70 years, and then another small peak at 120 years. This could indicate a mixture distribution. That mixture can either be a characteristic of the whole distribution or it might be caused by the mix of the different old use-case categories. We created the plot below to investigate this:

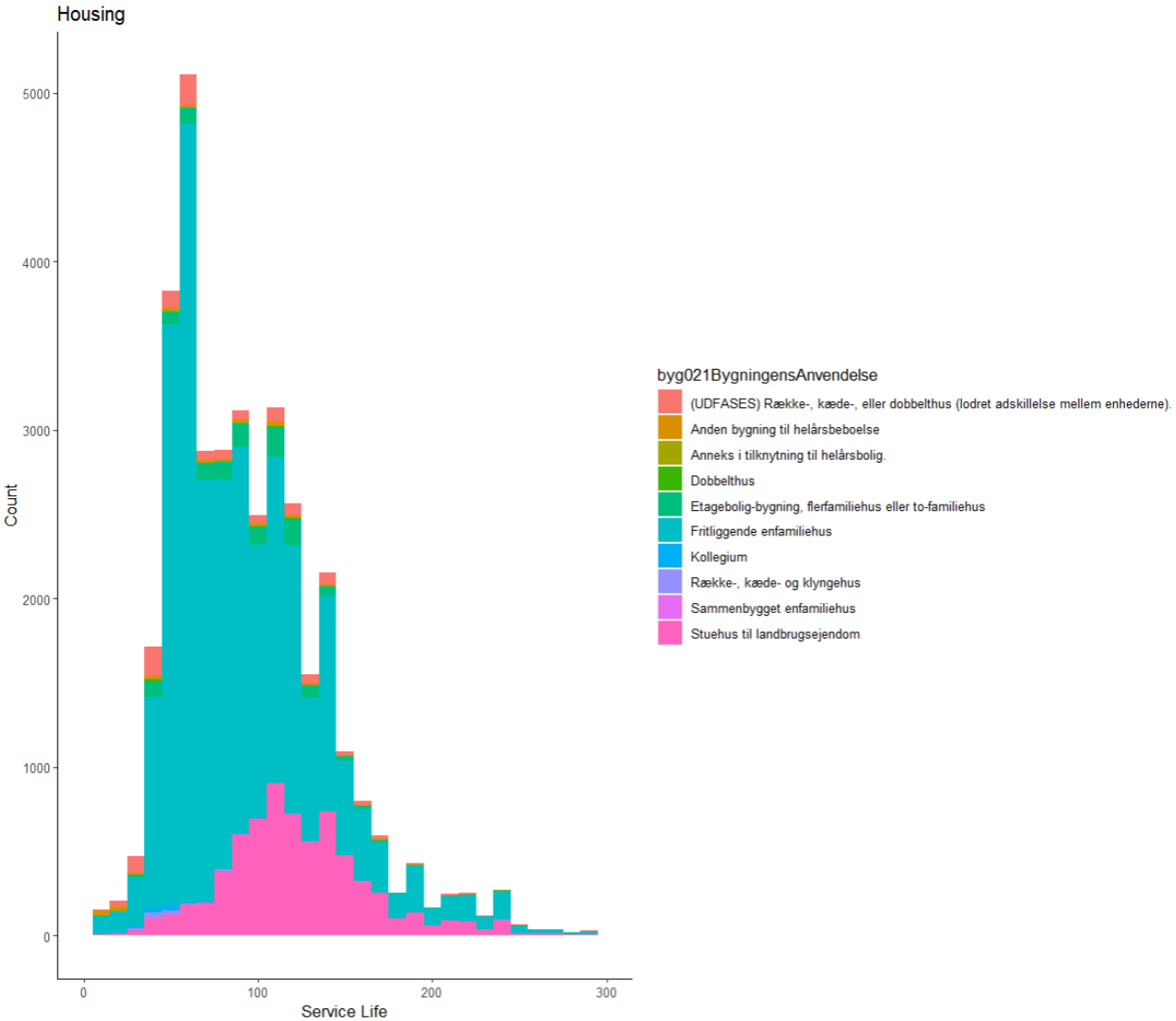


Figure A.4: Housing distribution of old categories

We see that the pink distribution "farmhouse for agricultural property" fills out some of the second peak and tail of the distribution, while single family houses fills out almost the rest. This could show that the different use-subcategories we bin together, may have slightly different decay processes. We are not too concerned with this particular case, as farmhouses make up a smaller proportion of the housing use-category. We also note that farmhouses are no longer a popular housing option to build and haven't been for the last centuries. While the other use-subcategories mostly follow the same distribution as the binned distribution, these make up a numerically small part of the use-category.

Leisure

For the use-category "leisure" we see a very sharp peak at 50-60 years of service life, with a quite sudden drop on both sides. The use-category "leisure" is mostly (86%) summer houses, a category which only became mainstream 60 to 70 years ago, explaining why only a few older buildings are found in this use-category. It is still somewhat right skewed with a slight tail, but all use-subcategories follow the same distribution.

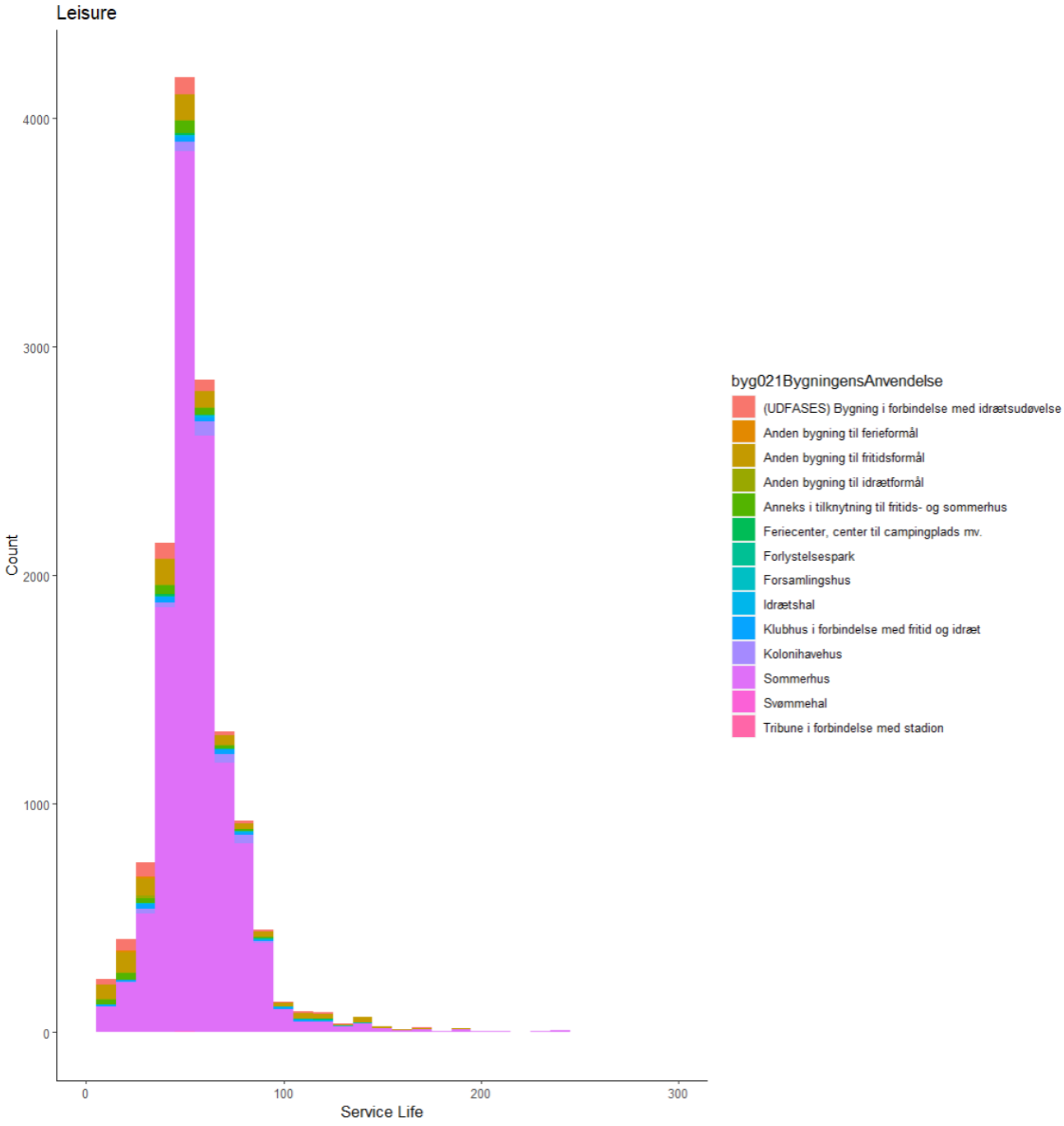


Figure A.5: Leisure distribution of old categories

Institution

For the use-category "institution" we see a peak around 50 years of service life. Once again we see either a right-skewed distribution or the possibility of a mixture distribution. Below the stacked histogram of the old use-category categories is plotted:

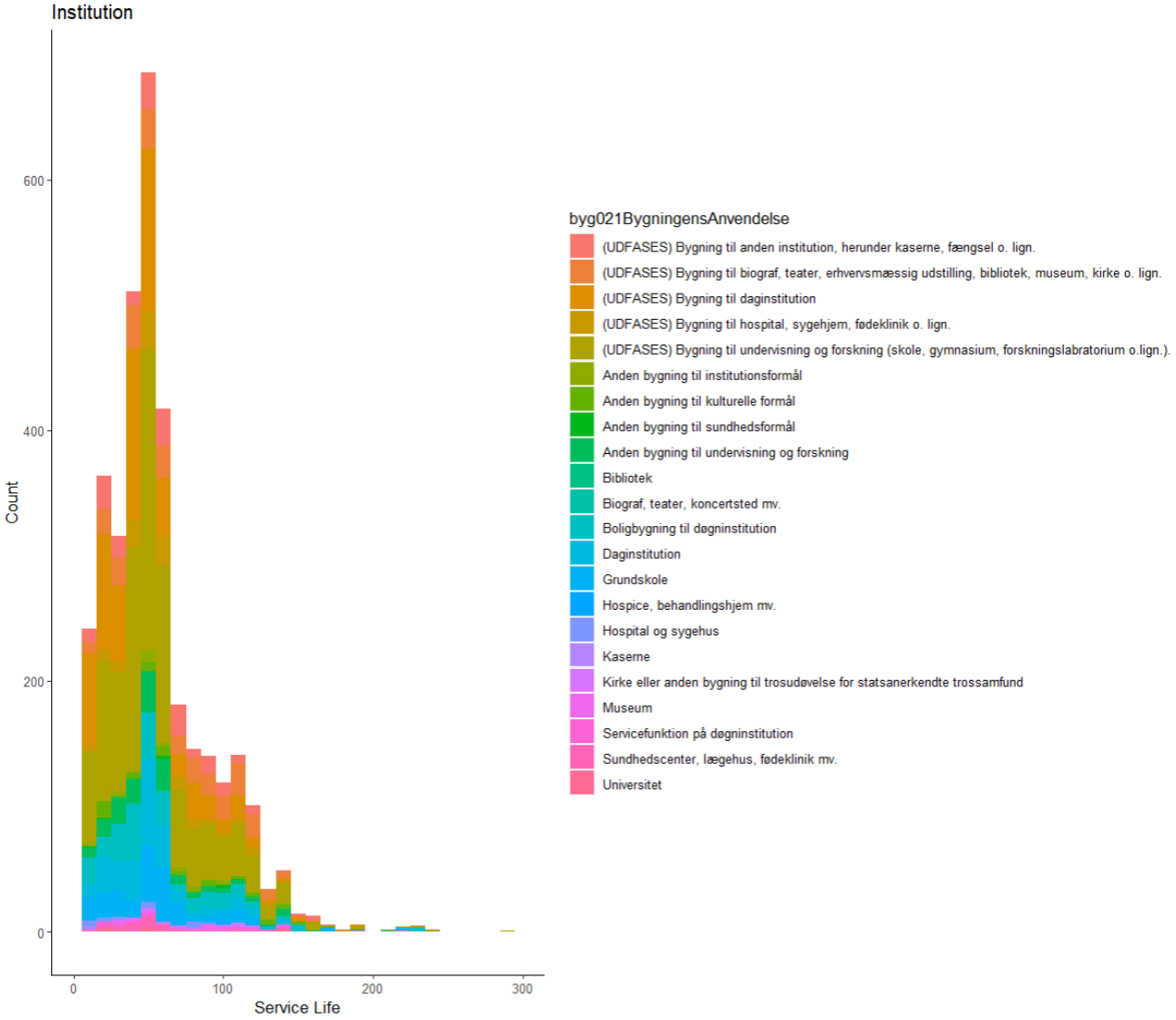


Figure A.6: Institution distribution of old categories

Here we see that the bimodal distribution tendency is not due to different old categories but rather something that is a general tendency in multiple of the categories.

Agriculture

For the use-category "agriculture" we see a clear bimodal distribution that peaks at 40-50 and 90-100 years of service life. The stacked histogram below shows that the bimodal distribution is consistent through the categories and thereby a general tendency of the agriculture distribution. This could be a sign that a mixture distribution is needed to fit this data for future modeling.

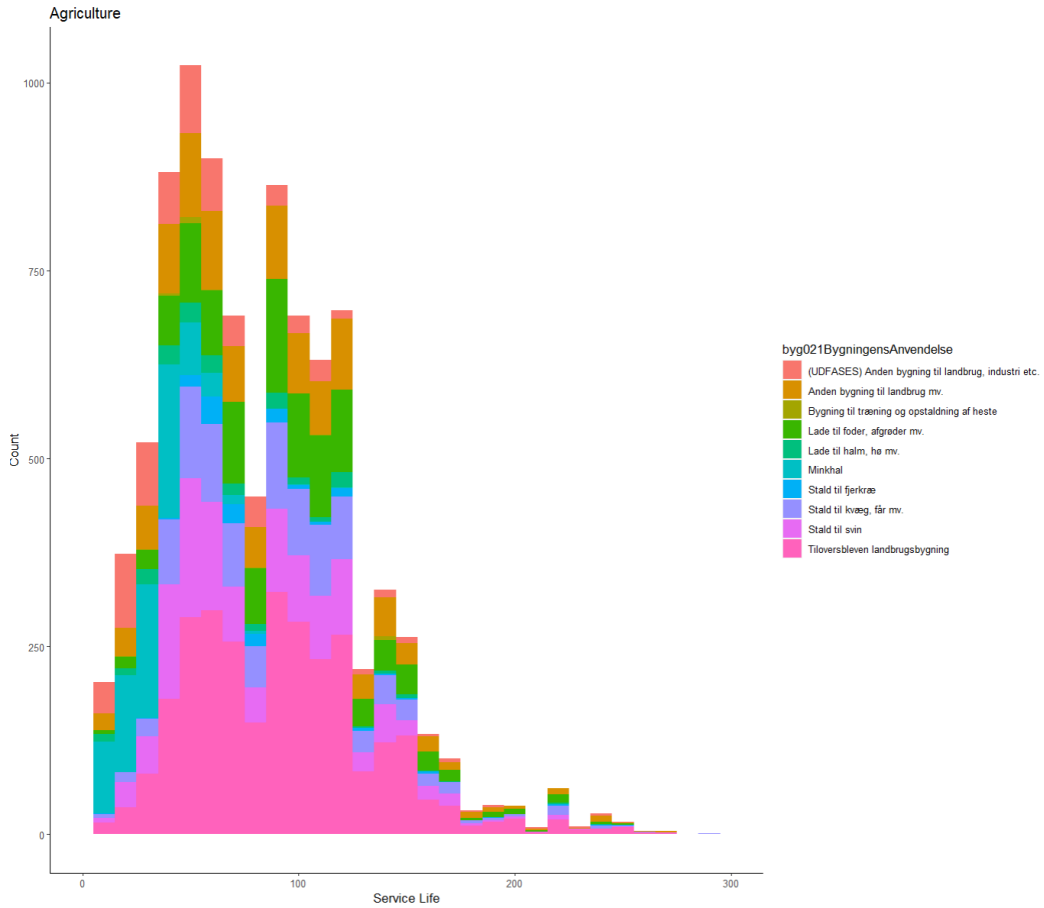


Figure A.7: Agriculture distribution of old categories

Production

For the use-category "production" we see a big peak around 50 years and another slightly flatter peak around 100 years, once again resulting in a bimodal distribution. The stacked histogram presented below shows that the buildings used for production of agriculture carry most of the general distribution while the rest of the categories mainly add to the spike around 50 years of service life.

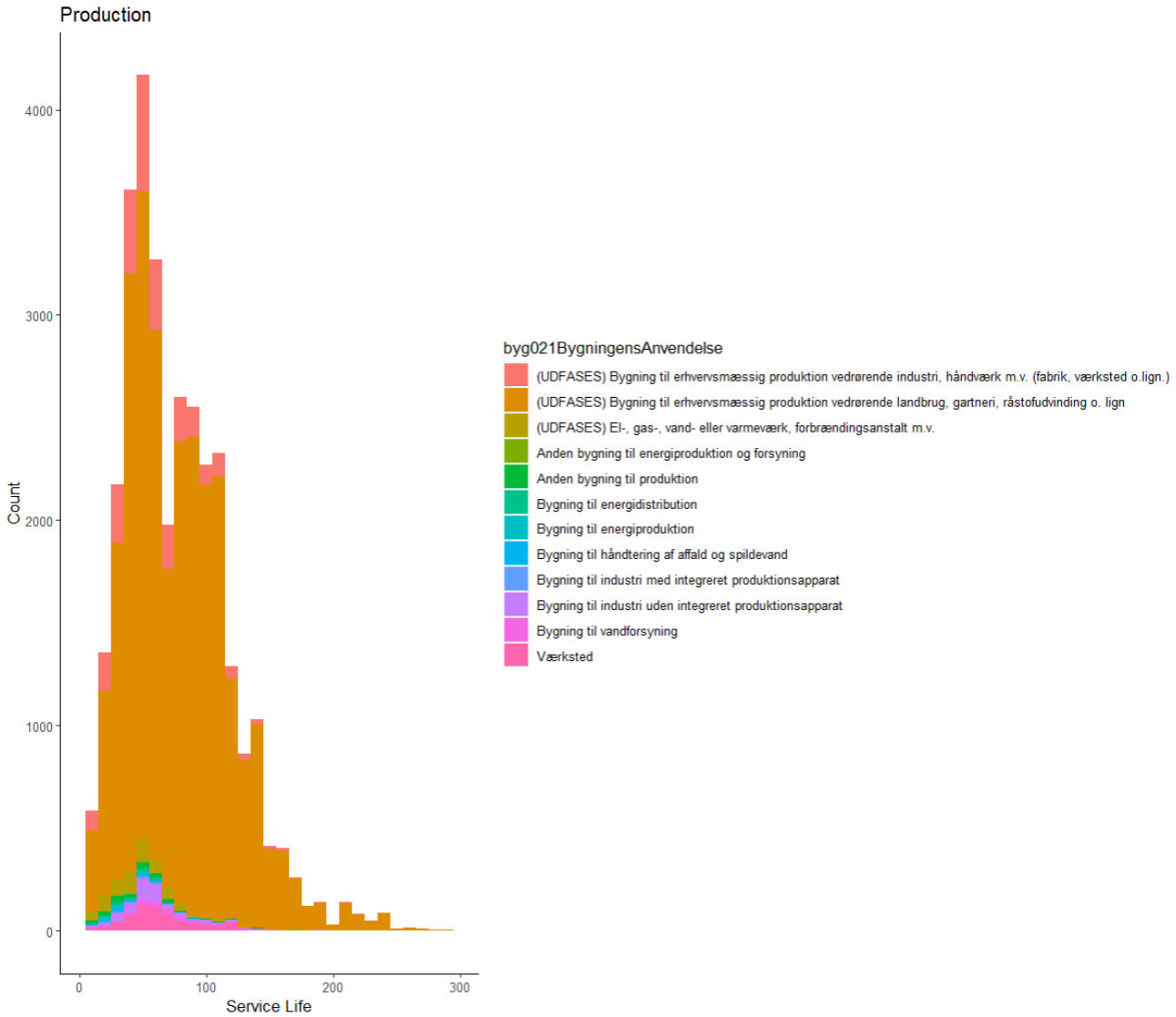


Figure A.8: Production distribution of old categories

Transport and Commerce

For the use-category "transport and commerce" we wee a clear peak at 50-60 years of service life and then a sharp drop on the right side into a heavy tail. Another way to view the "tail" could be to see it as a possible mixture distribution with the sharp peak as the first distribution and a wide symmetrical distribution as the other.

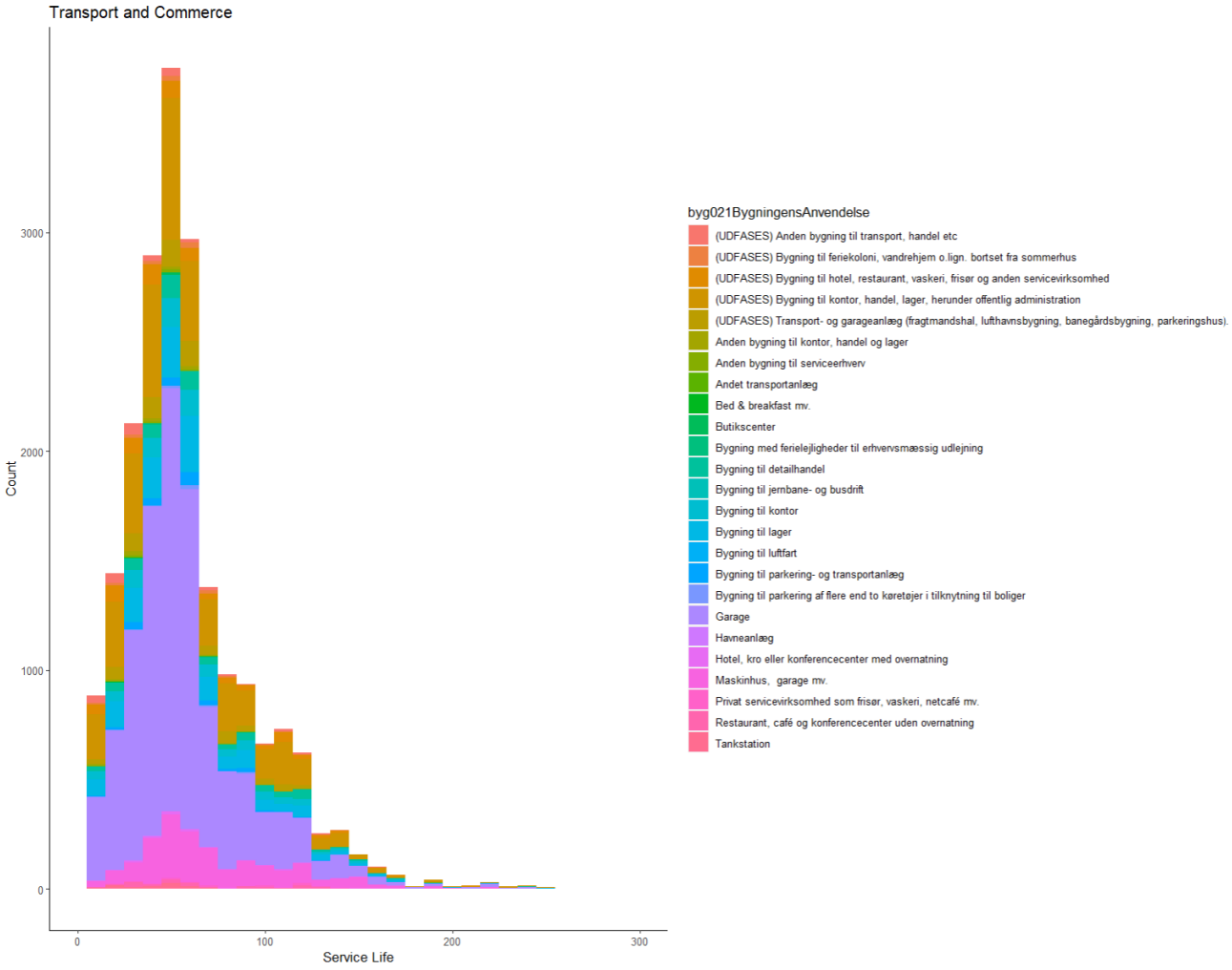


Figure A.9: Transport and Commerce distribution of old categories

Others

For the use-category "Others" we see a peak at 30-40 years of service life. Once again we see tendencies of a mixture distribution. By looking at the stacked histogram below we can see that the category "garages" is responsible for most of the peak, while the category "sheds" has a generally more flat structure, although still peaking around the same time, just with a heavy tail.

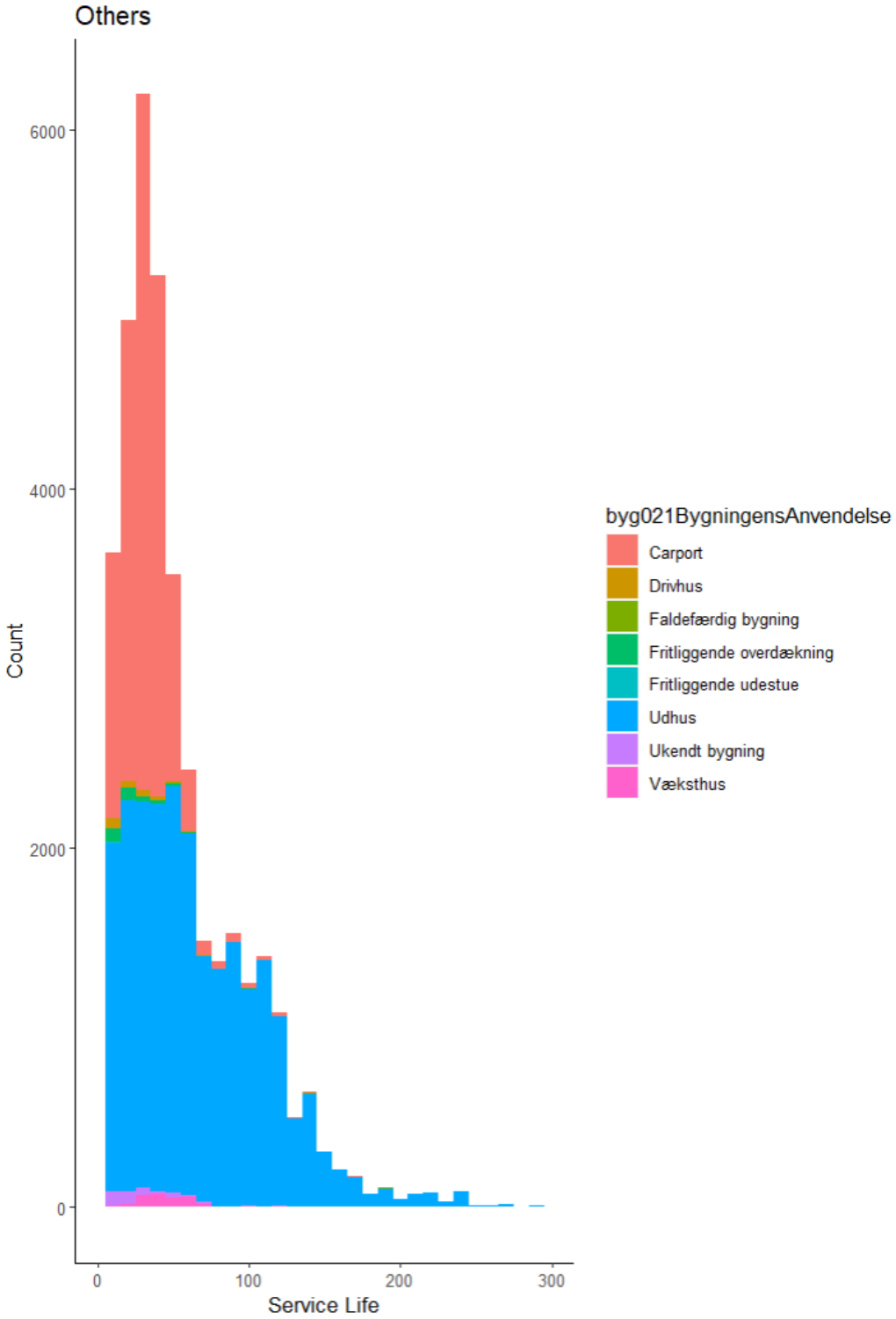


Figure A.10: "Unknown" distribution of old categories

A.4 Estimated parameters for the phase-type models

$$\overline{\mathbf{T}}_4 = \begin{bmatrix} -0.0406 & 0 & 0.0406 & 0 \\ 0 & -0.0428 & 0 & 0 \\ 0 & 0 & -0.0406 & 0.0406 \\ 0 & 0.0406 & 0 & -0.0406 \end{bmatrix} \quad \overline{\pi}_4 = \begin{bmatrix} 0.995 \\ 0.005 \\ 0.000 \\ 0.000 \end{bmatrix}$$

$$\overline{\mathbf{T}}_6 = \begin{bmatrix} -0.087 & 0.0864 & 0 & 0 & 0 & 0 \\ 0 & -0.0718 & 0.0718 & 0 & 0 & 0 \\ 0 & 0 & -0.0718 & 0.0718 & 0 & 0 \\ 0 & 0 & 0 & -0.0718 & 0.0718 & 0 \\ 0 & 0.0117 & 0 & 0 & -0.0718 & 0.0601 \\ 0 & 0 & 0 & 0 & 0 & -0.0515 \end{bmatrix} \quad \overline{\pi}_6 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\overline{\mathbf{T}}_8 = \begin{bmatrix} -0.0727 & 0.0721 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.1382 & 0.1382 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.1382 & 0.1382 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.1382 & 0.1382 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.1382 & 0.1382 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.1382 & 0.1382 & 0 \\ 0 & 0.0374 & 0 & 0 & 0 & 0 & -0.1382 & 0.1009 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.0411 \end{bmatrix} \quad \overline{\pi}_8 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\overline{\mathbf{T}}_{10} = \begin{bmatrix} -0.0355 & 0 & 0 & 0 & 0 & 0.0355 & 0 & 0 & 0 & 0 \\ 0 & -0.0947 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.1831 & 0.1831 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.1831 & 0.1831 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1316 & 0 & 0 & -0.1831 & 0.0515 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.2039 & 0.2039 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.2039 & 0.2023 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.1831 & 0.1831 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.1831 & 0.1831 \\ 0 & 0 & 0.1831 & 0 & 0 & 0 & 0 & 0 & 0 & -0.1831 \end{bmatrix} \quad \overline{\pi}_{10} = \begin{bmatrix} 0.994 \\ 0.006 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Technical
University of
Denmark

Richard Petersens Plads, 324, 220
2800 Kgs. Lyngby
Tlf. 45 25 25 25