Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Bachelor BSc^2 in Econometrics and Economics

Socioeconomic Life Expectancy Determinants: A Quantile Analysis Approach

Stijn Coremans (581362)

Supervisor:	Kathrin Gruber
Second assessor:	Pieter Schoonees
Date final version:	25th June 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This study investigates the socioeconomic determinants of life expectancy using a quantile regression approach, looking into pre- and post-COVID differences in specific. By employing adaptions of a quantile regression forests (QRF) with and without recursive feature elimination (RFE), various socioeconomic factors are analysed yielding insights into their respective importance in light of life expectancy inequalities. It is shown that economic determinants are of the greatest importance in predicting life expectancies, however, the importance of this role decreases post-COVID. Furthermore, the role of healthcare has become more important due to the COVID-19 pandemic. The contributions of this research to the existing literature are not only limited to providing insights into life expectancy inequalities and the development thereof but also offer valuable future research directions for further assessing the impacts of COVID-19 on socioeconomic inequalities.

Contents

1 Introduction

Life expectancy is a measure of lifespan as well as an indicator of socioeconomic inequality that influences our society. In addition to changing the fundamental social structure of our communities, COVID-19 may have also transformed the fundamental determinants of life expectancy inequalities as we recover from the turbulent environment of an international pandemic. One certainty in these uncertain times, however, is that socioeconomic factors have a significant impact on life expectancy (Mirowsky & Ross, 2000). This impact could be magnified by major changes in global economics that disproportionately affect the most vulnerable. In order to provide light on the interrelated factors influencing human health and lifespan and to provide important insights for policymakers as well as individuals, this research sets out on a mission to unravel these complexities and provide new insights into the post-COVID-19 state of inequality.

1.1 Current State of Research

Socioeconomic inequalities have been a long-standing field of research(Deaton & Paxson, 2001), leading to the great importance of researching the determinants of socioeconomic status to uncover these inequalities. The healthcare sector is one of the most notable settings in which socioeconomic status effects are closely monitored with life expectancy being a central topic of study (Mirowsky & Ross, 2000; Lallo & Raitano, 2018; Bär, Wouterse, Riumallo Herl, Van Ourti, & Van Doorslaer, 2021; Sanzenbacher, Webb, Cosgrove, & Orlova, 2021). Life expectancy differences of up to 5 years are not out of the ordinary for individuals of low and high socioeconomic status and insights into the development of these inequalities are thus of great importance. Furthermore, in light of the recent COVID-19 pandemic life expectancy determinants and their effects can have changed significantly (Strozza, Vigezzi, Callaway, & Aburto, 2024).

Research conducted by Mirowsky and Ross (2000) highlights the significant differences in subjective adult life expectancy between socioeconomic groups in the United States. They found an increase in subjective life expectancy of 0.7 years for each extra year of schooling, while a long period of economic hardship can decrease an individual's subjective life expectancy by 4 years. Expanding on these results, the research of Sanzenbacher et al. (2021) supports the idea that disparities in life expectancy are becoming more widespread in America. They highlight the nuanced effects of specific socioeconomic factors on the level of inequality observed, providing more insight into the relationship between socioeconomic status and life expectancy results. It is for example shown that even when taking an individual's education level relative to their birth cohort, inequalities based on socioeconomic status are persistent. A study conducted by Lallo and Raitano (2018) on the socioeconomic factors that influence life expectancy in Italy, also shows similar evidence of inequality.

Research on the same topic in the Netherlands provides a comprehensive analysis that goes beyond effects solely at the mean (Bär et al., 2021). Through the analysis of differences between various age groups, they uncover evolving patterns in the dynamics of inequality, most notably pointing out a decline in inequality among younger age groups in contrast to an alarming rise among older individuals. The growing inequality among older age groups may result from two factors: either the wealthier individuals benefit more from recent healthcare advancements, or the inequality may shift from young to old groups due to the aging of those members. A third possible explanation for this result is found by Luo, Zhang, Jin, and Wang (2009), which observe inequality in the access to healthcare for the elderly, possibly explaining the rising inequality among older individuals as they generally rely relatively more on healthcare compared to younger individuals.

Research from Strozza et al. (2024) shows that the COVID-19 pandemic has sparked changes in life expectancy and its determinants. It is shown that post-COVID-19 life expectancies for both men and women in Denmark have relatively decreased in groups of lower socioeconomic status compared to those in higher socioeconomic status groups. Their results highlight the pandemic's unequal effect on lower socioeconomic groups and show the complex dynamics in the context of recent international events like the COVID-19 pandemic.

As previously noted, Bär et al. (2021) in their research on life expectancy determinants adhere to an empirical strategy, which looks beyond effects at the mean and provides insights into effect sizes across different age groups. A comparable methodology is a quantile regression as proposed by Koenker and Bassett Jr (1978), which looks at effect sizes across the distribution of the outcome variable, revealing information that ordinary regressions analysing effects at the mean would have missed. This well-established methodology, however, can be further improved by utilizing random forests and recursive feature elimination (Breiman, 2001; Zhou, Zhou, Zhou, Yang, & Luo, 2014) and as shown by Meinshausen and Ridgeway (2006) these techniques can be combined to improve the predictive accuracy of these quantile regressions.

1.2 Research Outline

The question central to this research is: *"What are the socioeconomic factors influencing life expectancy"*. This research direction provides insights into inequalities, which could also be present outside of solely life expectancy. The methods used for analysing these inequalities build on the quantile regression forests introduced by Meinshausen and Ridgeway (2006), which are adapted to the case of discrete outcome variables. Firstly, a benchmark is created for assessing the performance of these new models, for which widely used datasets contained in R are used to improve the comparability of the performance of these models. The research subsequently moves on to analysing socioeconomic life expectancy determinants using the previously tested models and microdata on individuals in the Netherlands from the Central Bureau of Statistics. Data ranging from 2015 - 2022 is used where effect sizes across multiple quantiles are estimated for the entire dataset and solely for the pre- and post-COVID-19 eras, providing insights into the effects of COVID-19 on life expectancy as well as inequality.

Firstly, it is shown that quantile regression forest based models outperform alternative methods. After this, the socioeconomic life expectancy determinants analysis is carried out, This analysis shows that economic determinants are the most important socioeconomic status category for life expectancy inequalities, however, there is a large uncertainty in predictions especially for lower ages of passing, highlighting the hardships in predicting an individual's life expectancy. Furthermore, the pre- and post-COVID analysis indicate that the importance of socioeconomic determinants has changed, showing decreasing importance for economic and social determinants, while healthcare is becoming increasingly important.

The remainder of this paper starts with an overview of the models employed for the analysis in Section 2. This is followed Sections 3 and 4, in which the data samples are presented and the results are discussed, respectively. Lastly, in Section 5 the main findings are summarized and discussed, and possible future research directions are presented.

2 Methodology

This section introduces the selection of models used in this research. Firstly, the quantile regression is introduced, followed by an explanation of the random forests and how these methods can be combined. After this, the recursive feature elimination algorithm is discussed as well as an adaptation on the quantile regression forest (Meinshausen & Ridgeway, 2006) to handle discrete outcome variables. Lastly, this section is concluded by presenting the model performance evaluation metrics.

2.1 Quantile Regression

Quantile regressions provide a more nuanced understanding of the relationships between predictor variables and outcomes (Yu, Lu, & Stander, 2003). Researchers can investigate potential heterogeneity in these effects more thoroughly and gain a greater knowledge of how different variables affect outcomes across different segments of the joint distribution of the data. This versatility has shown to be extremely beneficial in a variety of research fields, such as environmental studies, healthcare, and finance by Koenker (2017). Showing the ability of quantile regressions to find patterns that might otherwise go unnoticed when concentrating only on the mean.

The ability to create prediction intervals is an additional benefit of quantile regressions (Meinshausen & Ridgeway, 2006). Wider prediction confidence intervals indicate less precision in the prediction, offering important insights into the accuracy of predictions. This accuracy indicator helps researchers better understand the reliability of their forecasts and models.

Koenker and Bassett Jr (1978) introduced the quantile regression approach and the basis for these quantiles comes from the conditional cumulative distribution function of outcome Y:

$$F(y|X = x) = P(Y \leq y|X = x). \tag{1}$$

Consequently, the definition of the α-quantile is the probability being equal to α that outcome Y is smaller than $Q_\alpha(x)$ conditional on a fixed X:

$$Q_\alpha(x) = inf\{y : F(y|X = x) \geq \alpha\}. \tag{2}$$

The definition is for the case of a continuous outcome variable, as the distribution function needs to be continuous for the equation to hold. A second application of quantile regressions is that of constructing prediction intervals. Equation (3) displays a 95% prediction interval of outcome variable Y, which can be interpreted as Y being inside this interval at the 95% level for a given X:

$$I(x) = [Q_{0.025}(x), Q_{0.975}(x)]. \tag{3}$$

2.2 Random Forests

By using the principle of building multiple decision trees, random forests can predict results based on related explanatory data (Breiman, 2001). This methodology, which is closely related to boosting, has shown to be an effective tool for both regression and classification tasks. Nonetheless, unlike boosting, which makes use of data residuals from previously built trees, random forests use the raw data to create each tree (Meinshausen & Ridgeway, 2006). Meinshausen and Ridgeway (2006) provide an excellent example of combining random forests and quantile regressions to create a quantile regression forest. This novel method utilizes random forests' ability to determine the conditional distribution of outcome variables. Additionally, random forests can determine a range of quantiles with a single model, saving computational time, particularly for large datasets. An alternative for random forests such as XGBoost, for example, requires training a separate model for each quantile by design (Zhang, Quan, & Srinivasan, 2018).

Breiman (2001) shows how applying boosting to random forests can significantly improve performance, especially when it comes to improving predictive accuracy for larger datasets. Additionally, Granitto, Furlanello, Biasioli, and Gasperi (2006) shows how feature selection can be optimized by using feature importance rankings to reduce computation time and improve performance by limiting the number of features. It is concluded by Granitto et al. (2006) that recursive feature elimination works best when combined with random forest compared to other machine learning methods such as support vector machines. A possible explanation for this could come from the built-in ability of random forests to rank the importance of features in predicting outcomes.

Random forests in the case of a regression operate by averaging over a multitude of decision trees (Breiman, 2001). The decision trees are trained using a random selection of explanatory variables to split on and each tree uses a random subset of the data, called a bag. This research uses the notation used in Meinshausen and Ridgeway (2006) which is based on Breiman (2001). A tree is denoted by $T(\theta)$ and has corresponding leaves $l = 1, .., L$, where the vector θ contains the random split variable selection used in each node of the respective tree. B is defined as the space of X and B consists of multiple rectangular subspaces defined as R_l, which are linked to the leaves l. If $x \in B$ is dropped down tree $T(\theta)$ one and only one leaf is obtained where $x \in R_l$, this leaf is denoted by $l(x, \theta)$. To obtain a prediction from $T(\theta)$ for x a weighted average of the observed values at the corresponding leaf $l(x, \theta)$ is used. The Equation for the weights is provided in (4) and the corresponding prediction is shown in Equation (5):

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_j \in R_{l(x,\theta)}\}} \quad \text{for } i = 1, \ldots, n, \tag{4}$$

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta) Y_i. \tag{5}$$

X_i corresponds to an observation and the sum of the weights is constrained to one, i.e., $\sum_i^n w_i = 1$. Consequently, the prediction can thus be seen as the weighted average of Y_i given

$X = x$. Random forests use the previously discussed single-tree predictions by averaging over k trees. The Equations (6) and (7) display the determination of the weights and subsequent prediction, respectively:

$$w_i(x) = \frac{1}{k}\sum_{t=1}^{k} w_i(x, \theta_t) \quad \text{for } i = 1, \ldots, n, \tag{6}$$

$$\hat{\mu}(x) = E(Y|X = x) = \sum_{i=1}^{n} w_i(x)Y_i. \tag{7}$$

With $t = 1, .., k$ corresponding to the respective trees $T(\theta)$ and the vectors θ_t are independent and identically distributed. The prediction is an approximation of the conditional mean of the outcome $E(Y|X = x)$.

The above-provided definition of random forests can be used in combination with quantile regression to create Quantile Regression Forests (QRF) (Meinshausen & Ridgeway, 2006). Equations (8) and (9) show how the distribution function of Y can be rewritten:

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x), \tag{8}$$

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)1_{\{Y_i \leq y\}}. \tag{9}$$

The estimate of the conditional distribution $\hat{F}(y|X = x)$ can subsequently be used to create estimates for the quantiles.

2.3 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a backward selection method, that iteratively evaluates each feature's relevance and keeps only those that are thought to be the most explanatory for the outcome variable. By eliminating less important features, this iterative process helps to improve the model's predictive accuracy (Guyon, Weston, Barnhill, & Vapnik, 2002). RFE works best when combined with random forests, and does not require any (fine-)tuning. Random Forests Recursive Feature Elimination (RF-RFE) is an algorithm proposed by Zhou et al. (2014) and the algorithm is specified below.

Algorithm 1 Recursive Feature Elimination using Random Forest

```
Require: Explanatory training set $X$, outcome training set $Y$, and the initial model trained
    with all features included
Initialize lists $F \leftarrow$ {list of initial features}, $R \leftarrow$ {empty list to store the removed features}
while $F$ is not empty or a threshold for removing features is not reached do
    Update $X$ to only contain features specified in $F$
    Train a random forest model on $X$ and $Y$
    Calculate importance scores of the features in $F$
    Define $f_{least\_important}$ as the least important feature in $F$
    Append $f_{least\_important}$ to $R$
    Remove $f_{least\_important}$ from $F$
end while
return $R$
```

In Algorithm 1, it can be seen that one feature is removed in each iteration of the recursion leading to the final output R, which contains the removed features. Combining the RF-RFE algorithm with QRF leads to the Quantile Regression Forest with Recursive Feature Elimination (QRF-RFE) proposed by Zhang et al. (2018). This model uses the RF-RFE algorithm to select which features to exclude and uses the updated feature selection to construct a new QRF. The new QRF-RFE model is then subsequently used for constructing the quantiles in the same manner as in Section 2.2.

2.4 Discrete Quantile Regression

The quantile regression and quantile regression forests discussed in Meinshausen and Ridgeway (2006) concern a continuous outcome, however, little to no attention has been given to the case of discrete quantile regressions. Discrete quantile regression can be implemented with jittering, which involves adding a variable U to the outcome variable. This jittered variable generally has a uniform distribution between 0 and 1 (Carcaiso & Grilli, 2023), transforming discrete integers to decimals. This method produces a new continuous outcome variable with a bounded domain that can be utilized in continuous quantile regressions.

An alternative to jittering is proposed by Geraci and Farcomeni (2022) this approach adapts the original distribution function of the discrete outcome variable such that it can be used in quantile regressions. The adapted distribution is called the mid-distribution function. This method has been shown to outperform jittering-based approaches for multiple discrete variable types. Furthermore, as the quantile regression forests proposed in Meinshausen and Ridgeway (2006) utilize the distribution function of the outcome variable obtained from regression forests, these mid-distributions can be of great value in extending the quantile regression forests to predicting discrete outcomes.

The new quantile regression model using the mid-cumulative distribution function is named mid-Quantile Regression (mid-QR) and introduced by Geraci and Farcomeni (2022). The method uses the original cumulative distribution function to construct a new continuous cumulative distribution, however, for this research, the method is slightly altered to combine the mid-QR with QRF.

The conditional cumulative distribution function $G(y|X = x)$ is constructed as specified in Equation (10). This function consists of two parts with the first being the original conditional cumulative distribution function and the second being a down-shifting factor using the probability density function:

$$G(y|X = x) = F(y|X = x) - 0.5 * f(y|X = x). \tag{10}$$

Alternatively to what has been done in Geraci and Farcomeni (2022), this research proposes to use the distribution obtained by the previously discussed RF in Section 2.2. Using the $\hat{F}(y|X = x)$ in Equation (10), $\hat{G}(y|X = x)$ can be obtained and subsequently used in a QR in the same way as proposed for the QRF. To the best of our knowledge, this model has not been proposed in earlier works and is thus named mid-Quantile regression forests (mid-QRF) or mid-Quantile regression forests with Recursive Feature Elimination (mid-QRF-RFE) in case REF is applied.

2.5 Evaluation Metrics

To assess the quality and performance of the previously proposed models the same evaluation metrics, which are employed by (Meinshausen & Ridgeway, 2006), are used in light of comparability. The quantiles used in the analysis of the performance of the models are thus $\alpha \in \{0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995\}$. The loss function indicating the accuracy of the model is specified in Equation (11) in which q is the prediction of the model. The average result of this loss function is used in combination with a 5-fold cross-validation design for training and testing.

$$L_\alpha = \begin{cases} \alpha \mid y - q \mid & y > q \\ (1 - \alpha) \mid y - q \mid & y \leq q \end{cases} \tag{11}$$

Moreover, bootstraps are performed to create 95% confidence intervals of the average loss. These confidence intervals are subsequently used in assessing significant differences between the proposed models.

3 Data

In this section, the data samples used in this paper are presented. The first data sample discussed is used to assess the predictive performance of the employed models, while the second dataset is analysed to retrieve insights into socioeconomic life expectancy determinants.

3.1 Model Performance Analysis

To assess the performance of the models proposed in Section 2, this research firstly applies the models to the same datasets as in the paper of Meinshausen and Ridgeway (2006). The results retrieved from this analysis are used as a benchmark for the performance of the models. Meinshausen and Ridgeway (2006) use five datasets taken from the R packages *mlbench* and *alr3* (now updated to the *alr4* package) (R Core Team, 2023), however, only four of these are still available. The datasets included in the *mlbench* and *alr4* packages are listed below including a short description of the contents (Blake & Merz, 1998; Weisberg, 2014; Leisch & Dimitriadou, 2024).

- *BostonHousing*: Housing data from a census conducted in 1970 in Boston. The data contains 506 entries with 14 variables of which the target variable is *medv: median value of owner-occupied homes in USD 1000's*
- *Ozone*: Ozone pollution data from Los Angeles in 1976 containing 366 entries and 13 variables with the target variable being *Daily maximum one-hour-average ozone reading.*
- *BigMac2003*: Labor time data for the production of Big Macs containing 69 entries and 10 variables with the target variable being *BigMac: Minutes of labor to purchase a Big Mac.*
- *fuel2001*: Fuel consumption in the United States containing 51 entries and 6 variables, after combining *gallons sold* and *estimated miles driven* to create the target variable *average gas-mileage* (Meinshausen & Ridgeway, 2006).

3.2 Socioeconomic Life Expectancy Determinants

For analysing socioeconomic life expectancy determinants microdata from the Central Bureau of Statistics (CBS) is used. It must thus be noted that results obtained in this research are based on calculations by the author using non-public microdata from Statistics Netherlands. Under certain conditions, these microdata are accessible for statistical and scientific research[1].

The data used ranges from 2015-2022 and is on an individual level. The target variable is *Age of Passing* and as explanatory variables a set of variables indicating socioeconomic status (SES) are selected. SES variables can be classified into the five categories: (1) economic, (2) education, (3) environment, (4) social, and (5) healthcare (Rój & Jankowiak, 2021). The variables selected in this research for these respective categories are listed in Table 1.

[1] For further information: `microdata@cbs.nl`.

Table 1: Explanatory socioeconomic determinants by socioeconomic category

Category	**Variables**
Economic	(1) household wealth
Education	(1) proportion low education
	(1) proportion middle education
	(1) proportion high education
	(2) amount of primary school students
	(3) amount of secondary school students
Environment	(1) population density
	(2) nearby trade and catering
	(3) nearby government, education, and healthcare
	(4) nearby culture and attractions
Social	(1) household size
	(2) unemployment benefits
	(3) social assistance benefits
	(4) disability benefits
Healthcare	(1) chronically ill
	(2) social support act
	(3) long-term care
Control Variables	(1) gender (women)

The outcome variable *Age of Passing* is defined as the age an individual has reached in the year of passing. *Proportion* ***level*** *education*, *amount of* ***school type*** *students*, *nearby* ***facilities***, *population density*, *unemployment benefits*, *social assistance benefits*, and *disability benefits* are taken CBS StatLine, which provides macro-data from the Netherlands (Centraal Bureau voor de Statistiek, 2024). These variables are linked to individuals via their municipality of residence. The variables of the sort *nearby* ***facilities*** concern the number of facilities in an individual's respective municipality. Additionally, *unemployment benefits*, *social assistance benefits*, and *disability benefits* are the amount of individuals receiving such benefits in an individual's municipality. *Chronically ill* is a binary variable where individuals are classified as being chronically ill based on their medicine usage in combination with the index proposed by Huber, Szucs, Rapold, and Reich (2013). Additionally, the variables *social support act* and *long-term care* are also binary variables and indicate whether an individual receives care out of the Dutch care policies wet maatschappelijke ondersteuning and wet langdurige zorg, respectively.

3.2.1 Descriptives and Sample

The entire dataset used in this research is of the size $n = 1.390.975$ with the pre and post-COVID subsets being of the size $n = 659.454$ and $n = 563.836$, respectively. As the COVID pandemic started in 2019, the pre and post-COVID samples exclude individuals who have passed in 2019. Table 2 provides an overview of the average values of each variable for the respective samples. The full, pre-COVID, and post-COVID samples are very similar in their descriptives, however, there are some notable differences. These differences can be found in the variables on nearby facilities and benefits received, where it can be noted that there is a relative increase in nearby facilities for the post-COVID sample compared to the pre-COVID sample. Furthermore, the number of people receiving unemployment and social assistance benefits has decreased, while the amount of individuals receiving disability benefits has stayed constant. A possible explanation for these differences can be due to a more favourable economic climate post-COVID leading to more newly opened nearby facilities and less demand for unemployment and social assistance benefits.

Table 2: Sample averages for the full, pre-COVID, and post-COVID samples

Variable	**Full Sample**	**Pre-COVID**	**Post-COVID**
Age of Passing	78.52	78.35	78.79
nearby trade and catering	2765.87	2632.81	2977.39
nearby government, education, and healthcare	2440.02	2143.41	2885.37
nearby culture and attractions	2211.59	2051.58	2434.44
unemployment benefits	2326.15	2670.26	1891.96
social assistance benefits	5385.19	5605.16	5104.98
disability benefits	5780.25	5787.97	5762.75
proportion low education	29.07	30.20	27.30
proportion middle education	41.95	41.86	42.19
proportion high education	28.98	27.94	30.52
population density	1561.07	1556.19	1560.67
amount of primary school students	11068.06	11124.18	10935.21
amount of secondary school students	6900.95	6889.53	6877.15
gender (women)	0.51	0.51	0.50
household wealth	232757473.8	231789162.3	234499508.8
household size	1.89	1.87	1.94
long-term care	0.33	0.33	0.35
social support act	0.22	0.21	0.24
chronically ill	0.74	0.74	0.73
sample size	1390975	659454	563836

Due to the computational complexity of the models proposed in Section 2, a random subset of the datasets of size $n = 10.000$ is taken from the non-missing value entries of the entire sample. The descriptive statistics of these subsets are very similar to those of the entire datasets and are provided in Appendix A.

4 Results

The results presented in this section consist of two parts: a performance analysis and a life expectancy determinants analysis. All computations are conducted in R (R Core Team, 2023) and an overview of the performed runs is contained in Appendix B. R version 4.3.0 is used in combination with the *quantregForest* package (Meinshausen, 2017).

4.1 Model Performance Analysis

The datasets presented in Section 3.1 are used to investigate the performance of the QRF and QRF-RFE models. A selection of benchmark models which are also used in Meinshausen and Ridgeway (2006) are compared with the random forest based models. Linear Quantile Regression (LQR), linear quantile regression with interactions (QQR), and three Regression

Tree based models with a second-degree polynomial without interactions (TRP), multiple linear terms (TRM), and solely a constant term (TRC) make up the benchmark models. The QQR model is constructed starting from the LQR model and adding interaction terms via forward selection until the model does not improve any further. Based on the loss function defined in Section 2.5, a 95% confidence interval of the average loss of the predictions is constructed for each model and their quantiles. For every model, 100 bootstraps are performed in order to determine these 95% confidence intervals. The seven quantiles used in this analysis are $\alpha \in \{0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995\}$. For training and testing the models, a 5-fold cross-validation split is used. This entails that the model is trained using 80% of the available data, and predictions are made using the remaining 20% of the data. Subsequently, the average loss is computed using these predictions. For the random forest based models the *mtry* parameter is maintained at its default value, and the number of trees used in the QRF and QRF-RFE models is fixed at 1000. Furthermore, the threshold for the RFE is set at 75% meaning that 25% of the explanatory variables are removed.

Figure 1 shows the 95% confidence intervals of the employed models. It can be seen that the TRC, TRM, and TRP models are excluded from this figure and the figure containing these models is presented in Appendix A. The TRC, TRM, and TRP models are excluded as the confidence intervals for these models are much higher than the other models and this leads to the confidence intervals of the other models not being visible on the graph. Normally this indicates that the TRC, TRM, and TRP models are largely outperformed, however, closer inspection of the loss calculations reveals a different result. Due to a lack of data availability at the leaf nodes, some leaf models can not be estimated sufficiently to provide reliable results leading to some predictions having a loss comparable to that of the other models, while other predictions lead to a loss of over 100 significantly increasing the average loss.

Inspection of Figure 1 reveals that the QRF and QRF-RFE models outperform the LQR and QQR models on all occasions, as the confidence intervals of these two sets of models do not overlap. The LQR and QQR models have relatively similar performance and are not significantly different in most cases, however, the LQR can on some occasions perform close to the QRF and QRF-RFE models, as can be seen for the very extreme quantiles of the *fuel2001* dataset. Furthermore, it becomes clear from the figure that the QRF and QRF-RFE models are very similar in performance with only a significant difference in performance for the 0.95 quantile of the *fuel2001* dataset, where the QRF-RFE outperforms the QRF.

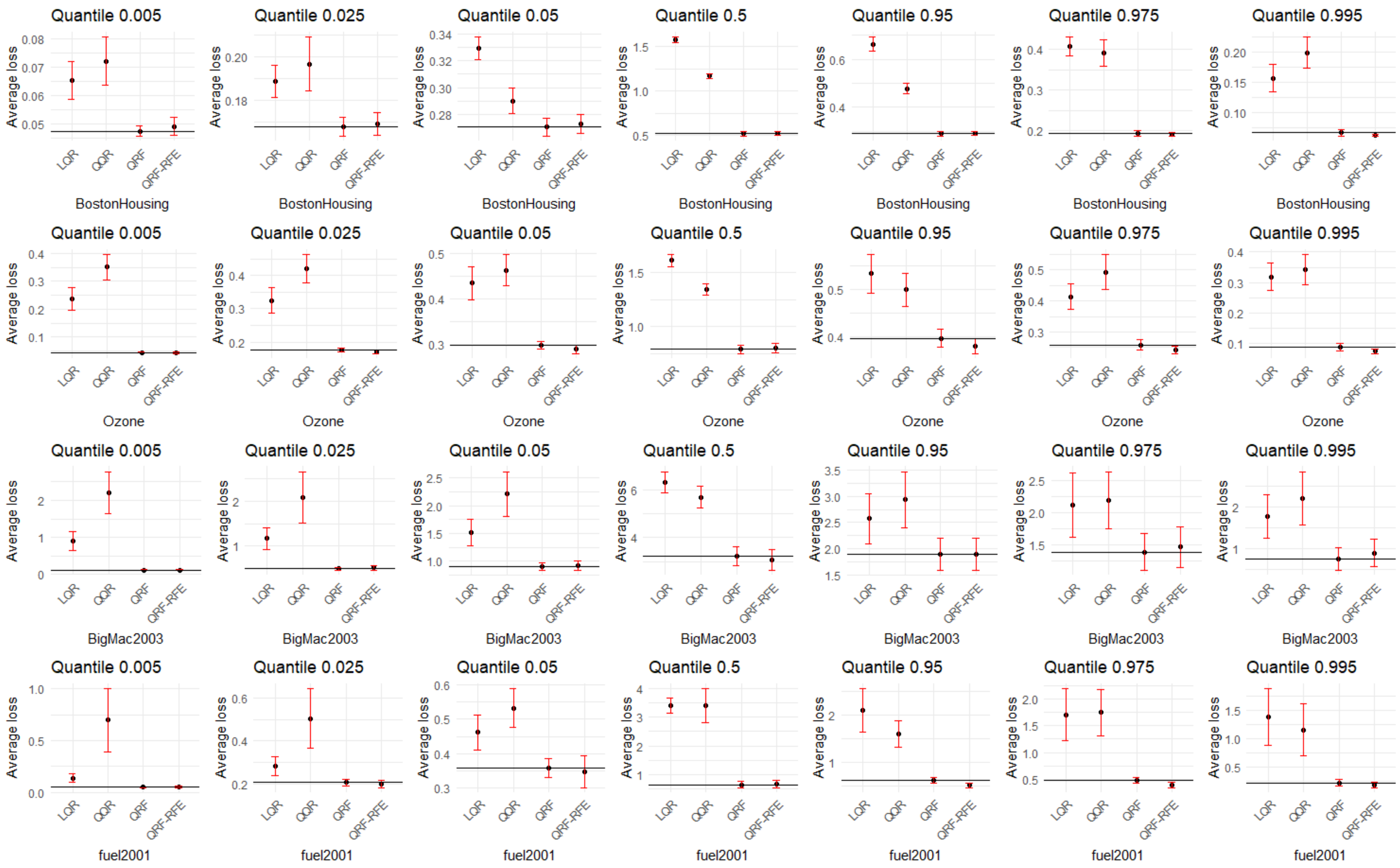

Figure 1: 95% Confidence intervals of the average loss for the inspected datasets, models, and quantiles

To analyse the performance of the random forest based models in the discrete outcome variable setting the previously used datasets are transformed to have a discrete outcome variable. The datasets are discretized using truncation, where the values are cut off at the decimal point leaving only integers. These transformed datasets are then subsequently used by the QRF, QRF-RFE, mid-QRF, and mid-QRF-RFE models to make predictions. Via the same method as previously described, these predictions are used in constructing the 95% confidence intervals for the average loss, as can be seen in Figure 2.

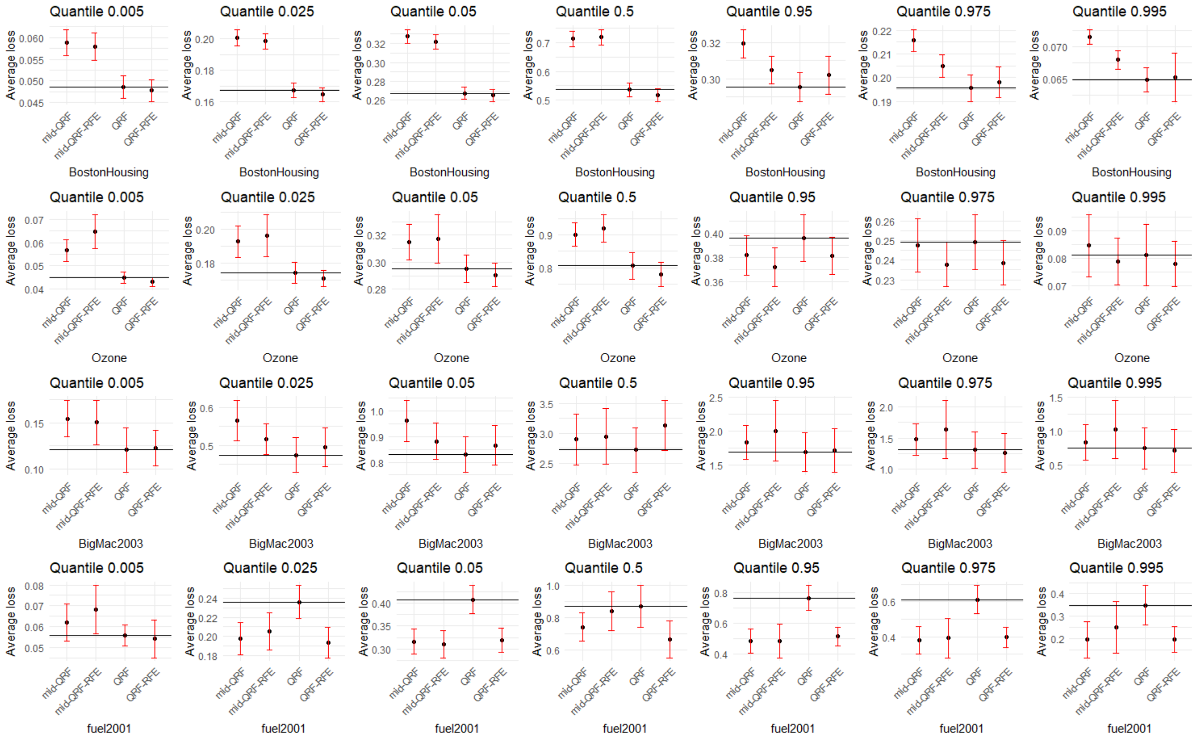

Figure 2: 95% Confidence intervals of the average loss for the employed datasets and quantiles in the discrete outcome variable case

Looking into the performance of the models for the discrete outcome variable case Figure 2 reveals that the four models have on average a similar performance. The models do not differ significantly on most occasions, however, the mid-QRF and mid-QRF-RFE models are outperformed by the QRF and QRF-RFE in some cases as can be seen for the *BostonHousing* dataset. On the other hand, the mid-QRF and mid-QRF-RFE models outperform the QRF model for the *fuel2001* dataset, while not being significantly better than the QRF-RFE model. In conclusion, There thus seems to be no generally outperforming model with slight variations in performance for different datasets and quantiles.

4.2 Socioeconomic Life Expectancy Determinants

To investigate socioeconomic life expectancy determinants, the dataset presented in Section 3.2 is analysed using the QRF, QRF-RFE, mid-QRF, and mid-QRF-RFE models. Similar to the model performance analysis, these models use a 5-fold cross-validation split with 1000 trees and *mtry* set at the default value. Additionally, the threshold for the RFE is kept at 75% and the quantiles investigated are $\alpha \in \{0.005, 0.025, 0.05, 0.25, 0.5, 0, 75, 0.95, 0.975, 0.995\}$.

Figure 3 provides an overview of the average loss of the predictions for the employed models on the full sample. The QRF, QRF-RFE, and mid-QRF models are very similar in predictive accuracy with the mid-QRF-RFE being outperformed slightly. Further research is, however, necessary to determine if this difference in performance is also statistically significant. Looking into performance differences across quantiles it becomes evident that the predictive accuracy of all the models is higher for the higher quantiles as indicated by the lower observed average losses. It can thus be concluded that life expectancies at the top of the outcome variable distribution can be better predicted, possibly due to more data availability. The average losses of the pre- and post-COVID models show similar results and are presented in Appendix A.

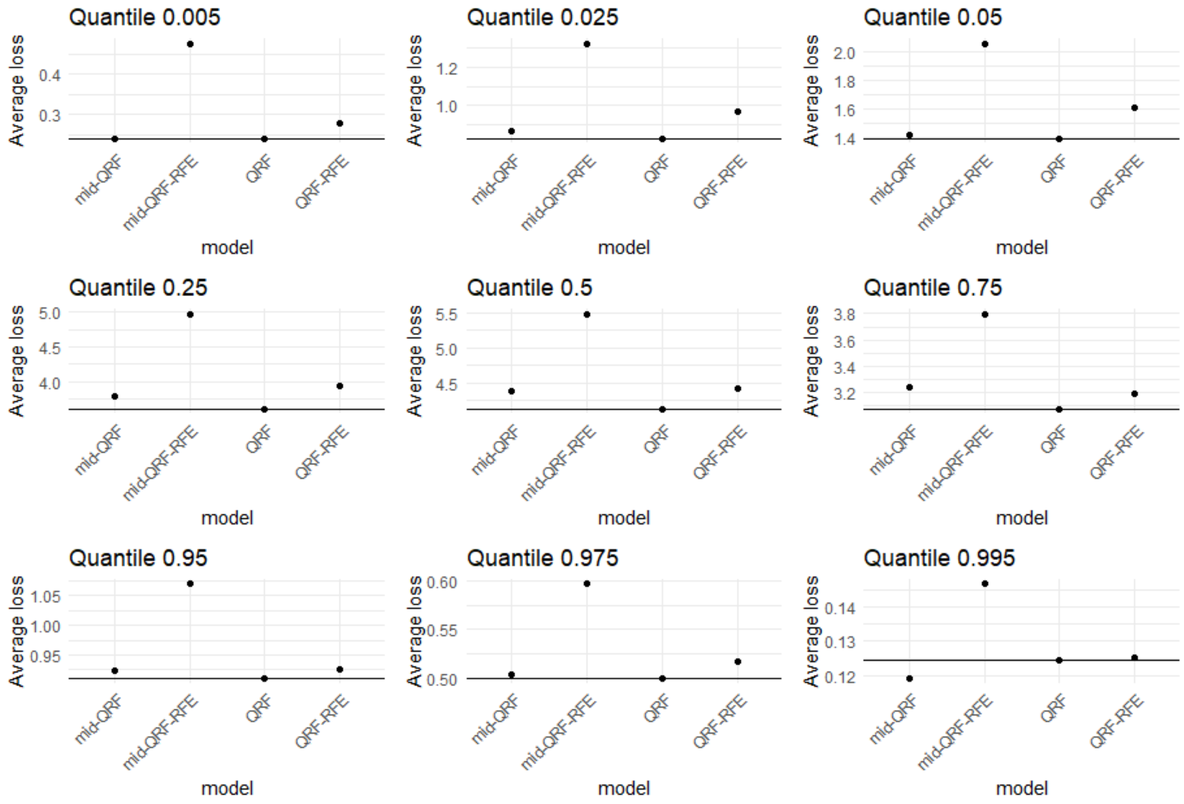

Figure 3: The average loss of the random forest based models on the total life expectancy dataset

The 95% prediction confidence intervals for the full sample are presented in Figure 9 where the actual outcomes are ordered from small to large with their respective confidence interval. The results for the pre- and post-COVID samples are again similar to the full sample results and are thus presented in Appendix A. The lowest outcomes lie outside of their prediction confidence interval on a relatively large number of occasions compared to the higher outcomes, showing the uncertainty in predicting low ages of passing. Furthermore, it can be observed that the ranges of the confidence intervals relatively decrease for higher outcomes indicating that these values can be predicted more precisely. However, the confidence interval ranges do not differ much throughout the outcome distribution, highlighting the uncertainty in predicting life expectancy.

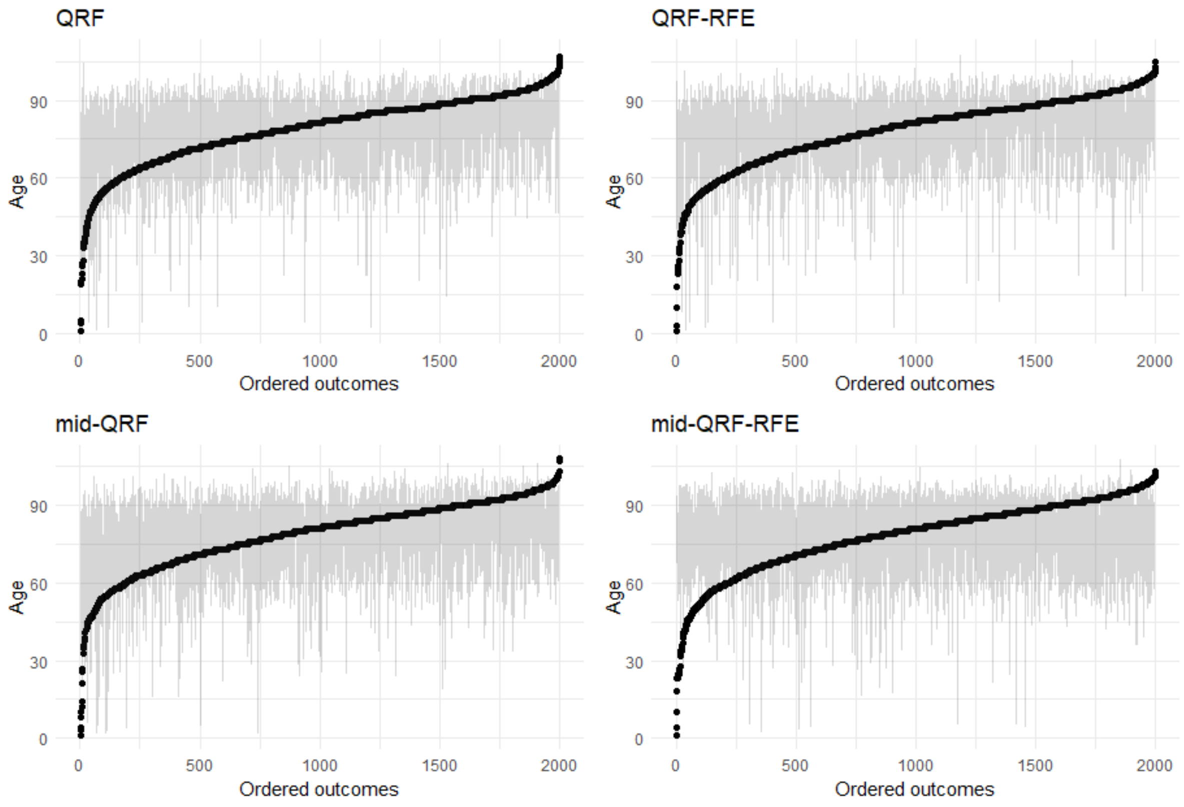

Figure 4: The 95% prediction confidence intervals of the random forest based models on the total life expectancy dataset

Table 3 shows the importance of the socioeconomic variables in making predictions of the QRF model. The QRF model is selected to analyse the socioeconomic determinants of life expectancy, as this model has the lowest average loss on most occasions. The importance estimates for the full sample reveal that household wealth is the most important feature in determining life expectancy. followed by long-term care, household size, and population density. Furthermore, variables related to education level and social benefits are relatively more important than variables related to facilities in the nearby environment. This thus shows that economic determinants are most important in explaining life expectancy inequalities, while environmental determinants are relatively unimportant. Education and social determinants are approximately equal in importance. The high estimates of *household size* and *population density* highlight the importance of an individual's social network in growing old, while *long-term care* indicates the importance of healthcare for a higher life expectancy.

The pre- and post-COVID importance estimates displayed in Table 3 are quite similar at first glance, however, some important differences can be seen on closer inspection. *Household wealth* has become relatively less important post-COVID, while the variables related to healthcare have become more important. Furthermore, *household size* and *population density* are relatively more important pre-COVID. These differences indicate a shift in the socioeconomic determinants of life expectancy due to COVID-19, where inequalities in life expectancy are relatively more likely to arise due to differences in received healthcare provisions. Additionally, economic determinants as well as an individual's social network have become less important post-COVID.

Table 3: Importance estimates (scaled to 100) of the QRF model for the full, pre-COVID, and post-COVID samples (percentages)

Variable	**Full Sample (%)**	**Pre-COVID (%)**	**Post-COVID (%)**
Household wealth	17.48	20.29	18.76
Household size	11.10	12.41	12.06
Long-term care	13.84	12.36	13.19
Population density	5.20	5.15	5.01
Proportion low education	4.62	4.45	4.39
Proportion middle education	4.52	4.42	4.46
Unemployment benefits	4.45	4.11	3.96
Proportion high education	4.39	4.26	4.20
Social assistance benefits	4.33	3.92	3.95
Disability benefits	4.21	3.85	3.92
Amount of secondary school students	3.93	3.50	3.49
Amount of primary school students	3.91	3.67	3.56
Nearby government, education, and healthcare	3.86	3.64	3.55
Nearby trade and catering	3.81	3.57	3.86
Nearby culture and attractions	3.74	3.43	3.41
Chronically ill	2.02	2.44	3.46
Social support act	1.93	1.83	1.90
Gender (women)	2.67	2.71	2.85

5 Conclusion

The research question central to this study is: *"What are the socioeconomic factors influencing life expectancy"*. To answer this question this research builds upon the QRF as proposed in Meinshausen and Ridgeway (2006) to uncover insights into the importance of socioeconomic determinants in life expectancy inequalities. A sample of $n = 1.390.975$ individuals from the Netherlands who passed away between 2015-2022 is employed for this cause, where not only the full sample but also the pre- and post-COVID subsamples are analysed, providing insights into changes in socioeconomic inequalities due to COVID-19.

This research starts by looking into the performance of the QRF model as well as the newly proposed QRF-RFE model by comparing the predictive accuracy of these models to the benchmark models: LQR, QQR, TRC, TRM, TRP. These models are compared for multiple quantiles and multiple datasets contained in R packages *mlbench* and *alr4* (Blake & Merz, 1998; Weisberg, 2014; R Core Team, 2023; Leisch & Dimitriadou, 2024). It is shown that the random forest based models outperform the benchmarks on all occasions, while not significantly differing from each other. Subsequently, the mid-QRF and mid-QRF-RFE models are introduced to be used in the case of a discrete outcome variable, however, the four random forest based models showed similar predictive accuracies across all quantiles and datasets. This result thus indicates the robustness of the QRF model in the case of discrete outcome variables.

After the performance analysis, the QRF, QRF-RFE, mid-QRF, and mid-QRF-RFE models are used in predicting life expectancy using socioeconomic determinants. The mid-QRF-RFE model is slightly outperformed by the other models in this case, however, recursively selecting the least relevant feature might run into problems when there are a lot of correlated variables in high-dimensional datasets (Darst, Malecki, & Engelman, 2018), underscoring the need for additional research into the statistical significance of performance differences of the models.

Furthermore, it is shown that the accuracy of predictions is relatively higher for outcomes at the higher part of the distribution, while predictions at the lower part of the outcome distribution are more likely to be wrong. Overall, however, the 95% prediction confidence intervals illustrate the challenges in predicting life expectancy as these intervals are shown to be quite large.

The importance of the socioeconomic determinants in making life expectancy predictions shows that economic determinants are the largest contributor to life expectancy inequalities for both the pre- and post-COVID cases. However, it can be seen that there is a shift in the respective contribution of the socioeconomic variables, with economic determinants becoming less important post-COVID. Furthermore, healthcare-related determinants have shown to be more important post-COVID while an individual's social network has become less important. These results illustrate how COVID-19 has fundamentally changed the socioeconomic inequalities in life expectancy and highlight the need for additional research into the effects of COVID-19 on our society.

5.1 Future research

The results of this research are promising, however, due to the usage of random forests effect sizes can not be estimated. Future research can improve upon this by for example also employing the LQR and QQR model. While these models are shown to be outperformed in predictive accuracy, they can provide insights into not only the most important socioeconomic determinants but also the magnitude of the effect these determinants have on life expectancy. Furthermore, due to data availability education level and social benefits are retrieved on the municipality level instead of the individual level. Incorporating individual education levels as well as the individual usage of social benefits can provide a better understanding of the effects these determinants have on life expectancy. Subsequently, looking specifically into the lower ages of passing can help improve the predictive accuracy of these outcomes and provide insights into life expectancy inequality differences for younger and older individuals.

Bibliography

Bär, M., Wouterse, B., Riumallo Herl, C., Van Ourti, T., & Van Doorslaer, E. (2021). Diverging mortality inequality trends among young and old in the netherlands. *Fiscal Studies*, *42*(1), 79–101.

Blake, C. L., & Merz, C. J. (1998). Uci repository of machine learning databases [Computer software manual]. Irvine, CA. (Formerly available from `http://www.ics.uci.edu/ mlearn/MLRepository.html`)

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Carcaiso, V., & Grilli, L. (2023). Quantile regression for count data: jittering versus regression coefficients modelling in the analysis of credits earned by university students after remote teaching. *Statistical Methods & Applications*, *32*(4), 1061–1082.

Centraal Bureau voor de Statistiek. (2024). *Netherlands macroeconomic data.* https://opendata.cbs.nl/statline//CBS/en/. (Accessed: 2024-06-07)

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, *19*, 1–6.

Deaton, A. S., & Paxson, C. (2001). Mortality, education, income, and inequality among american cohorts. In *Themes in the economics of aging* (pp. 129–170). University of Chicago Press.

Geraci, M., & Farcomeni, A. (2022). Mid-quantile regression for discrete responses. *Statistical Methods in Medical Research*, *31*(5), 821–838.

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83–90.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389–422.

Huber, C. A., Szucs, T. D., Rapold, R., & Reich, O. (2013). Identifying patients with chronic conditions using pharmacy data in switzerland: an updated mapping approach to the classification of medications. *BMC public health*, *13*, 1–10.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual review of economics*, *9*, 155–176.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lallo, C., & Raitano, M. (2018). Life expectancy inequalities in the elderly by socioeconomic status: evidence from italy. *Population health metrics*, *16*, 1–21.

Leisch, F., & Dimitriadou, E. (2024). mlbench: Machine learning benchmark problems [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mlbench (R package version 2.1-5)

Luo, J., Zhang, X., Jin, C., & Wang, D. (2009). Inequality of access to health care among the urban elderly in northwestern china. *Health Policy*, *93*(2-3), 111–117.

Meinshausen, N. (2017). quantregforest: Quantile regression forests [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=quantregForest (R package version 1.3-7)

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, *7*(6).

Mirowsky, J., & Ross, C. E. (2000). Socioeconomic status and subjective life expectancy. *Social Psychology Quarterly*, 133–151.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rój, J., & Jankowiak, M. (2021). Socioeconomic determinants of health and their unequal distribution in poland. *International Journal of Environmental Research and Public Health*, *18*(20), 10856.

Sanzenbacher, G. T., Webb, A., Cosgrove, C. M., & Orlova, N. (2021). Rising inequality in life expectancy by socioeconomic status. *North American Actuarial Journal*, *25*(sup1), S566–S581.

Strozza, C., Vigezzi, S., Callaway, J., & Aburto, J. M. (2024). The impact of covid-19 on life expectancy across socioeconomic groups in denmark. *Population Health Metrics*, *22*(1), 3.

Weisberg, S. (2014). *Applied linear regression* (Fourth ed.). Hoboken NJ: Wiley. Retrieved from http://z.umn.edu/alr4ed

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D: The Statistician*, *52*(3), 331–350.

Zhang, W., Quan, H., & Srinivasan, D. (2018). Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. *Energy*, *160*, 810–819.

Zhou, Q., Zhou, H., Zhou, Q., Yang, F., & Luo, L. (2014). Structure damage detection based on random forest recursive feature elimination. *Mechanical Systems and Signal Processing*, *46*(1), 82–90.

A Additional Tables and Figures

Table 4: Sample averages for the full, pre-COVID, and post-COVID random sub-samples

Variable	Full Sample	Pre-COVID	Post-COVID
Age	78.59	78.35	78.76
Nearby trade and catering	2855.70	2614.60	3024.16
Nearby government, education, and healthcare	2520.71	2131.28	2929.92
Nearby culture and attractions	2288.39	2045.24	2485.06
Unemployment benefits	2427.35	2643.79	1919.51
Social assistance benefits	5651.61	5531.56	5192.25
Disability benefits	5985.49	5743.40	5847.57
Proportion low education	28.95	30.17	27.38
Proportion middle education	41.85	41.84	42.19
Proportion high education	29.21	27.98	30.44
Population density	1592.79	1549.59	1571.33
Amount of primary school students	11450.04	11028.14	11084.53
Amount of secondary school students	7121.81	6832.52	6960.59
Gender (women)	0.50	0.51	0.50
Household wealth	250190913.7	180180682.5	180228037.3
Household size	1.90	1.87	1.93
Long-term care	0.33	0.32	0.34
Social support act	0.22	0.22	0.24
Chronically ill	0.74	0.74	0.73
sample size	10000	10000	10000

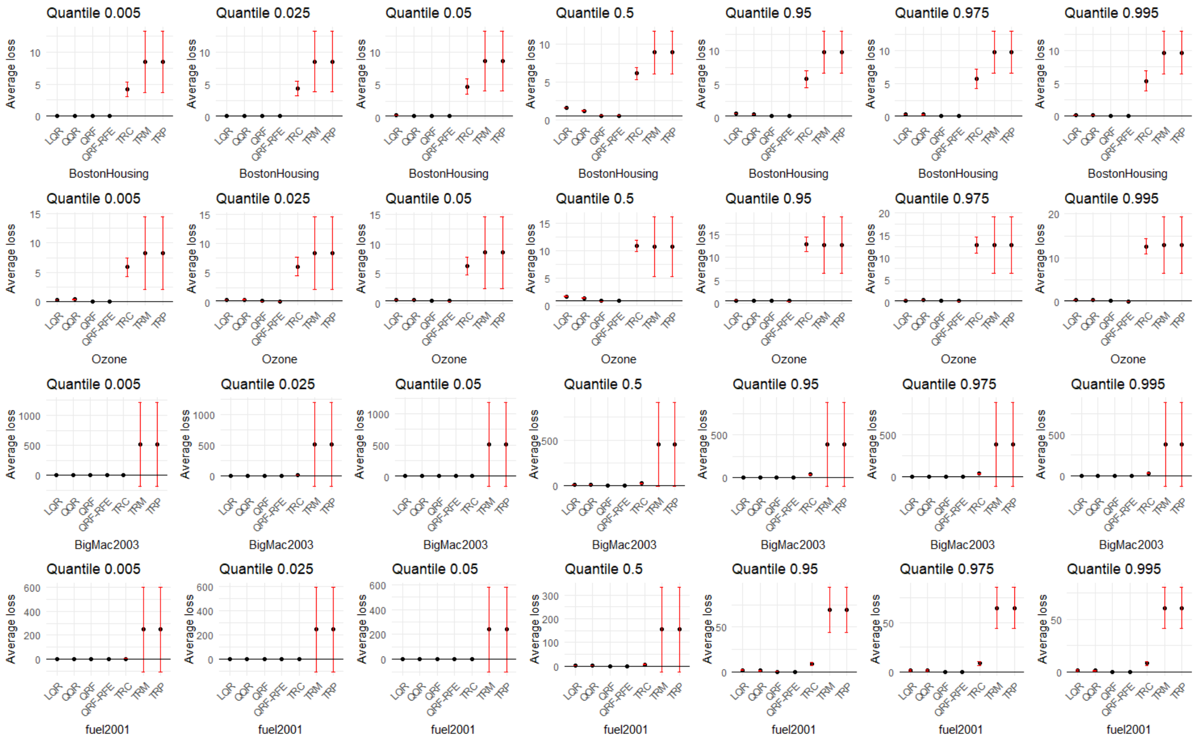

Figure 5: 95% Confidence intervals of the average loss for all datasets, models, and quantiles

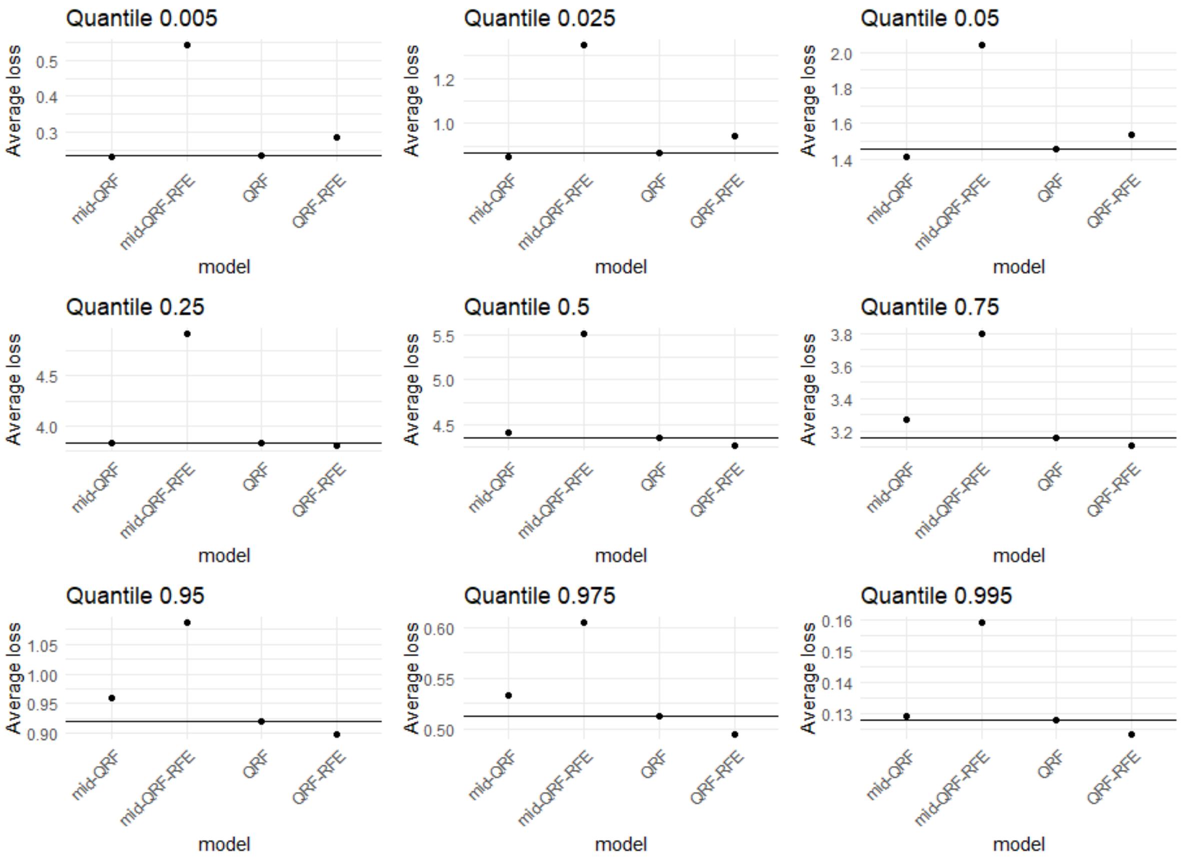

Figure 6: The average loss of the random forest based models on the pre-COVID life expectancy dataset

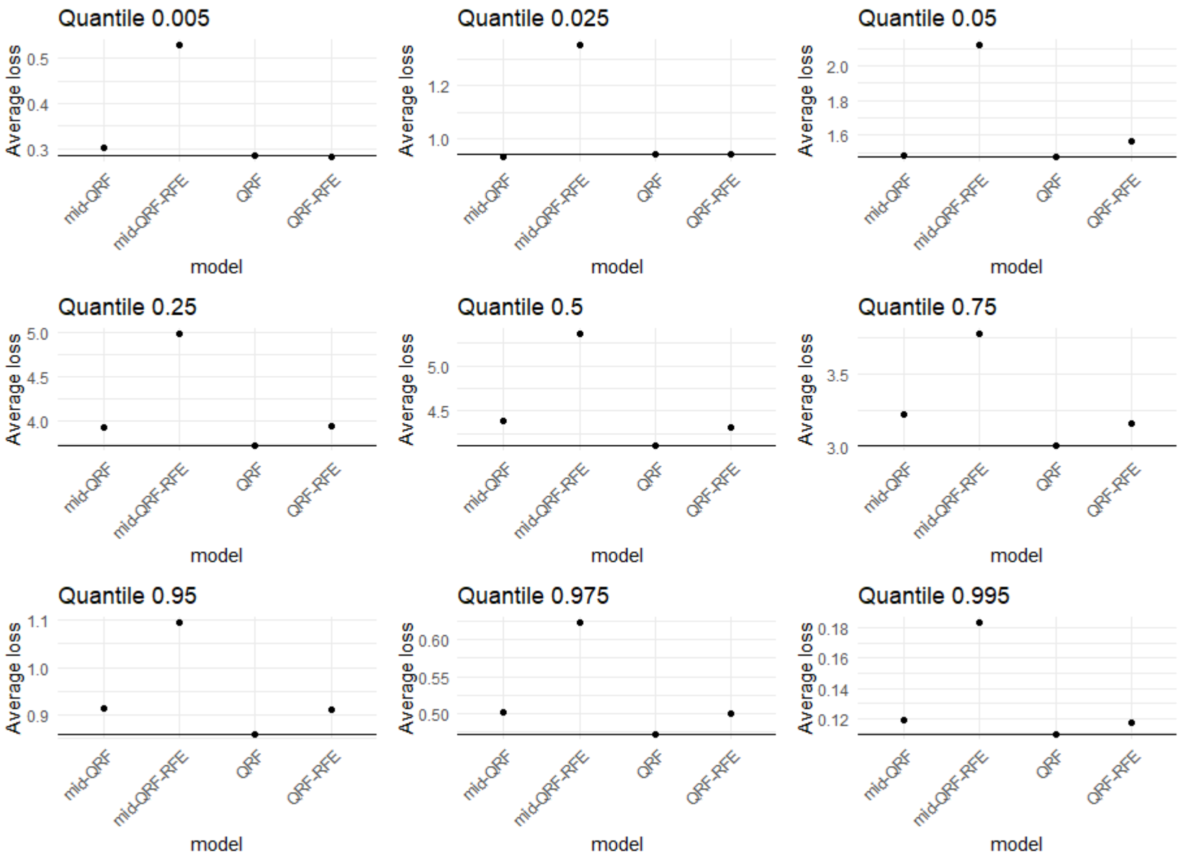

Figure 7: The average loss of the random forest based models on the post-COVID life expectancy dataset

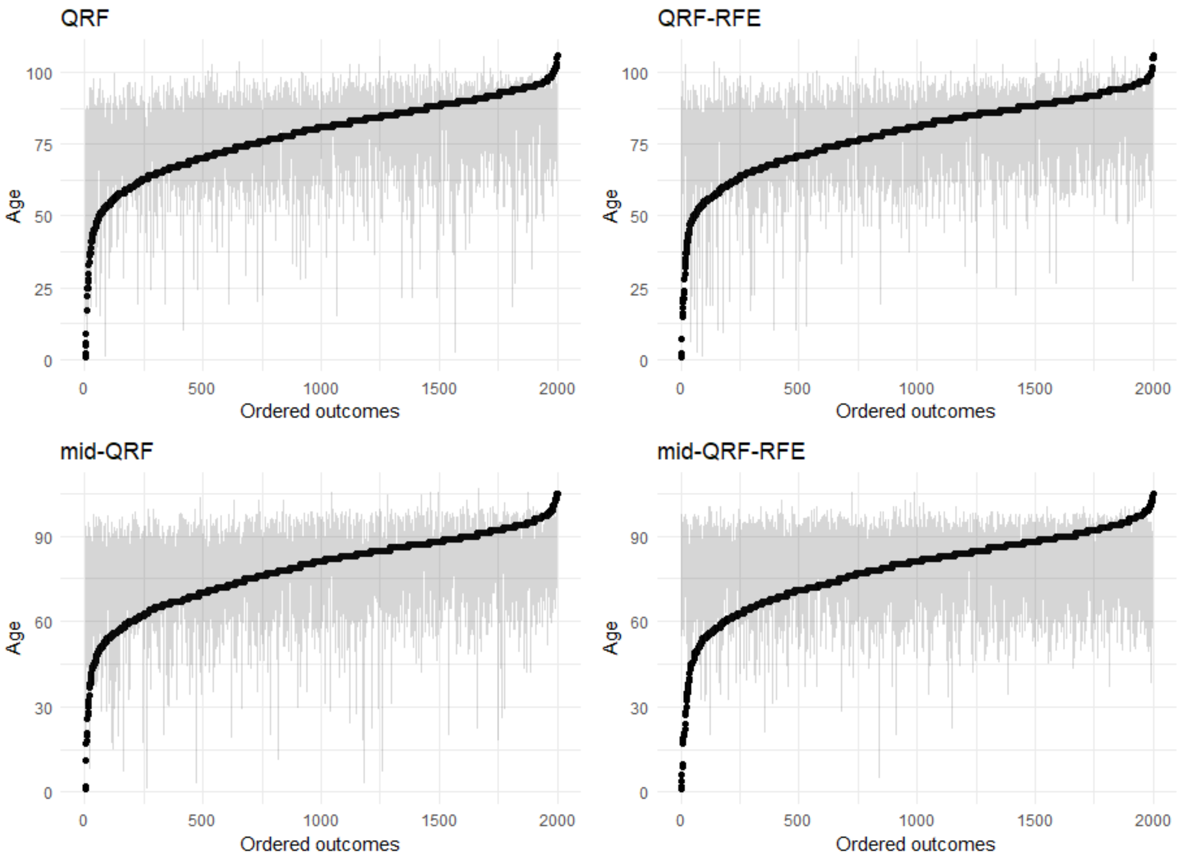

Figure 8: The 95% prediction confidence intervals of the random forest based models on the pre-COVID life expectancy dataset

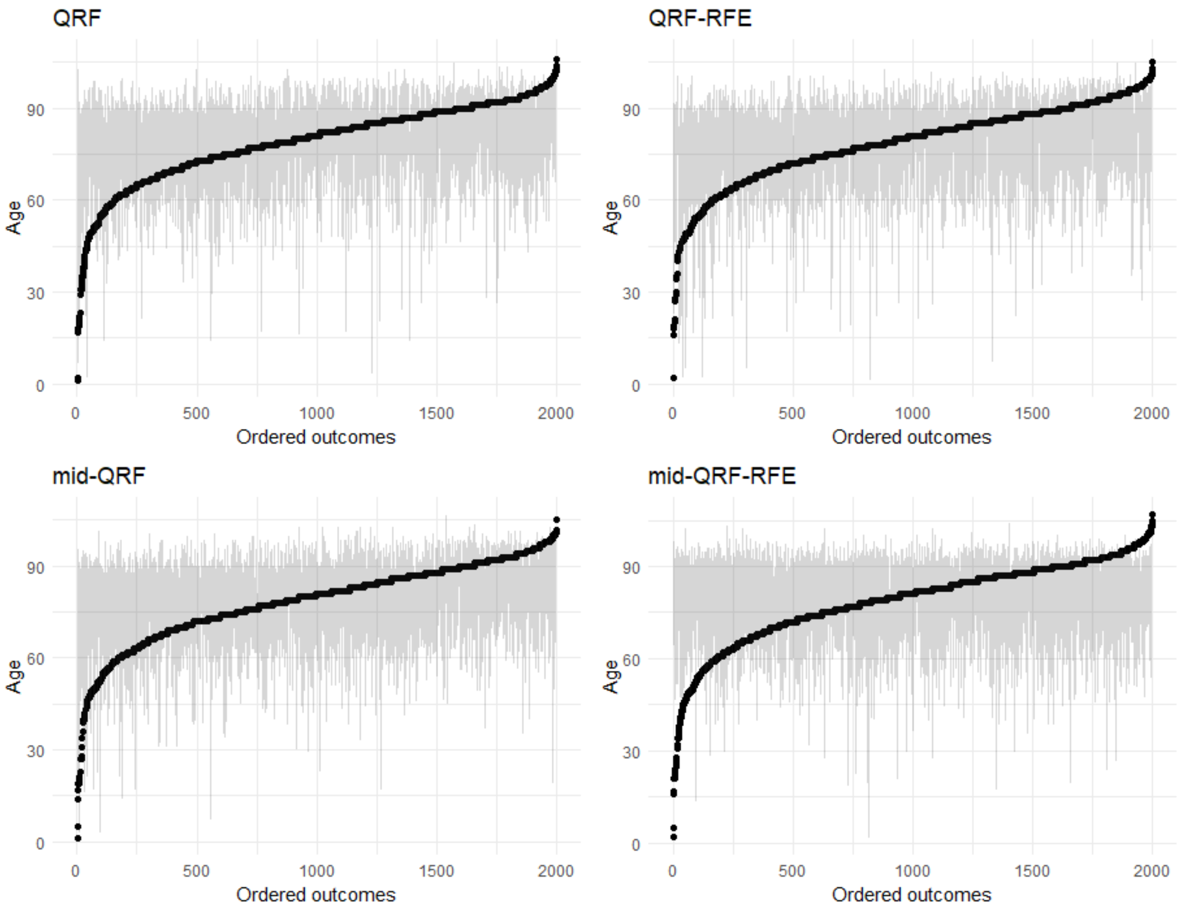

Figure 9: The 95% prediction confidence intervals of the random forest based models on the post-COVID life expectancy dataset

B Programming Overview

An overview of the programming runs performed during this research is presented in this section. R version 4.3.0 is used in combination with the *quantregForest* package (Meinshausen, 2017), which contains quantile regression functionalities. Table 5 contains an overview of the code files with their respective result section. The files with the name *ReplicationCode_.....* all run the respective model on the respective dataset as specified in the name, ultimately providing the average loss confidence intervals shown in Section 4.1. Additionally, *Replication_plot.R* and *save_results_code.R* are used to provide an interpretable overview of the model performance analysis results. The remainder of the code files is used for the results in Section 4.2 where the distinction between 3 file types can be made: (1) code files for creating the data sample, (2) code files for running the models, and (3) code files for generating the figures and descriptive statistics presented in this paper.

Table 5: Overview of the analysis code files

Filename	**Section**
address dataframe.R	4.2
descriptives.R	4.2
df_household.R	4.2
df_household_wealth_link.R	4.2
Life expectancy dataframe.R	4.2
mid-QRF-RFE.R	4.2
mid-QRF.R	4.2
QRF-RFE.R	4.2
QRF.R	4.2
ReplicationCode_LQR_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_LQR_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_LQR_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_LQR_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_mid-QRF-RFE_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_mid-QRF-RFE_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_mid-QRF-RFE_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_mid-QRF-RFE_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_mid-QRF_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_mid-QRF_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_mid-QRF_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_mid-QRF_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_QQR_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_QQR_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_QQR_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_QQR_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_QRF-RFE_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_QRF-RFE_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_QRF-RFE_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_QRF-RFE_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_QRF_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_QRF_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_QRF_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_QRF_ConfidenceInterval_Ozone.R	4.1
ReplicationCode_TRC_TRM_TRP_ConfidenceInterval_BigMac2003.R	4.1
ReplicationCode_TRC_TRM_TRP_ConfidenceInterval_BostonHousing.R	4.1
ReplicationCode_TRC_TRM_TRP_ConfidenceInterval_fuel2001.R	4.1
ReplicationCode_TRC_TRM_TRP_ConfidenceInterval_Ozone.R	4.1
Replication_plot.R	4.1
result_plots_tables.R	4.2
save_results_code.R	4.1