Bachelor's Thesis

Bayesian neural networks for age-period-cohort models

Department of Statistics
Ludwig-Maximilians-University Munich

by Markus Ewert

supervised by Prof. Dr. Volker Schmid
Munich, January 28, 2021

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und dabei keine anderen als die angegebenen Hilfsmittel benutzt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde bisher weder gesamt noch in Teilen einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, 28.01.2021 .....

Ort, Datum

M. Ewert .....

Markus Ewert

Abstract

The area of machine learning and especially deep learning had its breakthrough in the past years. Researchers in this area provided empirical evidence for the success of those methodologies. For this reason, combining methods from this field with Age-Period-Cohort models would be an interesting approach. Therefore, the goal of this thesis is to analyze how those models could profit from using *Bayesian Neural Networks*. It presents two conceptual approaches for this purpose and conducts computational experiments to obtain an impression of their the empirical performance. Thereby, it investigates this from two points of view - the concepts should be able to predict future incidences and explain the influence of a single variable on the variable of interest. While the experimental section provides empirical evidence that a *Bayesian Neural Network* has a superior predictive performance compared to established APC models, an ensemble method combining both model types fails to show the second subject of analysis, which is the explanation of the effects. Therefore, this thesis addresses also open challenges, which solutions would improve the performances of those models.

Keywords: *Bayesian Neural Network*, *Age-Period-Cohort Models*, *Cohort Analysis*

Contents

List of Figures

List of Tables

1 Introduction

Many researchers in social and behavioral sciences face the challenge of investigating unobservable phenomenons, and thus, they are unable to measure those, which makes an analysis often infeasible. Therefore, they utilize measurable variables as surrogates to get an impression of the latent ones (Hobcraft, Menken, & Preston, 1985). The cohort analysis, which analyzes age, period, and cohort effects, is a commonly used example for such a method in various research fields (e.g., Lopez et al., 2006; W. Mason & Wolfinger, 2001). While inferring from a surrogate to the real effect of a variable is generally not trivial, this method suffers in addition to that also from the identification problem (O'Brien, 2011; Rodgers, 1982): Each of those variables can be described as a linear combination of the remaining two preventing the identifiability of the real effects. However, the estimation of those effects is necessary for further inference purposes. Thus, many researchers developed methods to cope with this problem (e.g., V. Schmid & Held, 2007; Yang & Land, 2006; Yang, Schulhofer-Wohl, Fu, & Land, 2008), however, there does not exist an approach that solves it. A more recent stream in the literature combines established models with methods from other research areas such as deep learning (Breeden & Leonova, 2019).

1.1 Problem Statement and Research Questions

The goal of this thesis is to investigate the use of a *Bayesian Deep Neural Network* (BDNN) could improve the estimation of *Age-Period-Cohort* (APC) effects using established APC models. Even though some of these models could profit from incorporating a BDNN directly into their estimation processes, for instance, it could substitute the principal component regression in the *intrinsic estimator* (Yang et al., 2008), it would not enhance the estimation of the effects because there is no theoretical justification stating that a BDNN can solve the identification problem. For this reason, this thesis limits itself in combing both approaches from an external perspective only and treats them as black-boxes. Thus, the concepts in this thesis are generally applicable to various contexts and not limited to a specific APC model and its restrictions.

As many researchers proposed methods for solving the identification problem in APC models, these proposals simultaneously created a discussion about the estimability of the effects. A central result of this discussion is that only the non-linear effects of the variables are estimable (Rodgers, 1982). Since neural networks are non-linear models (Hastie, Tibshirani, & Friedman, 2009), they depict a potential alternative to established APC models that mostly belong to the family of linear regression models. Moreover, the Bayesian character of this model class allows to quantify the uncertainty of its estimation leading to additional insights into the analysis. The primary requirement for being an alternative approach is the ability to estimate the real APC effects correctly. It is also necessary to develop a method that extracts those from the model in an interpretable manner for further inference purposes. Moreover, the resulting model should utilize the estimated effects to predict future cases of the target variable. Taken together, this leads to the first research question of this thesis:

> **Research Question 1**: Is a BDNN an alternative for estimating APC effects, and can it predict the number of future incidences?

According to Hobcraft et al. (1985), the APC effects serve as proxies for measuring latent variables. Additionally, they conclude that the relationships between those variables are usually insufficient because the effects fail to explain the complexity of the underlying concepts. Therefore, there exist approaches in the literature that split the three variables into subsets. For instance, Yang and Land (2006) developed a hierarchical model that uses census data on an individual level incorporating additional variables to the analysis compared to an aggregated data set. Another example is the approach of Breeden and Leonova (2019), who use the estimation of a Bayesian APC model as the input for a neural network to compensate lack of long-term data that prevents making meaningful forecasts. Their empirical analysis showed that this approach successfully competes with established survival models. Taken together, the research community is interested in incorporating more surrogate variables into the APC analysis to extend the interpretation value of these models. As there exists first empirical evidence for the successful integration of a neural network in these methods, it would be interesting to investigate how a BDNN performs in this setting, to also obtain a measurement for the uncertainty of the resulting estimation. That leads to the following research questions:

> **Research Question 2**: Can an ensemble method combining the estimations of an established APC model and a BDNN assess sub-effects correctly?

To approach these research questions, this thesis uses a real-world data set describing the mortality of *chronic obstructive pulmonary disease* (COPD) in England and Wales from 1942 to 1996 and simulated examples. While the COPD data serves mainly for answering the first research question from a practical perspective, controlling the simulation of data allows to investigate constellations thoroughly that would not occur often in the real world. Each of these data sets builds the foundation for an experiment: The COPD data allows the implementation and optimization of a single BDNN to compare its estimated effects with the ones of an APC model. Additionally, it will predict the further development of COPD mortality on examples that are not part of the training process. Since the effects are known in advance in the second data set, it enables the investigation of the correctness of their relevance assessment.

1.2 Structure

This thesis is structured as follows. The next chapter will provide a brief overview of the most important topics of the literature that influence the development of the concepts in this thesis. The third chapter presents the derivation of the conceptual approaches. Afterwards, it will justify the general design of the experiments and their parametrization. Subsequently, it will describe the results of these experiments and discuss them in the light of the research questions. The final chapter concludes with the most crucial insights and gives intuitions for future research.

2 Theoretical Background

This thesis develops methods estimating APC effects on a given target variable using BDNNs. Additionally, those methods serve as predictive models that describe the behavior of the target variable in future periods. The derivation of these methods combines theoretical concepts from multiple areas of literature. The research activities addressing the development of the APC models compose the first field. Its major contribution consists of insights on the identification problem and the development of estimation methods. The second area is the literature of deep learning, which investigates the development, training, and utilization of BDNNs. This section provides a brief overview of these topics.

2.1 Age-Period-Cohort Models

APC models are a class of methods with many applications, especially in social and behavioral sciences. Their primary goal is to estimate age, period, and cohort effects on a variable of interest, whereas these effects are indicators for other latent variables (Hobcraft et al., 1985). The major challenge in the estimation of those effects is the identification problem, which prevents finding a unique solution by using a linear modeling approach (K. O. Mason, Mason, Winsborough, & Poole, 1973; Rodgers, 1982). There are many examples for APC models that try to tackle this challenge in many different ways. However, none of them solves the problem. This section presents the classical APC model and its variations.

2.1.1 The classical APC model

The general idea of APC models is to investigate the time-dependent influences of age, period, and cohort variables on a target variable. Thereby, none of these three variables does have any effect on the studied phenomena, but their underlying concepts have one: While age is an indicator for results of social influences on the physiology of an individual, period and cohort describe phenomena that influence the variable of interest, independent of an association with an age group, in a specific period or cohort (Hobcraft et al., 1985; Yang et al., 2008). It is important to note that each of these three variables can be expressed as a linear combination of the other two (K. O. Mason et al., 1973; Rodgers, 1982). For instance, it holds that $\text{Age} = \text{Period} - \text{Cohort}$. This relationship depicts a challenge in the development of an estimation model.

The classical APC model is the linear model shown in equation 1 (K. O. Mason et al., 1973). It estimates the relationship between the dependent variable Y and the APC variables. Thereby, α is the intercept, the β_i are the coefficients of the corresponding APC variables, and ϵ is a random error term. The regular way to approach this model would be to calculate the ordinary least square (OLS) estimator (Fahrmeir, Kneib, Lang, & Marx, 2007). However, the linear dependence between the independent variables leads to a singular data matrix without full rank, which prevents the determination of a unique estimator using the OLS method (Rodgers, 1982; Yang et al., 2008). From a formal perspective, there exists a function describing the real-world in the form of the model in equation 1, and thus, it is impossible to describe this *Data Generating Function* (DGF) without a unique estimator ap-

proximating it because researchers would need to select the right function out of an infinitely large set.

$$Y = \alpha + \beta_1 A + \beta_2 P + \beta_3 C + \epsilon \tag{1}$$

To overcome this identification problem, K. O. Mason et al. (1973) and Fienberg and Mason (1979) conclude that it is necessary to resolve the linear dependence between age, period, and cohort. Thus, they propose to either ignore one of the variables or restrict the model in the sense that the effects for belonging into one or another age/ period/ cohort group are equal. While they note that the first idea is not a meaningful approach in many disciplines, the model that results from the second one would be in many contexts too simple. One reason for this is that it is impossible to add interaction terms into the given model (Rodgers, 1982).

2.1.2 Advanced Methods

Many researchers followed the idea of the classical APC model and developed a multitude of examples that extend it. For instance, Yang and Land (2006) developed a hierarchical APC model to incorporate individual data into the estimation. Moving from an aggregated view to an individual level leads to new opportunities for the research community because it enables to include individual variables into the model. As each of the observations of the individual variables can be made at a specific period and each of them belongs to a specific cohort, they assume that the latter variables are quasi-independent, and thus, they add for each period and cohort a random effect to the model (Bell & Jones, 2014), which leads to the general model in equation 2. $Age_{i_{(j_1 j_2)}}$ is the age variable of individual i in period j_1 and cohort j_2. The error term ϵ and the random effects for period $u_{1_{j_1}}$ and cohort $u_{2_{j_2}}$ are normally distributed with a zero mean.

$$\begin{aligned} Y &= \beta_0 + \beta_1 Age_{i_{(j_1 j_2)}} + \epsilon \\ \beta_0 &= \alpha + u_{1_{j_1}} + u_{2_{j_2}} \\ \epsilon &\sim N(0, \sigma_\epsilon^2), u_{1_{j_1}} \sim N(0, \sigma_{u_1}^2), u_{2_{j_2}} \sim N(0, \sigma_{u_2}^2) \end{aligned} \tag{2}$$

Besides adding additional variables and transforming the problem into an individual dimension, other approaches focus on improving the estimation process to determine the real DGF. For instance, the intrinsic estimator of Yang et al. (2008) separates the design matrix into two subspaces that are orthogonal to each other. While the authors argue that the first subspace is independent of the variable of interest, and thus, not relevant for the estimation, the other one can be estimated using a principal component regression. This approach removes the dependence between the design matrix and the estimation of the regression coefficients of the APC model. Although the authors state that their estimator has desirable statistical properties, there exist theoretical justifications that its application is not suitable in every situation (e.g., Luo, 2013; O'Brien, 2011), and thus, it is not a general solution for estimating APC models.

The *Bayesian Age-Period-Cohort Modeling and Prediction* (BAMP) package in *R* implements another method for estimating the parameters of an APC model using

a Bayesian approach (V. Schmid & Held, 2007). The authors use a binomial logit model as shown in equation 3: They assume that the variable of interest, which counts the number of incidence cases, follows a binomial distribution. The first parameter of this distribution, which is the population size of age group i at period j, equals the logit of the incidence probability p_{ij}. Moreover, they formulate it as the sum of an intercept μ and the respective effects of the age θ_i, period ϕ_j, and cohort ψ_k variables.

$$\begin{aligned} y_i j &\sim B(n_{ij}, p_{ij}) \\ n_{ij} &= log(\frac{p_{ij}}{1-p_{ij}}) \\ &= \mu + \theta_i + \phi_j + \psi_k \end{aligned} \tag{3}$$

Since this method is a Bayesian method, it is necessary to specify a prior distribution for each of these parameters. In accordance with V. Schmid and Held (2007) and Berzuini and Clayton (1994), a random walk prior is a reasonable choice for this purpose. However, it depends on the final applications to determine the order of these prior distributions. The first order for the age parameter implies a constant trend and is defined as follows, whereas κ^{-1} is a precision parameter that is assumed to be Gamma distributed (this definition also holds for the other two variables):

$$\begin{aligned} p(\theta_1) &\propto const. \\ \theta_i \mid \theta_{i-1}, \theta_{i-2} &\sim N(\theta_{i-1}, \kappa^{-1}), \text{ for } i = 2, \cdots, I \end{aligned} \tag{4}$$

The second-order random walk prior distribution describes a linear trend for the age parameter which follows the definition (which is also valid for the remaining variables):

$$\begin{aligned} p(\theta_1) = p(\theta_2) &\propto const. \\ \theta_i \mid \theta_{i-1}, \theta_{i-2} &\sim N(2\theta_{i-1} - \theta_{j-2}, \kappa^{-1}), \text{ for } i = 3, \cdots, I \end{aligned} \tag{5}$$

According to Rodgers (1982), only the non-linear components are estimable in the class of APC models. Consequently, only effects with non-linear priors are identifiable assuming non-linear likelihoods, and thus, the effects with a second-order random walk prior are generally not estimable. Taking all groups of a variable together leads to their joint distribution, whereas $\mathbf{R}$ is the precision matrix of the respective prior distribution.

$$\begin{aligned} \boldsymbol{\Theta} &= (\theta_1, \cdots, \theta_I) \\ p(\boldsymbol{\Theta}) &\propto \lambda^{rg(\mathbf{R})/2} exp(-\frac{\lambda}{2}\boldsymbol{\Theta}'\mathbf{R}\boldsymbol{\Theta}) \end{aligned} \tag{6}$$

The use of random walks as prior distributions has the advantage that they enable a projection of the effects. For this purpose, practitioners calculate the effect for the next point in time using equation 7 and plug the result then into the binomial distribution in equation 3 to obtain an estimation of the number of incidences in the next period.

$$\begin{aligned} \theta_{I+1} &\sim N(\theta_I, \kappa^{-1}) && \text{for random walk 1,} \\ \theta_{I+1} &\sim N(2\theta_I - \theta_{I-1}, \kappa^{-1}) && \text{for random walk 2} \end{aligned} \tag{7}$$

Overall, the definition of these prior distributions allows to define the posteriori distribution using Bayes theorem with the data likelihood $f(y \mid \mu, \theta, \phi, \psi) = \prod_m^M f(y_m \mid \mu, \theta, \phi, \psi)$ that is shown in equation 8. Afterwards, it is possible to calculate the full conditional of each parameter vector to obtain an estimation for the corresponding effect. However, these are generally no standard distributions, and thus, the software samples from the posteriori using a Metropolis-Hastings algorithm (V. Schmid & Held, 2007). Besides the estimation of the effects, BAMP provides an option for adding a term for global-heterogeneity by adding another variable to the regression model in equation 3. This parameter is assumed to be normally distributed with a zero mean and a precision parameter that follows a Gamma distribution.

$$p(\mu, \theta, \phi, \psi \mid y) \propto f(y \mid \mu, \theta, \phi, \psi) p(f(y \mid \mu, \theta, \phi, \psi)) \tag{8}$$

Although it is possible to estimate the effects of the age, period, and cohort groups, their interpretation is not trivial because this method does not solve the identification problem. Nonetheless, there exist many examples that utilize this method in various fields, such as cancer research (Lopez et al., 2006), other general medical fields (Kypridemos et al., 2016), and social sciences (Odagiri, Uchida, & Nakano, 2011). Moreover, BAMP builds the basis for ensemble methods, like in the work of Breeden and Leonova (2019). They create an economic forecast model by estimating age, period, and cohort effects but face the challenge that the available data has incompatible time scales. An explanation is that organizations have better possibilities to store internal information than external ones leading to this discrepancy. Breeden and Leonova (2019) approach this problem in two steps: In the first one, they calculate a Bayesian APC model using BAMP and use these estimates as an input for training a *Deep Neural Network* (DNN). Thereby, the input nodes of those effects that are supported by long-term data are directly connected to the output node by a fixed weight matrix that consists of ones. In the other case, the effects are split into sub-variables, which describe the overall effect. This step bridges the gap between the difference in the time scales of the data because the DNN extrapolates the small time frame into a larger one. Overall, it enables a financial forecasting model, which their empirical analysis has shown.

2.2 Deep Learning

A substantial part of this thesis is the development of BDNNs. This section provides an overview of the foundations of this class of models. Thereby, it presents the general idea of training and using a DNN. Subsequently, it summarizes the state of the art in research about BDNNs and how they relate to DNNs. Afterwards, it introduces the idea of Bayesian optimization methods that support the finding of suitable hyperparameters. Finally, the section discusses the interpretability of those models and the consequences for practitioners.

2.2.1 Formulating Deep Neural Networks

In the past years, the area of machine learning, and especially the field of deep learning, had its breakthrough in the literature. The methods that belong to this area find a successful application in various domains, such as image recognition (e.g., Russakovsky et al., 2015) and natural language processing (e.g., Conneau, Schwenk, Barrault, & Lecun, 2017). According to Hastie et al. (2009) and Ripley (1996), the foundation of deep learning is the DNN: This model is a subclass of neural networks, which are again a class of non-linear statistical models. Their general idea is to feed an input vector through a network structure that consists of an input layer, multiple hidden layers, where each of them comprises several hidden units, and an output layer. In a fully connected DNN, all hidden units of a layer are connected to all hidden units of the subsequent layer. Figure 1 provides an example of the structure of these models. Each hidden unit takes the output vector of the previous layer as an input to a linear function $f(z) = W^T z + b$, where W is a weight matrix and b a bias vector. Both parameters are randomly chosen while initializing the model. After calculating this affine transformation of the input, the hidden units insert the result into a non-linear activation function. There exist a multitude of these function in literature, such as the *Rectified Linear Unit* (ReLU) function that is defined as $g(x) = max(0, x)$. The choice for such a function depends on the individual application.

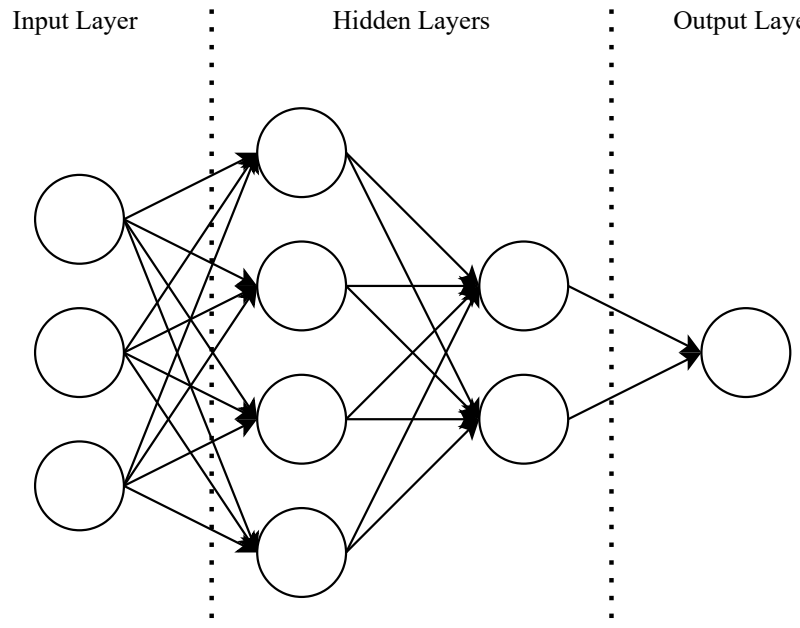

Figure 1: Sstructure of a two layer neural network.

A DNN requires a training procedure to adjust the model parameters to the given data, such that it can predict an output for a given input. One of the most basic learning algorithms for this purpose is the stochastic gradient-descent algorithm (Hastie et al., 2009; Ripley, 1996). It consists of four steps: First, the algorithm performs a forward-feed through the network, i.e., it inserts the input vector through the input-layer into the first hidden layer, and then the resulting output vector iteratively through the remaining hidden layers until the model yields a final output from the output layer. Afterwards, it calculates the loss using an arbitrary loss func-

tion L to obtain a measurement of the prediction error of the model at the current iteration. The choice of this function depends on the application. For instance, the root-mean-squared error would be a suitable choice in regression contexts. The third step is the backpropagation step, which goal is to determine the gradients of all model parameters with respect to the loss using the chain rule. In the final step, the algorithm updates the model parameters by subtracting the product of a *learning rate* α that determines to which extend the algorithm converges to an optimum and the gradient of the model parameter. It follows for each weight matrix W and bias vector b: $W \leftarrow W - \alpha\frac{\delta L}{\delta W}$ and $b \leftarrow b - \alpha\frac{\delta L}{\delta b}$. The algorithm repeats those steps for all available input vectors multiple times until it is perceived as optimal. The actual number is called the number of *epochs*.

The sole application of the stochastic gradient-descent algorithm could lead to a reasonable prediction performance of the training data but fails to generalize with unobserved data that is not part of the training procedure. There exist several approaches to counteract this overfitting problem. One example is the use of regularization techniques. Their idea is to include a penalty term in the update functions (Hastie et al., 2009; Ripley, 1996) - it penalizes "large weights that do not contribute a correspondingly large reduction in the error" (Witten & Frank, 2002) leading to a reduction of their importance regarding the prediction, which in turn reduces the degree of overfitting that results from the corresponding hidden units. Another example is the dropout method - it chooses a set of hidden units that the training algorithm ignores in each iteration with a predefined probability (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Thereby, it reduces the influences of single hidden units on the final model, which in turn reduces the degree of overfitting.

There are many examples of other learning algorithms in the literature that promise even better empirical performances. One of these variations introduces the idea of mini-batches: Instead of inserting a single input vector to the network in each iteration, it is possible to increase the sample size and calculate then the mean of the losses of those samples. Following Masters and Luschi (2018), the size of these batches should not be too large to ensure that the training algorithm converges fast and robust to the optimum. Another example is *Adam* (Kingma & Ba, 2014): Its central assumption is that the loss function is stochastic. Hence, it aims at finding the minimum of the expected loss function. To achieve this, the algorithm determines the first two moments $\hat{m}$ and $\hat{\nu}$ of the gradient of the loss and uses those for the update of the parameters: $W \leftarrow W - \alpha\frac{\hat{m}}{\sqrt{\hat{\nu}}+\epsilon}$, where ϵ is the step size and depicts a hyperparameter of the method that users have to tailor to their specific problem. This algorithm has the advantage of adjusting the learning rate dynamically, such that it prevents getting stuck at a local rather than the global minimum.

2.2.2 Bayesian Deep Neural Networks

The described training procedures in the section above result in models that yield a point estimation for each input. While this is sufficient in many applications, others require information about the uncertainty of the resulting estimations. Thus, the major goal of a BDNN is to obtain a Bayesian estimation of a probability distribution for each output label. Thus, the smaller the variance of this posteriori, the

smaller is the uncertainty of the estimation. There are several proposals to achieve this goal in literature. However, they generally have the same foundation: The training algorithm assumes that the model parameters are sampled from a probability distribution. Thus, it is possible to assign each of them a prior distribution. Subsequently, the algorithm can calculate the data-likelihood using initial data samples and combine the priors and the likelihood through Bayes theorem to the posteriori (MacKay, 1992; Neal, 1996).

The conceptual ideas in this thesis do not have special requirements for the learning algorithm of the used BDNNs, and thus, the choice of a training algorithm is generally irrelevant. Since the experimental part requires to implement one, it summarizes the VOGN optimizer (M. E. Khan et al., 2018) in the following and utilizes it for the implementation later on: The central issue in applying Bayes theorem to the given problem is the calculation of the normalization constant. Additionally, the likelihood and the prior distributions are usually nonconjugate. Consequently, the optimizer approximates the posteriori using a distribution such that the normalizing constant is computable. The normal distribution is a suitable candidate for this purpose (c.f. Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015). Based on these assumptions, M. E. Khan et al. (2018) derive update rules for the parameters of the prior distributions.

The natural-gradient variational inference method (M. Khan & Lin, 2017) builds the foundation for the VOGN optimizer. The idea of this method is to use a variational objective function

$$\mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{N} \mathbb{E}_q[log\ p(D_i \mid \theta)] + \mathbb{E}_q[log\ \frac{p(\theta)}{q(\theta)}]$$

for the calculation of the weight updates. $P(D_i \mid \theta)$ is the data likelihood, $p(\theta)$ the prior of the model parameters, and $q(\theta)$ the approximation of the posteriori. Overall, it results in the following update rule under the consideration of a Gaussian approximation, where θ are the model parameters and μ and σ^2 the parameters of the prior distributions:

$$\begin{aligned} \mu_{t+1} &\leftarrow \mu_t + \beta_t \sigma_{t+1}^2 \circ [\hat{\nabla}_\mu \mathcal{L}], \\ \sigma_{t+1}^{-2} &\leftarrow \sigma_t^{-1} - 2\beta_t [\hat{\nabla}_{\sigma^2} \mathcal{L}_t]. \end{aligned} \tag{9}$$

M. E. Khan et al. (2018) translate this update rule into the context of a BDNN. Thereby they define the objective of the training as minimizing the negative log-likelihood $-log\, p(D_i \mid \theta)$ such that the objective function is $f(\theta) = \frac{1}{N} \sum_{i=1}^{N} -log\, p(D_i \mid \theta)$. The use of mini-batches requires to determine the mean of the sum of the gradients of this objective, such that $\hat{g}(\theta) = \frac{1}{M} \sum_{i \in M} \nabla_\theta f_i(\theta)$, where M is the number of examples in the mini batch $\mathcal{M}$. Taken together, this leads to the following update rule:

$$\begin{aligned} \mu_{t+1} &\leftarrow \mu_t - \beta_t (\hat{g}(\theta_t) + \tilde{\lambda}\mu_t)/(s_{t+1} + \tilde{\lambda}), \\ s_{t+1} &\leftarrow (1 - \beta_t) s_t + \beta_t diag[\hat{\nabla}_{\theta\theta}^2 f(\theta_t)]. \end{aligned} \tag{10}$$

This update rule could lead to negative variances in the update, and thus, the authors propose to use the Generalized Gauss-Newton approximation for the Hessian matrix $\nabla^2_{\theta\theta} f(\theta_t)$. It follows that $\hat{h}(\theta_t) = \frac{1}{M} \sum_{i \in \mathcal{M}} [\nabla_{\theta_j} f_i(\theta)]^2 \approx \nabla^2_{\theta_j \theta_j} f(\theta_t)$. This results in the final version of the update rules:

$$\begin{aligned} \mu_{t+1} &\leftarrow \mu_t - \beta_t (\hat{g}(\theta_t) + \tilde{\lambda}\mu_t)/(s_{t+1} + \tilde{\lambda}), \\ s_{t+1} &\leftarrow (1 - \beta_t) s_t + \beta_t \hat{h}(\theta_t). \end{aligned} \tag{11}$$

Since the network stores the parameters of the distributions of the weights and biases rather than the values themselves, it is necessary to perform multiple forward-feeds through the network to approximate the parameters of the posteriori distribution. It is also important to note, that the update rules of the VOGN optimizer could potentially be extended using regularization techniques to further improve the predictive power of the network. However, this is not necessary for this thesis, and thus, will not be further explained.

2.2.3 Optimizing and Using Deep Learning

DNNs, as well as BDNNs, require the definition of hyperparameters to tailor them for specific applications. There exist empirical evidence that the choice of these parameters has a crucial impact on the success of the model. For instance, e.g., Pinto, Doukhan, DiCarlo, and Cox (2009) conclude this fact while comparing many biological vision models. However, they also summarize that the process of choosing those parameters is not trivial due to the potentially high search space. Although many authors still use a manual approach for defining the set of hyperparameters, literature developed methods to automate this process (Bardenet, Brendel, Kégl, & Sebag, 2013). It is possible to distinguish between two types of those algorithms. While the first type systematically tests the search space, such as grid search and random search algorithms (Bergstra & Bengio, 2012), the second type utilizes a probability distribution to predict meaningful candidates in the search space (Snoek, Larochelle, & Adams, 2012).

A famous example of the latter type is the family of *Bayesian Optimization* (BO) algorithms. Generally speaking, their goal is to determine the global minimum of an unknown, stochastic function that is not trivial to evaluate (Močkus, 1975). Following the description of Brochu, Cora, and De Freitas (2010), these algorithms utilize a surrogate model for the estimation of the objective function, which is easier to evaluate. This surrogate is a Bayesian model because it combines a prior distribution with the data likelihood to obtain this estimation. Many authors follow the proposal of Močkus (1975) to use a Gaussian process as the prior distribution for the surrogate model because the resulting posteriori is again a Gaussian process that is defined by its mean-function and covariance function (Močkus, 1975). Subsequently, sampling the surrogate $k(x)$ multiple times enables to approximate the location μ_x and scale parameter σ_x^2 of a Gaussian such that $k(x) \sim \mathcal{N}(\mu_y, \sigma_x^2)$. It is necessary to provide initial samples for creating this model, i.e., the method conducts a random search for a fixed number of steps to obtain these samples.

After generating the initial surrogate model, the algorithm uses its predictions to query an acquisition function to iteratively retrieve points in the search space that

should be evaluated next (Brochu et al., 2010). The goal of this phase is to optimize the surrogate model, and thus, the points should reduce the uncertainty at each point x in the search space, and simultaneously, explore new points to improve the overall prediction performance of the model. This is called the exploration-exploitation tradeoff. Among the variety of examples for acquisition functions, the *expected improvement criterion* is a reasonable candidate to approach this tradeoff (Brochu et al., 2010; Močkus, 1975; Snoek et al., 2012). Its idea is to estimate the expected value of the improvement function $I(x) = max\{0, k(x) - k(x^\star) + \zeta\}$, where $x^\star$ is the best objective value observed so far and $\zeta \geq 0$ a minimal improvement parameter that ensures reasonable runtimes of the algorithm by prohibiting only marginal steps. Taken together, the definition of the criterion, where D is the data, is

$$EI(x) = \mathbb{E}(I(x)) = \int_{-\infty}^{\infty} || \; k(x) - k(x^\star) + \zeta \; || \; P(k(x) \mid D) dk. \tag{12}$$

Since the calculation of an integral is not trivial from a computational perspective, there exists a closed form for this criterion (Brochu et al., 2010; Snoek et al., 2012):

$$EI(X) = \begin{cases} (\mu(x) - f(x^\star))\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) \leq 0 \end{cases} \tag{13}$$

Thereby, $Z = \frac{\mu(x) - k(x^\star)}{\sigma(x)}$, and $\Phi(Z)$ is the cumulative distribution function, and $\phi(Z)$ the probability density function. The algorithm extracts the next promising point by maximizing this criterion. Subsequently, it evaluates the point using the original objective function and adjusts the surrogate model. The algorithm repeats those steps for a fixed number of iterations until the set of hyperparameter is perceived as optimal.

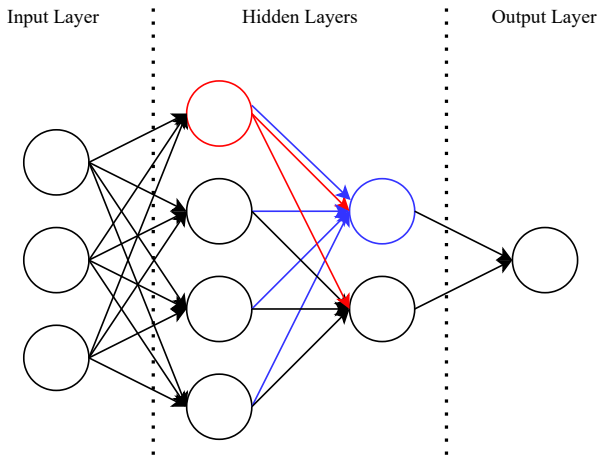

Figure 2: Example for calculating the relevance score of the red neuron using the LRP method. The red line divided by the sum of the blue lines at each node in the subsequent layer builds the degree to which the red node is relevant for the prediction of the blue node. The sum of those relevances is the relevance score of the red node.

Besides the optimization and usage of the predictive power of DNNs and BDNNs, a recent stream in the literature focuses on the interpretability of those models. The goal of this research field is to develop methods that explain the relationship between the inputs and the resulting predictions in the sense of explaining how a model makes a prediction, and thus, open the black box models. This step is necessary for various reasons. For instance, regulatory changes might require practitioners to provide users with an explanation of how a decision was made, or without an insight into the decision-making process, patients would not trust their physicians (Samek, Wiegand, & Müller, 2017). Whilst a classical sensitivity analysis that changes the value of a single input node ceteris paribus provides a first insight into explaining the relationship, it is insufficient in many contexts because it does not explain the actual function value but its variation. For this reason, there exist plenty of methods that try to explain it directly (Lundberg & Lee, 2017). Nonetheless, these measures are not directly comparable to the effect estimation of a linear regression model due to the non-linear character of a DNN.

One of these methods is the *Layer-wise relevance backpropagation* (LRP) (Bach et al., 2015): After training a DNN, the method calculates for each prediction separately for each input node a relevance score that describes to which degree the input was responsible for predicting the corresponding output. It achieves this by iteratively assigning these scores to all hidden units in each layer. Figure 2 demonstrates the calculation of this score for a single neuron - the score equals the sum of contributions the node has on the relevance scores of the nodes of the consecutive layer. Formally follows that the score of neuron i at layer l is

$$R_i^{(l)} = \sum_j R_{i \leftarrow j}^{(l+1)}. \tag{14}$$

The contribution of this neuron to another neuron j in layer $l+1$ equals the share of the affine transformation of the neuron of interest to the sum of the affine transformations of all neuron on layer l that are inserted into j multiplied by the score of j:

$$R_{i \leftarrow j}^{(l+1)} = \begin{cases} \frac{x_{ij} w_{ij}}{\sum_i x_{ij} w_{ij} + \epsilon} \cdot R_j^{(l+1)}, & \text{for } \sum_i x_{ij} w_{ij} \geq 0 \\ \frac{x_{ij} w_{ij}}{\sum_i x_{ij} w_{ij} - \epsilon} \cdot R_j^{(l+1)}, & \text{for } \sum_i x_{ij} w_{ij} < 0 \end{cases} \tag{15}$$

The stabilizer $\epsilon \geq 0$ ensures that there is no division by zero. The algorithm determines these scores for each layer starting at the output node until it assigns each input a score. It is noteworthy that these scores are only valid for a single prediction and that there does not exist an approach that aggregates them for an overall interpretation of the model. Thus, the interpretability of these models is still limited compared to other model classes such as linear regression.

3 Conceptual Approach

The goal of this thesis is to investigate how BDNNs could improve the predictive performance of APC models. Two requirements define the successful development of approaches that fulfill this goal. The first one is that the resulting models should be able to predict the development of the variable of interest into future periods. The second one requires the model to express the effect or the importance of an input variable on the output variable. This thesis develops two conceptual approaches to achieve this goal considering the two requirements. This chapter presents both approaches from a theoretical perspective. However, it extends the LRP method to ensure applicability in BDNNs, prior to this.

3.1 Bayesian LRP

While the LRP method aims at interpreting samples of DNNs, it does not apply to BDNNs without extending it. The reason is that the estimation of a BDNN is a posteriori distribution rather than a point estimation, and thus, feeding the same input through the network multiple times leads to different results. Consequently, applying the LRP method to this sampling process would also yield different relevance scores for the input nodes, which would bias further implication tasks. Therefore, it is necessary to extend this method. Additionally, the research community investigating APC models is not interested in the explanation of a single sample, but the overall effects of the input parameters on the target variable. As the method does also not fulfill this criterion, an extension requires also aggregating the sample view into a larger frame. This section proposes the *Bayesian LRP* (BLRP) method that incorporates these requirements.

The first challenge is to find robust relevance scores for the same inputs considering the posteriori distribution of the outputs. The general idea of BLRP is to sample from the posteriori multiple times until the distribution converges to a specific location and scale parameter. Thereby, it is necessary to store all outputs of all hidden units as well as all model parameters at each sampling step, which restricts the number of the sample size in terms of the available memory. Subsequently, the algorithm determines the relevance scores for each sample individually using the LRP method. Finally, it calculates the mean and standard deviation of the resulting relevance scores for each input node. While the mean is an estimation for the relevances of the inputs, the standard deviations provide a measurement of the uncertainty of this estimation. Consequently, the proposed method extends the LRP method from a Bayesian perspective.

The second challenge is to aggregate the results from the sampling process into a view that allows inferring about the general relevance of an input variable on the target variable. As mentioned before, it is not possible to derive linear effects like in a linear regression model of the independent variables on the dependent one because of the non-linear nature of BDNNs and the definition of the relevance scores that result from the LRP and BLRP methods. Nevertheless, the latter method equips researchers with the opportunity to obtain an impression on the relevance of an independent variable by investigating the distribution of its relevance scores per expression of this variable. On the one hand, these distributions give them

an intuition about the relevances of a single expression and the accompanying uncertainty. On the other hand, it allows comparing those distributions between the single expressions to gain insights about the development of a variable.

3.2 BAPCNN

The first approach is the naive one - it utilizes a BDNN as an alternative to established APC models and is called the *Bayesian APC Neural Network* (BAPCNN). In contrast to methods like BAMP that use the data on the Lexis diagram (V. Schmid & Held, 2007), it requires the data to comprise a set of quadruples in the form of (A, P, C, L), where A is the age index and P, C the respective index for period and cohort. Each of these variables could potentially be replaced with another set of variables that describe the effect in detail. For instance, the age variable could be replaced by the characteristics of an individual, such as its smoking and sleeping habits, and its actual age. Additionally, L describes the target variable for the three indexes. Besides the transformation of the data, a BDNN requires the definition of a training algorithm. As mentioned before, this thesis proposes to use the VOGN optimizer, whereas there is no theoretical restriction on using a different algorithm for this purpose. Following the derivation of this optimizer (M. E. Khan et al., 2018), the BAPCNN assumes that the network parameters are samples from a normal distribution leading to a Gaussian posteriori distribution. Moreover, it uses the average negative log-likelihood as its loss functions, where μ is the mean and τ is the precision parameter of a Gaussian distribution:

$$L = -\frac{1}{2} log(2 \cdot \pi) + log(\tau) - \tau \cdot (y - \mu)^2.$$

This approach has several advantages and disadvantages. The literature provides empirical evidence for the strong predictive performance of DNNs and BDNNs. Consequently, this approach should successfully predict cases of the target variable in future periods. However, it is necessary to show this using empirical experiments. One disadvantage is the number of hyperparameters of this method because their choice adds additional complexity to the method. Although a manual search would be possible, it is not satisfactory in many situations. Consequently, it makes sense to use an automated approach like the BO algorithm for identifying those. This step is essential because a poor choice of hyperparameters diminishes the predictive power of the model. Another challenge is the amount of available data. Previous empirical studies have shown that DNNs and BDNNs require large amounts of data to provide a reasonable prediction for unobserved data samples. Hence, overfitting the training data enables a better interpretation of the relevance scores that result from the BLRP method, but fails to generalize the resulting prediction into future periods. That depicts a tradeoff of the two subgoals of this thesis. Therefore, researchers could fit two models for this purpose, which is unnecessary in established APC models.

3.3 APC-Ensemble

The second approach adapts the idea of the hierarchical model of Breeden and Leonova (2019). While they intended to extrapolate a lack of data onto a larger

time horizon, the goal of this method is to use the estimation of an established APC model to get an insight on the relevance of effects that describe one of the three independent variables. Hence, this approach focuses not on the prediction of future incidences of the target variable but seeks to find an explanation for which components are relevant for building an effect. To achieve this, the algorithm firstly uses an established APC model and estimates it. Thereby, the choice of this method is generally irrelevant for further processing, but it is essential that the method can provide an effect estimation for each independent variable individually. Based on the resulting estimation, it is possible to extract the information of the dependent variable that only depends on one of the three variables by subtracting the effects of the remaining variables. This information builds the label for training a BDNN. The set of inputs for this model consists of variables that describe the overall effect, and thus, they depend on the domain of future applications. Optimizing and training this BDNN enables to retrieve the relevance of these inputs using the BLRP method.

This method splits an estimated effect into multiple sub-effects using a BDNN under the assumption of the correctness of the estimation of the APC model. Consequently, it cannot guarantee a reasonable description of the relevance of these sub-effects because none of the established APC models solves the identification problem such that there is no guarantee that it finds the effects describing the real DGF. Nevertheless, a domain expert could use the description of those sub-effects to judge the quality of the estimation of the APC model. Conversely, assuming an expert supports the estimation of an APC model, this hierarchical approach would provide researchers with even more information about potential influences that compose an aggregated effect.

4 Research Design

Each of the two conceptual approaches has a specific strength regarding the research questions. While the BAPCNN method promises a reasonable predictive power, the strength of the hierarchical approach lays in the interpretability of its estimation. Consequently, this thesis uses two types of experiments to investigate the empirical behavior of those methods. While the first type uses a real-world data set to gain an insight on the prediction performance of the BAPCNN method, the second one uses simulated data to observe how the hierarchical approach recognizes the relevance of the sub-effects that compose an overall effect. This section describes both types and their parametrization in detail.

4.1 COPD Experiment

The first experiment uses a dataset that describes the mortality of *Chronic obstructive pulmonary disease* (COPD) in England and Wales in the time between 1942 and 1996 for 15 age groups from 15 to 89, whereas each of these groups contains five years. It also contains the population data for those periods and the corresponding age groups. Moreover, it differentiates between the male and female population. This experiment aims at evaluating the predictions of the BAPCNN approach using this dataset. Thereby, it uses the first 45 periods for training an APC model and a BAPCNN for each gender group. Subsequently, it uses each model to predict the remaining ten periods and compares those predictions with the real number of incidences.

Lopez et al. (2006) use a similar dataset for predicting the mortality of COPD with a Bayesian APC model using BAMP. Consequently, it depicts a suitable candidate for representing the class of established APC models in this experiment. A similar data basis allows transferring the parametrization of the models in the paper into the present context. Table 1 summarizes those parameters: Lopez et al. (2006) assume a second order random walk as the prior distribution for all variables. Moreover, they specify that the precision parameter of those priors follows a Gamma distribution defined by the parameters a and b. Besides this, the authors include another parameter that measures the overdispersion in the model, which follows a Gaussian distribution with a precision parameter following a Gamma distribution. Finally, the estimation of the method uses 102000 *Markov Chain Monte Carlo* (MCMC) iterations, whereas 2000 of them describe the burn-in phase. Following the authors, the experiment in this thesis also implements the models using BAMP.

Besides the definition and estimation of the APC models, it is necessary to create and train a BAPCNN using the VOGN optimizer. That requires the transformation of the data, as described in section 3.2. Additionally, it also requires the definition of a set of hyperparameters for the training process. Since the choice of those is not trivial, this thesis uses the mlrMBO framework (Bischl et al., 2017) to run a BO algorithm leading to a reasonable set of hyperparameters. Table 2 summarizes the results of this procedure for both models. The *prior precision* and *precision init* parameters characterize the precision parameter of the prior distributions. Due to the Gaussian approximation in the VOGN optimizer, the calculation of the negative log-likelihood also requires the specification of a precision parameter, which is the

Parameter	Value
Prior distribution for Age	Random Walk 2
Prior distribution for Period	Random Walk 2
Prior distribution for Cohort	Random Walk 2
Number of MCMC iterations	100000
Burn in	2000
Gamma distribution of independent variables	a = 1, b = 0.0005
Gamma distribution of overdispersion parameter	a = 1, b = 0.05

Table 1: Overview on the used hyperparameters for training the Bayesian APC models in the COPD experiments.

noise precision parameter in the table. This thesis extends the torch package (Falbel & Luraschi, 2020) for implementing the VOGN optimizer.

Parameter	Value (female)	Value (male)
Prior precision	0.0074	0.0796
Precision init	47.75	30.694
Noise precision	0.01	0.01
Hidden units	(84, 109, 109)	(28, 70, 101)
Epochs	20	20
Batch size	16	16
Learning rate	0.01	0.01
Beta	0.999	0.999

Table 2: Overview on the used hyperparameters for training the BAPCNN models in the COPD experiments differentiated according to the underlying gender of interest.

After training both model types, the experiment uses the models to predict the next ten periods. Afterwards, it compares those predictions with each other considering the underlying gender in the data and with the real-world incidences. This comparison requires the definition of a metric that quantifies the predictive performance of the models. The mean absolute error

$$MAE(Z) := \frac{1}{\mid Z \mid} \sum_{i=1}^{|Z|} \mid z_i - z_i^\star \mid$$

and the root mean squared error

$$RMSE(Z) := \sqrt{\frac{1}{\mid Z \mid} \sum_{i=1}^{|Z|} (z_i - z_i^\star)^2}$$

are suitable metrics for this purpose. $Z = [z_1, z_2, \cdots, z_Z]^T$ and $Z^\star = [z_1^\star, z_2^\star, \cdots, z_Z^\star]^T$ are the input vector and the respective vector of real values. The smaller those metrics, the better is the predictive performance of the models.

4.2 Simulation

While the use of a real-world data set provides valuable insights on the empirical behavior of the prediction performance of the models, it does not allow to evaluate the estimation of the underlying effects that characterize the DGF because it is impossible to estimate them using the available tools due to the identification problem. For this reason, the second experiment conducts a simulation study that enables to control the definition of the DGF. Consequently, it is possible to validate that a model assesses the relevance of a variable correctly. This setting builds the basis for investigating the empirical behavior of the hierarchical approach.

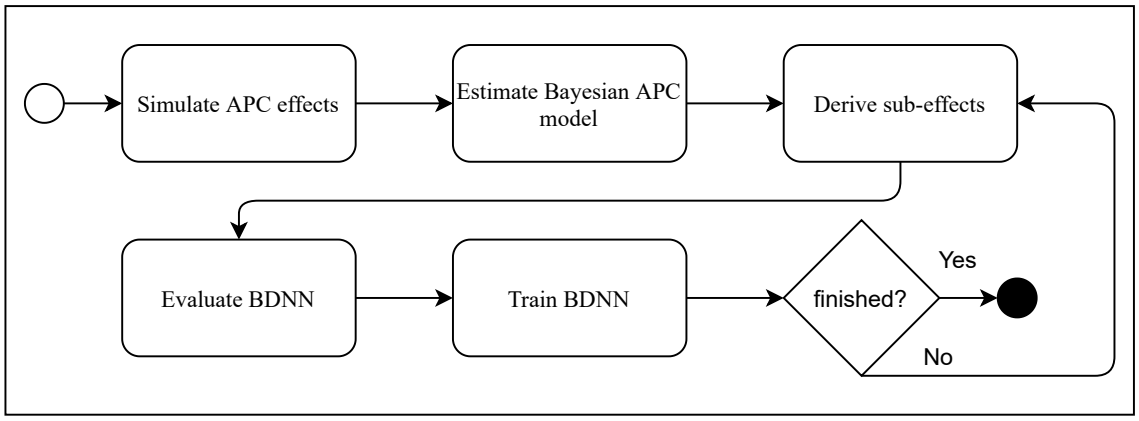

Figure 3: Summary of the simulation process.

Figure 3 shows the simulation process: The first step is to manually define the effects of age, period, and cohort. It uses the example of the data simulation function of the BAMP packages shown in figure 8 (V. J. Schmid, 2020). The next step is then to fit a Bayesian APC model to this data using the BAMP package. Subsequently, the core of the simulation starts. It splits the estimated cohort effect of the APC model randomly into three sub-effects such that the sum of those effects equals the overall cohort effect. These sub-effects build the input vector for a BDNN. The estimated cohort effect depicts the labels for those inputs during the training phase. After training the BDNN, the next step is to calculate the relevance scores for each input variable using the BLRP method. The simulation repeats this core 100 times to ensure statistical validity.

As the result of this BLRP method is an overview of the distribution of the relevance score per variable expression and the real values are only a single point, it is not trivial to find a distance measure to quantify the success of the estimation. Hence, an alternative is to plot the resulting distributions against the real values to investigate the empirical behavior of this method for each simulation run separately. It is infeasible to include all of these plots in the thesis itself, and thus, they will be made available in the electronic appendix together with the source code of both experiments.

The APC model, as well as the BDNNs, demand the definition of a set of hyperparameters. For the priors in the APC model, the simulation uses second-order random walk priors and the default number of iterations in the BAMP package, which are 55000 MCMC iterations with 5000 steps as the burn-in phase. The *apcSimulate* function of the BAMP package is used to generate the simulation data based on the

described effects in figure 8 in the appendix. This step assumes a constant population of one million residents. First manual experiments with the BDNN showed that there exist a set of hyperparameters that is suitable for all BDNNs independent from the simulation iteration. Table 3 summarizes this set.

Parameter	Value
Prior precision	11.55
Precision init	10.87
Noise precision	4.66
Hidden units	(28, 94, 85)
Epochs	10
Batch size	53
Learning rate	0.079
Beta	0.941

Table 3: Overview on the used hyperparameters for training the BDNNs in the simulation study.

5 Results

This section presents the results of the COPD experiment and the simulation study separately. Nevertheless, they both follow the same structure. First, they validate the successful training of the used models by presenting several diagnostic plots. This includes the check of the Bayesian APC models to converge using the *checkConvergence* function in the BAMP package (V. J. Schmid, 2020). Moreover, it consists of observing the development of the loss function of the BAPCNN model and plotting the prediction of the training data against the real labels. Afterwards, the sections describe the empirical behavior of the proposed models considering the goal of each experiment type.

5.1 Predictive Performance

The goal of conducting the COPD experiment was to investigate how the BAPCNN method competes against established APC models in predicting future incidence cases. The APC models for the female and the male dataset did not converge given their parametrization. Varying the number of MCMC iterations and the remaining hyperparameters did not lead to an improvement of this state. Consequently, the comparison of the methods is only partly possible. Figure 4 shows that the loss of both BAPCNN models converge to zero during the training phase, which indicates that this phase was completed successfully. The comparisons of the training labels and the prediction of those values after completing the training phase, which is shown in figure 6, support this statement because both plots imply that the models produce reasonable predictions on the training data.

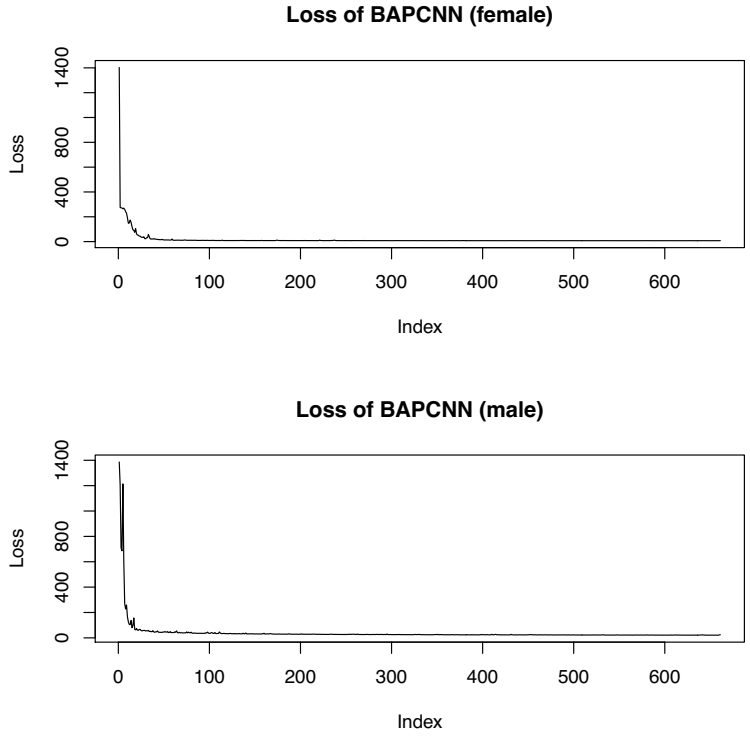

Figure 4: The value of the loss functions converges to zero with an increasing number of epochs for both BAPCNN models.

Figure 6 shows the predictions for the last ten periods of both model types and

differentiated according to the underlying gender of the models. While the black line demonstrates the real values for these periods, the orange and blue ones show the respective predictions of the BAPCNN and Bayesian APC model. It indicates that both model types can generally predict the number of incidences in future periods. However, it also shows that the predictions of the BAPCNN models are substantially closer to the real values than the prediction of the APC model independent of the underlying gender. The evaluation metrics in table 4 support this: They show that the predictions of both models are worse for the male dataset than the female one. Additionally, the MAE and the RMSE are both substantially lower for the BAPCNN model than for the APC models implying a better predictive performance of those models. Taken together, the BAPCNN method is a competitive alternative in the task of predicting future cases.

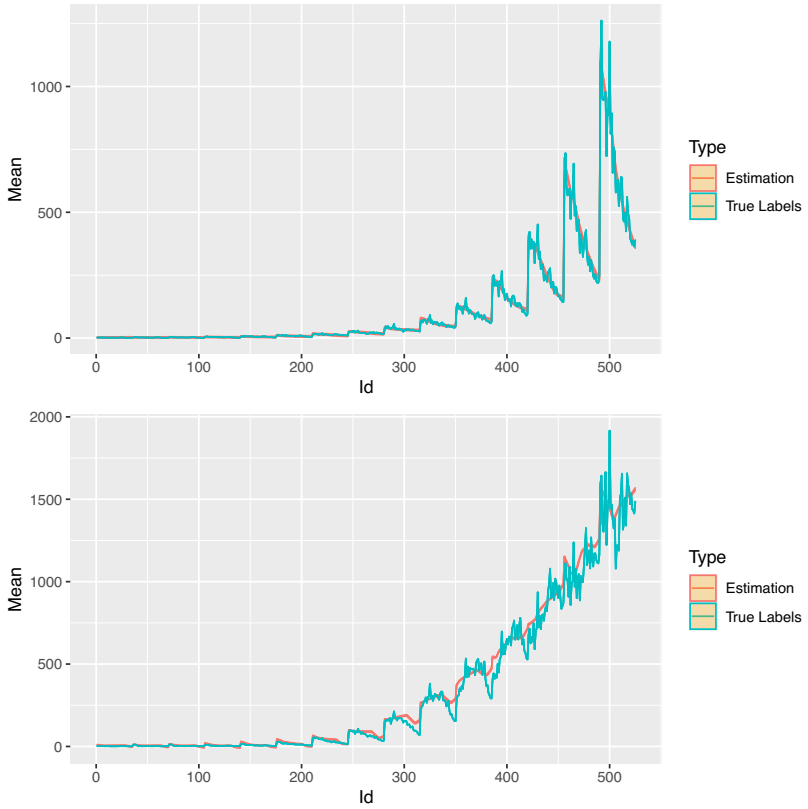

Figure 5: The plot shows the predictions of the BAPCNN models compared to the real values for the female and male model.

Model	MAE	RMSE
APC (female)	7915.519	7923.949
BAPCNN (female)	2255.025	2622.789
APC (male)	14089.21	14119.75
BAPCNN (male)	2603.235	2269.76

Table 4: Summary of the evaluation metrics that result from the COPD experiment differentiated according to the used model type.

5.2 Relevance of Sub-Effects

The second experiment aimed at investigating if the ensemble method predicts the relevance of sub effects correctly. To achieve this, the APC model and the BDNN must converge in their training phases. While the *checkConvergence* function of the BAMP package indicates that the APC model converges, the BDNN did not, given the previously identified parameterization. Although the application of the BO algorithm described earlier yielded a promising set of hyperparameters, the experiments showed that it only improved the prediction marginally. Another reason for this could be that there are too few training samples in the process, leading to the poor adjustment of the network parameters. One option to overcome this problem is to augment additional data and add it to the training set. A simple approach is to copy the input vectors and add a random standard Gaussian noise to it using the original labels, which also makes the prediction more robust against perturbations in the data (An, 1996). However, quintupling the number of samples also did not lead to sufficient results. Hence, the following assessments of the simulation are not directly applicable to other contexts.

As mentioned before, due to the non-linear nature of a BDNN it is impossible to extract the influences of an input variable on the target variable directly. Nevertheless, the BLRP method provides an opportunity to quantify this relationship. Figure 7 shows an example of the result for one sub-variable of a single simulation run. It represents the remaining runs because all of them provide very similar implications (compare the electronic appendix for all simulation results). The plot indicates that the median values of the relevance scores approximate the real sub-effects, which the red line indicates. Nevertheless, the plot also shows that the distributions of the scores are partly very broad, i.e., their variance is very high compared to the real effects, which makes the prediction not robust enough for practical implementations. Conversely, there are many examples for distributions with a low variance that approximate the real effects quite well. The convergence problem could be an explanation for these mixed accuracies. Either it leads to completely random predictions of the model, or the training process optimizes the model parameters only for a subset of inputs explaining the poor assessment of some inputs. All in all, the ensemble method requires further optimization in the training procedure, but generally assesses the relevance of inputs reasonably.

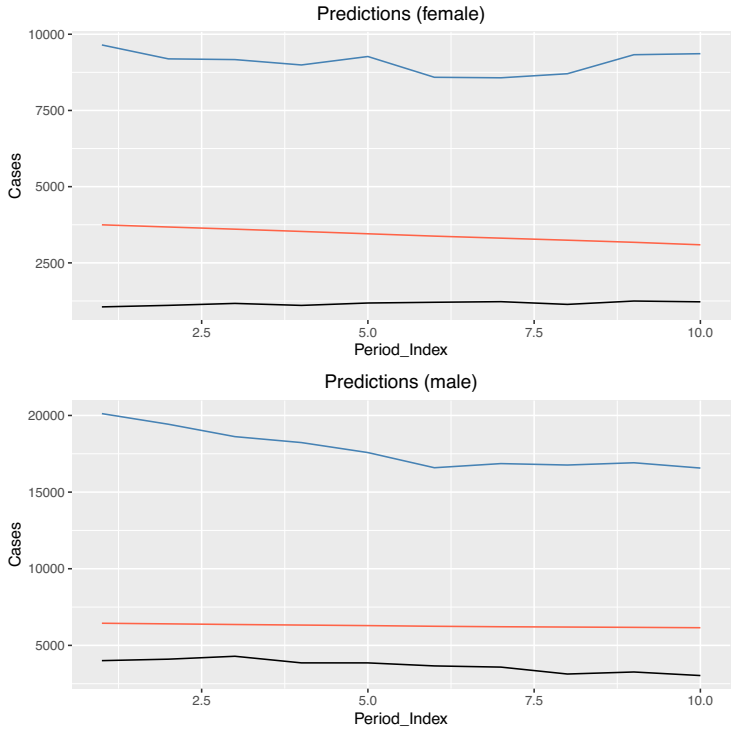

Figure 6: Overview of the prediction for all models compared to the real-world values and differentiated according to the underlying gender of the models. The black line shows the real values, and the orange/ blue one shows the respective predictions of the BAPCNN and Bayesian APC model.

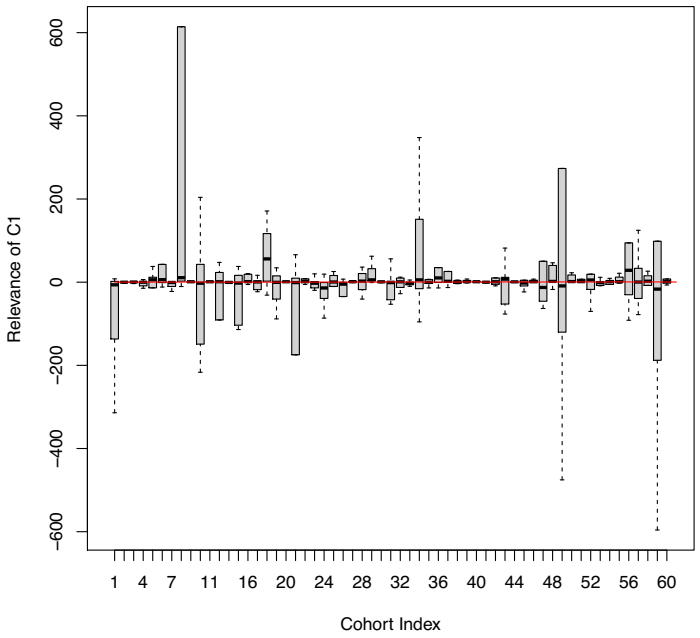

Figure 7: The figure shows an example for the assessment of cohort sub-effects using the ensemble method. While the red line indicates the simulated effect, the boxes show the distribution of relevance scores resulting from the application of the BLRP method.

6 Discussion and Conclusion

This section uses the results of the previous section and discusses them under consideration of the research questions. Thereby it identifies theoretical and practical limitations that restrict the generalizability of the implications that result from this discussion. Afterwards, the section concludes by summarizing the most important results of this thesis. Finally, it utilizes these findings to propose topics and open challenges for future research, which development would enhance the performance of the models in this thesis.

6.1 Discussion

The COPD experiment and the simulation study showed the empirical performance of the two conceptual approaches. The discussion of the approaches in the light of the two research questions requires the justification of the limitations that affect the generalizability of the following implications. There are mainly two sources for limitations in this thesis - the research design and the implementation challenges. The use of a single real-world data set is not representative for all real-world scenarios, and thus, BAPCNN models could fail to find a sufficient prediction in other contexts, or established APC models could outperform those. Moreover, a simulation allows to analyze specific data constellations, but it is unable to portray the complexity of the real-world. Thus, its validation would require analyzing the same constellations with real datasets, which is impossible, because there does not exist a method that generally identifies the true APC effects. Finally, the failing convergence of the models to their optimum also prevents the generalizability of the results because estimation errors could lead to the results at hand. Consequently, they would bias deriving further implications. The following answers the research questions considering these limitations.

The first research question asks whether a BDNN could be an alternative for the estimation of APC effects and if such a model could predict future incidences of the target variable. The previous section provides empirical evidence that the BAPCNN achieves sufficient prediction performance compared to the Bayesian APC model. However, the success of this method strongly depends on the identification of a set of hyperparameters because they lead to a better result of the training process. Although established APC models also require the definition of a suitable parametrization, the number of parameters is quite large for this class of models, which increases the complexity of finding such a set. The simulation study revealed another problem that could influence the predictive power of this approach - in domains where practitioners are unable to generate enough samples for the training phase, the model might fail to generalize such that its predictions are not sufficient. Although data augmentation methods such as the addition of a Gaussian noise could improve this situation, future research should investigate how the BAPCNN could be improved in these settings. Nonetheless, this approach depicts generally an alternative to established APC models considering only the ability to predict future incidences.

Besides this ability, a wholesome alternative would also be able to describe the single influences of the APC variables on the corresponding variable of interest. As this is

generally infeasible due to the identification problem, the second research question investigates if a BDNN can assess the influence of sub-effects on the overall effect correctly, given an estimation of this effect using an APC model in a hierarchical approach. Manual experiments in the COPD experiment setting showed that the BAPCNN is unable to assess those influences to a sufficient extent. The simulation study supports this view. Due to the failed convergence of the training phase, it was infeasible to evaluate those effects using the BLRP method. Overall, this thesis could not provide empirical evidence for the second research question. Yet, this does not prove that using BDNNs in the estimation of APC effects is not a valuable approach in general. It is the task of future research to apply the presented concepts to other data sets in various domains to obtain further impressions on this research question. Under the assumption that especially the ensemble method yields sufficient results in other contexts, it would also be interesting to analyze under which conditions a BDNN is able the assess the sub-effects correctly.

6.2 Conclusion

The goal of this thesis was to develop possibilities on how APC models could profit from BDNNs and analyze how those concepts would perform in a practical setting. To approach this, this thesis formulated two research questions that describe the primary tasks of APC models - those models should predict future incidences and assess the influences of variables correctly. Overall, it presented two conceptual approaches; each aims at fulfilling one of those goals. It conducted two types of experiments to determine how well these concepts achieve their goals. The first experiment uses a real-world data set to train a BAPCNN and predict the mortality of COPD for the next ten periods. The other one conducts a simulation study and utilizes a hierarchical approach to get an impression on the model's ability to assess sub-effects.

While the results in the experimental section provide empirical evidence for the prediction ability, it could not show that those models can also evaluate the influences of a sub-set of variables correctly. It would be interesting to see whether more complex formulations of the BDNN, such as a convolutional BDNN, or the application of other training algorithms would lead to different results. Those concepts could improve the estimation of the real APC effects and simultaneously enhance the predictive performance into future periods. Moreover, it could build the basis for another simulation study that investigates the assessment of the influences. This leads to another interesting idea for future researchers.

The evaluation of the influences using the BLRP method is not as expressive as describing the effects of the independent variables in a linear regression model. The primary challenge in these methods is that they evaluate only a single prediction rather than the influence of the variable. While this is a valuable property in some domains, it is insufficient in the area of cohort analysis. Thus, it is important that researchers in the field of interpretable machine learning continue their work to develop such a method in the future. Besides the lack of evaluation methods, the experimental section revealed a need for methods that automate the development of BDNNs. On the one hand, researchers, who did not have any contact with the area of deep learning will probably be deterred from the identification of the hy-

perparameters. Additionally, many social sciences lack data for a cohort analysis over longer periods, and thus, the number of training samples is very low. Consequently, it would be beneficial to develop methods in the future that would support those researchers in their tasks by automating these processes. However, this requires the development of improved methods. Independent from the proposals for future research, this thesis showed that BDNNs are generally an valuable addition for improving APC models.

A Simulated Effects

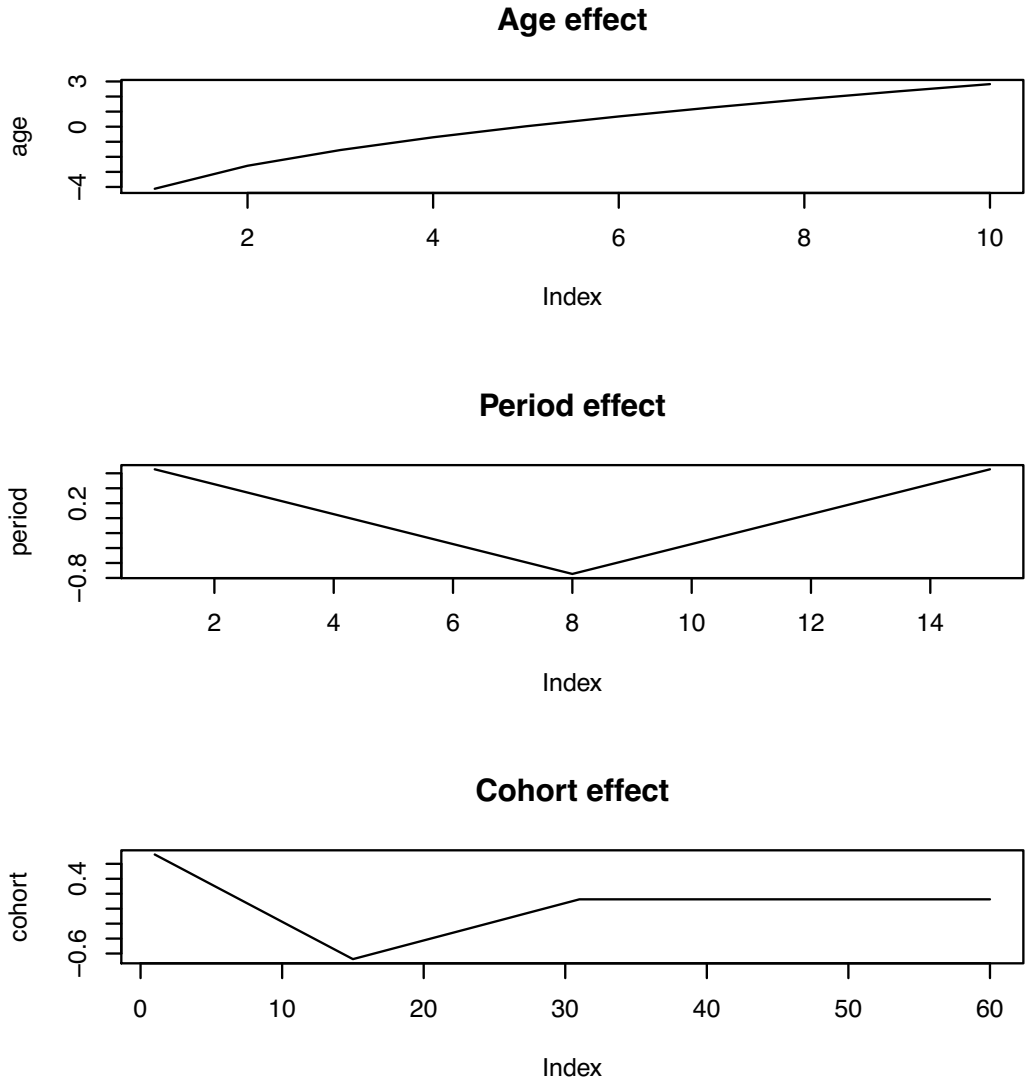

Figure 8: The figure shows the definition of example effects that are used in the simulation study. These effects are derived from the manual of the BAMP package (V. J. Schmid, 2020).

References

An, G. (1996). The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation*, *8*(3), 643–674. doi:10.1162/neco.1996.8.3.643

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), 1–46. doi:10.1371/journal.pone.0130140

Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013). Collaborative hyperparameter tuning. In *Proceedings of the 30th international conference on international conference on machine learning - volume 28* (II–199–II–207). Atlanta, GA, USA: JMLR.org.

Bell, A., & Jones, K. (2014). Another 'futile quest'? a simulation study of yang and land's hierarchical age-period-cohort model. *Demographic Research*, *30*, 333–360.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*(null), 281–305.

Berzuini, C., & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, *13*(8), 823–838. doi:https://doi.org/10.1002/sim.4780130804. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780130804

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. arXiv: 1703.03373 [stat]. Retrieved from https://arxiv.org/abs/1703.03373

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Breeden, J. L., & Leonova, E. (2019). When big data isn't enough: Solving the long-range forecasting problem in supervised learning. In *2019 international conference on modeling, simulation, optimization and numerical techniques (smont 2019)*. Atlantis Press.

Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2017). Very deep convolutional networks for text classification. arXiv: 1606.01781 [cs.CL]

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2007). *Regression*. Springer.

Falbel, D., & Luraschi, J. (2020). *Torch: Tensors and neural networks with 'gpu' acceleration*. R package version 0.2.0. Retrieved from https://CRAN.R-project.org/package=torch

Fienberg, S. E., & Mason, W. M. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological methodology*, *10*, 1–67.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hobcraft, J., Menken, J., & Preston, S. (1985). Age, period, and cohort effects in demography: A review. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem* (pp. 89–135). doi:10.1007/978-1-4613-8536-3_4

Khan, M., & Lin, W. (2017). Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th international conference on artificial intelligence and statistics* (Vol. 54, pp. 878–887). Fort Lauderdale, FL, USA: PMLR.

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., & Srivastava, A. (2018). Fast and scalable bayesian deep learning by weight-perturbation in adam. arXiv: 1806.04854 [stat.ML]

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv: 1412.6980 [cs.LG]

Kypridemos, C., Allen, K., Hickey, G. L., Guzman-Castillo, M., Bandosz, P., Buchan, I., ... O'Flaherty, M. (2016). Cardiovascular screening to reduce the burden from cardiovascular disease: Microsimulation study to quantify policy options. *BMJ*, *353*. doi:10.1136/bmj.i2793. eprint: https://www.bmj.com/content/353/bmj.i2793.full.pdf

Lopez, A. D., Shibuya, K., Rao, C., Mathers, C. D., Hansell, A. L., Held, L. S., ... Buist, S. (2006). Chronic obstructive pulmonary disease: Current burden and future projections. *European Respiratory Journal*, *27*(2), 397–412. doi:10.1183/09031936.06.00025805. eprint: https://erj.ersjournals.com/content/27/2/397.full.pdf

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *CoRR*, *abs/1705.07874*. arXiv: 1705.07874. Retrieved from http://arxiv.org/abs/1705.07874

Luo, L. (2013). Paradigm shift in age-period-cohort analysis: A response to yang and land, o'brien, held and riebler, and fienberg. *Demography*, *50*(6), 1985–1988. doi:10.1007/s13524-013-0263-8

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, *4*(3), 448–472.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, *38*(2), 242–258. Retrieved from http://www.jstor.org/stable/2094398

Mason, W., & Wolfinger, N. (2001). Cohort analysis. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 2189–2194). doi:https://doi.org/10.1016/B0-08-043076-7/00401-0

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. arXiv: 1804.07612 [cs.LG]

Močkus, J. (1975). On bayesian methods for seeking the extremum. In G. I. Marchuk (Ed.), *Optimization techniques ifip technical conference novosibirsk, july 1–7, 1974* (pp. 400–404). Berlin, Heidelberg: Springer Berlin Heidelberg.

Neal, R. M. (1996). Priors for infinite networks. In *Bayesian learning for neural networks* (pp. 29–53). Springer.

O'Brien, R. M. (2011). The age–period–cohort conundrum as two fundamental problems. *Quality & Quantity*, *45*(6), 1429–1444. doi:10.1007/s11135-010-9397-6

O'Brien, R. M. (2011). Intrinsic estimators as constrained estimators in age-period-cohort accounting models. *Sociological Methods & Research*, *40*(3), 467–470. doi:10.1177/0049124111415369

Odagiri, Y., Uchida, H., & Nakano, M. (2011). Gender differences in age, period, and birth-cohort effects on the suicide mortality rate in japan, 1985-2006. *Asia Pacific Journal of Public Health*, *23*(4), 581–587. PMID: 19861318. doi:10.1177/1010539509348242. eprint: https://doi.org/10.1177/1010539509348242

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLOS Computational Biology*, *5*(11), 1–12. doi:10.1371/journal.pcbi.1000579

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Rodgers, W. L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review*, *47*(6), 774–787. Retrieved from http://www.jstor.org/stable/2095213

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv: 1708.08296. Retrieved from http://arxiv.org/abs/1708.08296

Schmid, V., & Held, L. (2007). Bayesian age-period-cohort modeling and prediction - bamp. *Journal of Statistical Software, Articles*, *21*(8), 1–15. doi:10.18637/jss.v021.i08

Schmid, V. J. (2020). *bamp: Bayesian age-period-cohort modeling and prediction*. R package version 2.0.8. Retrieved from https://volkerschmid.github.io/bamp/

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Yang, Y., & Land, K. C. (2006). A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores. *Sociological methodology*, *36*(1), 75–97.

Yang, Y., Schulhofer-Wohl, S., Fu, W. J., & Land, K. C. (2008). The intrinsic estimator for age-period-cohort analysis: What it is and how to use it. *American Journal of Sociology*, *113*(6), 1697–1736. Retrieved from http://www.jstor.org/stable/10.1086/587154