Estimating Dynamic Causal Effects using an Approximated Bayesian Long Short-Term Memory Network

Tony Su[1] and Raad Khraishi[1,2]

[1]University College London, United Kingdom
[2]Data Science & Innovation, NatWest Group, United Kingdom

November 27, 2024

Abstract

This paper proposes a novel approach for computing the dynamic causal effects of policies and shocks through estimating the counterfactual period. An approximated Bayesian Long Short-Term Memory (LSTM) network will be used to compute the counterfactual, yielding period-wise and cumulative causal effects. Epistemic uncertainty will also be captured by simulating posterior sampling through Monte Carlo dropout, providing an estimate of the model's uncertainty. The methodology is applied to the estimation of the dynamic causal effect of the global financial crisis on the Dutch economy, where the analysis shows a 7.6% contraction in real GDP per capita (RGDPpc) by 2008, growing to 13.1% by 2014. The cumulative RGDPpc decline is estimated at $15,656, with a total GDP loss of approximately $261 billion from 2008 to 2014.

1 Introduction

The field of causal inference and causal machine learning has witnessed significant growth over the past decade, but the focus largely remains on static treatment effects. The issue of confounding variables is at the core of causal inference, which is resolved in traditional static settings with randomised control trials (RCT), matching propensity scores, among other methods. Subsequently, methods such as generalised random forests and double machine learning, which excel at computing average and heterogeneous treatment effects in static contexts, are widely adopted in the literature [1, 2, 3].

Despite their success in applied domains, these methods fall short in time-dependent contexts where the effects of interventions evolve dynamically. In order to perform causal inference in time-dependent contexts, where the observational data lacks an adequate amount of control and treated units, counterfactual periods are estimated to compute the treatment's causal effect. Metrics like the average treatment effect (ATE) can be misleading in such scenarios. For instance, a strong initial policy effect followed by a gradual decline might be misinterpreted as a medium-sized intervention. Static metrics fail to capture the nuance of treatments with delayed effects, such as vaccines or monetary policy, where impacts do not materialise immediately [4], or treatments with feedback loops, which are ubiquitous in complex systems such as the economy. To understand the treatment's effect holistically, it requires an analysis of how the impact evolves across various horizons, capturing the initial, lagged, and cumulative effects of the intervention. This approach provides a nuanced understanding of the short-term and long-term effects, which static metrics often overlook.

The global economy has witnessed two major financial crises in the past two decades, and an understanding of the impact of such shocks is of utmost importance to policy-making, and the welfare of citizens worldwide. However, the observed economic data often exhibits complex, non-linear patterns, requiring increasingly sophisticated architectures to capture the intricate patterns. Deep learning algorithms have demonstrated considerable success in the time-series domain, and should be leveraged in econometrics and causal inference. However, the application of deep learning algorithms in causal inference frameworks is relatively sparse, and the scope of existing methods is primarily on individualised heterogeneous effects. To address this gap in the literature, we apply an approximated Bayesian LSTM to the interrupted time-series analysis framework.

Our contribution to the interrupted time-series analysis framework is the novel application of a deep learning algorithm, where a counterfactual period is estimated, and causal effects over various horizons are computed. The dynamic causal effect is measured as the difference between the observed ground truth and the hypothetical counterfactual period where the treatment had not been applied. The proposed model is free from constraints faced by traditional statistical models, such as Bayesian Structural Time-Series (BSTS) and Autoregressive Integrated Moving Average (ARIMA) models, previously used to estimate the counterfactual period. Statistical models are effective at capturing local trends and seasonalities apparent in commercial data that exhibit less complex patterns. Their compatibility for tasks characterised by trends and seasonalities that can be clearly decomposed and modularised is apparent. But for a complex system such as the economy, where business cycles are aperiodic and effects of shocks are non-linear, deep learning algorithms are better suited for their ability to capture long-term and non-linear temporal dependencies. Our model will show greater fidelity to the pre-treatment economy, and will extrapolate a more realistic counterfactual period.

Additionally, our model addresses the need for handling model uncertainty by constructing an uncertainty interval, providing a quantified measure of confidence in the estimated causal effects. At inference time, with Monte Carlo (MC) dropout enabled, the prediction produced during each forward pass approximates sampling once from the posterior distribution. Over many samples, we are able to construct a large enough sample to approximate the posterior distribution, from which we can compute an uncertainty interval using the parameters derived from the posterior distribution. MC dropout circumvents the computational burden of maintaining a large number of posterior distributions as faced by Bayesian neural networks, and provides computationally inexpensive probabilistic estimates. In sensitive domains where dynamic causal inference is applied, prediction uncertainty is of considerable importance as it provides an indicator of the model's reliability, and guides decision-making with far-reaching consequences.

The remainder of this paper is structured as follows. Section 2 provides an overview of the existing literature, and a description of the the framework, the choice of model, and the evaluation process is provided in Section 3. Section 4 presents the results of our experiment, and Section 5 concludes with a discussion of the implications and limitations of our findings.

2 Related Work

Across academic disciplines, dynamic causal effects are computed using various methods. Difference-in-Difference (DiD) models and panel data models are favoured among political scientists, where the continuously changing effects of policy over time is of interest, while Bayesian hierarchical models account for heterogeneity [5] [6] [7]. Marginal Structural Models (MSMs) and survival analysis are frequently employed in epidemiology and medicine to analyse the impact of dynamic treatments effects while accounting for time-varying covariates [8] [9]. However, these panel data models do not align with our objectives and the nature of the analysis we aim to perform. In contrast, Bayesian Structural Time Series (BSTS) models are widely used, offering a probabilistic framework for handling dynamic treatments.

Vector Autoregressive Models: Stock and Watson

Alongside Local Projections and Dynamic Stochastic General Equilibrium (DSGE) models, Vector Autoregressive (VAR) models are most favoured amongst macroeconomists [10] [11]. VAR models capture the linear interdependencies among multiple time series, with each variable modelled as a linear function of lagged values of itself and the other variables in the system, and the systems of equations are solved to obtain the coefficient matrix.

For a system with k variables, the VAR model of order p (VAR(p)) is expressed as:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \epsilon_t \tag{1}$$

Where:

- Y_t is the vector of endogenous variables at time t.
- $A_1, A_2, \ldots, A_p$ are coefficient matrices.
- ϵ_t is a vector of white noise error terms.
- p is the number of lags.

However, VARs are 'reduced-form' models, often requiring further constraints to identify structural shocks. Structural Vector Autoregression (SVAR) models are posed as a solution by imposing additional constraints that are driven by economic theory, to allow causal inferences to be made beyond the correlations captured by reduced-form VARs [12].

SVAR model of order p (SVAR(p)) is expressed as:

$$BY_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \eta_t \tag{2}$$

Where:

- B is the matrix of contemporaneous relationships between the variables.
- Y_t is the vector of endogenous variables at time t.
- $A_1, A_2, \ldots, A_p$ are coefficient matrices.
- η_t is a vector of structural shocks (uncorrelated errors).
- p is the number of lags.

The structural shocks η_t are related to the reduced-form residuals ϵ_t by:

$$\epsilon_t = B^{-1} \eta_t \tag{3}$$

The matrix B is used to impose theoretical constraints on the system, including but not exclusive to restrictions about the timing or theory, and this is instrumental in justifying the identification of structural shocks and provides robust theoretical grounding.

CausalImpact

In the CausalImpact package, the task of computing dynamic causal effects is structured with a framework where a BSTS is used to produce probabilistic estimates of the counterfactual periods [13]. The framework is robust and easily implementable, but there still exist limitations and gaps that we aim to addresss.

The BSTS model is a state-space model for forecasting probabilistic time-series data. It is defined by two equations:

$$y_t = Z_t^\top \alpha_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2) \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q_t) \tag{2}$$

Where:

- y_t is the scalar observation.
- α_t is the d-dimensional latent state vector.
- Z_t, T_t, R_t, and Q_t are model matrices that define the system structure.
- ϵ_t and η_t are independent Gaussian noise terms with variances σ_t^2 and Q_t, respectively.

The observation equation (1) relates the observed data y_t to the hidden state α_t, while the state equation (2) governs the evolution of α_t over time.

BSTS models are modular and customisable, accounting for trends, seasonality, and other components by including additive variables. While they provide robust and interpretable probabilistic forecasts, BSTS models face shortcomings that can be mitigated through employing a different methodology. BSTS produces counterfactual estimations through a latent state, which models components as mentioned above. While this is effective for data with identifiable patterns, such as an explicit periodic seasonal pattern, for complex data with unstructured temporal dependencies, a more sophisticated architecture may exhibit higher efficacy when applied to complex dynamic systems, such as business cycles with aperiodic cycles. Additionally, non-linear effects and feedback loops may be difficult to capture as part of the BSTS's modular components, which assumes fixed lengths and patterns of trends and seasonalities. Additionally, the performance of BSTS models will deteriorate asymptotically with the number of covariates, and this bottleneck arises from the difficulty in modelling and maintaining a non-trivial amount of posterior distributions. Sophisticated sampling techniques like Markov Chain Monte Carlo (MCMC) methods are employed to model the posterior distribution, offering robust theoretical grounding at the expense of high computational overhead.

3 Problem Statement

The challenges and subsequent gaps in computing dynamic causal effects can be categorised into three main clusters:

3.1 Uncertainty Quantification in Machine Learning Predictions

A significant challenge in machine learning algorithms is the lack of uncertainty quantification in predictions, particularly algorithms geared towards time-series forecasting, where point estimates are produced. In sensitive domains where machine learning is being applied to such as healthcare, education, and economics, it is crucial to provide probabilistic estimates that capture the variability and uncertainty that arises from the stochastic nature of machine learning algorithms [14]. Incorporating uncertainty quantification also adds a layer of interpretability to the estimates produced by the model, which is a well known flaw of opaque 'black box' models.

3.2 Lack of Dynamic Causal Models

Causal inference methods and causal machine learning algorithms are heavily focused on static treatments, creating a gap in our ability to handle dynamic treatments. Such static treatments are prevalent in microeconomics and adjacent fields which heavily utilises techniques and methods from causal inference, which makes it unsurprising that causal machine learning algorithms are also geared towards static microeconomic problems. As such, the dynamic nature which is inherent in various treatments and shocks in sensitive domains are neglected, and they're not exclusive to macroeconomic problems.

3.3 Inflexibility of Econometric Models

Traditional econometric models lack the flexibility required to capture the complex, nonlinear dynamics that emerge in the economy [15]. The economy functions as a complex system, characterised by feedback loops, path dependencies, and interdependent relationships, which cannot be adequately modelled using static equilibrium models [16]. Instead, dynamic approaches are required to account for the evolving interactions and emergent properties inherent in such systems. Although solutions to these gaps have been proposed to some extent in their respective fields, a unified framework has yet to be developed which utilises these tools in unison to better model and estimate dynamic causal effects.

4 Objectives

The primary objectives of this paper can be distilled into two key goals, each addressing the challenges outlined earlier:

- Develop a robust framework for estimating dynamic causal effects that includes probabilistic estimates to quantify uncertainty.
- Apply these methods to real-world macroeconomic shocks and policies, assessing their effectiveness under varying conditions by performing counterfactual analysis to compute dynamic causal effects.

In pursuit of these objectives, our research is guided by the following key questions:

1. How do Bayesian deep learning models perform in estimating dynamic treatment effects compared to traditional econometric models?
2. Can Bayesian deep learning models provide more accurate forecasts of macroeconomic variables in response to policy interventions, while also quantifying uncertainty?
3. What are the trade-offs between interpretability, predictive accuracy, and uncertainty estimation when using machine learning for dynamic causal inference?

4.1 Significance

This research is of considerable importance to macroeconomic policy analysis and related fields, where an understanding of the dynamic impact of interventions is critical for well informed decision-making. Orthodox methods often fall short in environments characterised by non-linearity, structural breaks, and complex interactions between variables, which is the case for many macroeconomic issues where intricate feedback loops and interdependence between variables are prevalent. By integrating Bayesian deep learning into causal analysis, we aim to provide a more accurate, flexible, and probabilistic-driven approach to estimating dynamic treatment effects. This has far-reaching implications not only in macroeconomics but also in related fields such as environmental science, neuroscience, and political science, where dynamic causal analysis is an invaluable method in the toolboxes of economists and political scientists alike.

5 Methodology

We propose a novel methodology to compute dynamic causal effects by estimating the counterfactual period using an approximated Bayesian LSTM network. The LSTM network will be used to forecast the counterfactual period, and the epistemic uncertainty of our model, as distinct from aleatoric uncertainty, will be captured by employing Monte Carlo dropout during inference [17]. This approach approximates sampling from the posterior distribution by performing multiple forward passes with dropout enabled, where different nodes are 'active' during each forward pass, and a sample from the approximate posterior distribution is taken each time. The Bayesian nature of the LSTM allows for probabilistic estimates, which entails the mean prediction surrounded by a 99% uncertainty interval, both of which are estimated parameters from the posterior distribution.

5.1 Model Framework

The proposed methodology applies deep learning algorithms in interrupted time series analysis. Interrupted time series analysis is a quasi-experimental design used in econometrics to analyse the data prior to and following the implementation of a treatment [18]. This is done by establishing a baseline trend and comparing it with the ground truth data a posteriori. The baseline trend is constructed by extrapolating the pre-treatment data to forecast values that would be observed in the absence of the treatment, in the period after the treatment is implemented. Subsequently, the treatment's effect over various horizons can be inferred from a comparison of the counterfactual and the ground truth.

This is a significant advancement from traditional interrupted time series analysis where regression and statistical methods are used to forecast the counterfactual values. More specifically, linear regression models with dummy variables are used to distinguish between pre-treatment and post-treatment periods. Though there has been use of statistical methods such as ARIMA models, deep learning adds a layer of complexity to the modelling process, better capturing non-linearities that are present in volatile financial and economic data, and most importantly, complex patterns with long term dependencies.

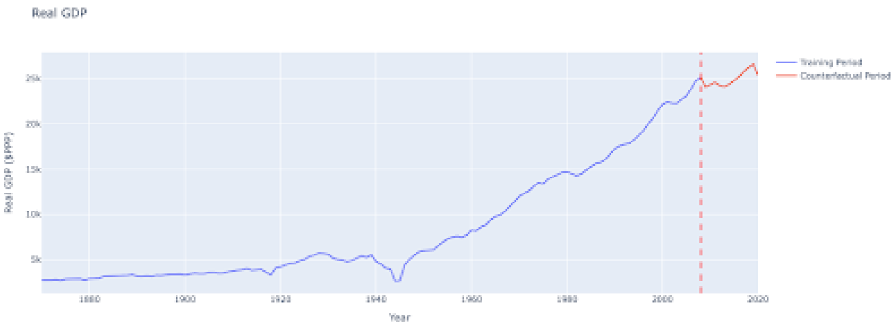

Figure 1: Real GDP (in PPP) over time, illustrating the training period (blue line) and the counterfactual period (red line). The vertical dashed red line marks the intervention point in 2008, after which the model estimates the counterfactual real GDP assuming no impact from the intervention (e.g., financial crisis).

There exists a crucial caveat, and it entails the difference between forecasting and counterfactual estimation. This paper is only concerned with the latter as there are diverging implications, motivations, and use cases, hence we make the distinction. Forecasting is an immensely difficult task compared to estimating a counterfactual due to the inherent uncertainty present in the future. The nature of our methodology reflects the difference: shocks are difficult to forecast inductively, and the best we can do is to propagate observed data into future periods. This is particularly the case with 'Black Swan' events, whose causal effect are the primary object of interest [19].

Time series forecasting is essentially a curve-fitting exercise, where the basis for future prediction is derived from historical data. Subsequently, for a forecast to be accurate, it is imperative that the statistical distribution of the time series is stationary. Otherwise, it is difficult to extrapolate the model into the future if there exists structural breaks in the data. Counterfactual estimation permits us to make the crucial assumption of stationarity as our objective is to extrapolate the behaviour observed prior to the treatment in order to estimate a counterfactual period.

5.1.1 Model Architecture

For our model, we employ the LSTM network for univariate forecasting. The structure of the LSTM allows the network to receive input and output that are of different lengths, which is suitable for time series forecasting where the length of the forecast is significantly shorter than the data on which the model is trained on. The LSTM only consists of one layer, with a hidden size of two to show that a relatively simple neural network will perform well without unnecessary complexity. The formulation and mechanics of the LSTM is provided below, and this will form the basis of our model.

Long Short-Term Memory (LSTM) Formulation

An LSTM is a type of recurrent neural network that addresses the vanishing gradient problem by introducing gating mechanisms. For a sequence input $\{x_1, x_2, \ldots, x_T\}$, where x_t is the input at time step t, the LSTM computes a hidden state h_t and a cell state c_t at each time step.

LSTM Equations

The LSTM uses three gates: the input gate i_t, forget gate f_t, and output gate o_t, along with a candidate cell state $\tilde{c}_t$. The equations are as follows:

$$\begin{aligned}
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) && \text{(Forget Gate)} \\
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) && \text{(Input Gate)} \\
\tilde{c}_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) && \text{(Candidate Cell State)} \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t && \text{(Cell State Update)} \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) && \text{(Output Gate)} \\
h_t &= o_t \odot \tanh(c_t) && \text{(Hidden State)}
\end{aligned}$$

The model utilises the cell and hidden states to better captures intricate temporal dependencies which are often driven by unorthodox factors that are not strictly statistical, such as herd behaviour and other behavioural dynamics. Though Monte Carlo dropout is applied during inference to approximate posterior sampling, our model's lack of complexity did not warrant the use of regularisation methods.

The LSTM is preferred over the BSTS, as used in the CausalImpact framework, due to its better handling of long-term dependencies and nonlinearities alongside its more efficient and scalable computation capabilities, particularly when dealing with high dimensional spaces. The rigid modular structure of the BSTS, which is better suited for consistent periodic patterns such as seasonality patterns for businesses, is more aligned to the needs of less complex systems. Therefore, while BSTS is useful for small-scale, interpretable models with clear structural components, a Bayesian LSTM is better suited for modern, large-scale forecasting tasks.

5.1.2 Approximated Bayesian Inference

In order to quantify the model uncertainty of our predictions, we approximate Bayesian Inference using Monte Carlo dropout during inference, which is distinct from dropout during training. We maintain the dropout probability at 0.01 for stability, and follows a similar framework to how dropout would be applied during training. We run 50,000 forward passes with Monte Carlo Dropout applied, where the deactivation of random neurons produces a distribution of stochastic predictions each time, and this simulates posterior sampling [17]. The posterior distribution $p(\theta \mid D)$, as provided below, represents the updated belief about the parameters θ after observing data D, combining prior knowledge $p(\theta)$ and the likelihood $p(D \mid \theta)$.

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

where:

- $p(\theta \mid D)$ is the posterior distribution of the parameters given the data.
- $p(D \mid \theta)$ is the likelihood of the data given the parameters.
- $p(\theta)$ is the prior distribution of the parameters, representing our belief about the parameters before observing the data.
- $p(D)$ is the marginal likelihood, also called the evidence, which normalizes the posterior.

Samples from the approximated posterior distribution are then used to compute the mean and standard deviation parameters, and a 99% uncertainty interval is then constructed from the parameters. This produces a forecast that provides robustness to epistmeic uncertainty at low computational costs. We opted for an approximated approach as it is more tractable and computationally inexpensive as opposed to a Bayesian Neural Network [17]. A Bayesian LSTM will require complex sampling techniques such as Variational Inference or Markov Chain Monte Carlo to learn posterior distributions over the parameters, as opposed to point estimates as is done in a frequentist deep learning algorithm.

Additionally, the successful modelling of a posterior that is faithful to the true distribution is delicate and relies heavily on the correct specification of the prior distribution [20]. Misspecification often leads to biased estimates of model parameters, and MC Dropout may be susceptible to misspecification as it assume a unimodal Gaussian distribution for prior distributions. Scalability is another advantage of MC Dropout especially when training deep neural networks with a non-trivial amount of parameters. This advantage derives from our use of an approximated Bayesian approach with a frequentist neural network, which circumvents the burden of maintaining posterior distributions over a large number of parameters. However, Monte Carlo Dropout lacks the theoretical rigour and grounding that Bayesian neural networks provide, but it provides a computationally inexpensive and efficient approximate solution.

5.1.3 Dataset

The methodology will be applied to the Jordà-Schularick-Taylor Macrohistory Dataset, which contains data concerning major macroeconomic variables across OECD countries [21]. For the purpose of this study, the Dutch economy will be analysed along with data spanning from 1870 to 2020. We will perform univariate forecasting, hence we will be using past values of Real GDP per capita (PPP, 1990 Int$, Maddison) to forecast the counterfactual period, enabling us to compute dynamic causal effects by comparing the data with the ground truth.

5.2 Model Training and Counterfactual Estimation

5.2.1 Training Loop

The training of the Bayesian LSTM requires the minimisation of the model's loss function to model the temporal dependencies of the pre-treatment period. This is done iteratively over 500 epochs where the model receives input in the form of time-series data with one input feature and a sequence length of 4, and produces a single output which represents the prediction for next period's value. The model's chosen loss function is the Mean Squared Error (MSE), and the model's parameters will be optimised with the Adam optimiser. The specific steps in the training loop are as follows:

- **Initialisation:** The model is initialised with hidden and cell states of zeroes, a learning rate of 0.01, and a hidden layer with 2 nodes. Input sequences of length 4 will be fed as input into the LSTM, and the step ahead forecast will be produced for the following periods. The model parameters are then optimised as part of the training loop over 500 epochs.
- **Forward Pass:** With each forward pass, the LSTM produces a context vector in the form of a hidden state, which is a latent state, which is maintained alongside the cell state. The LSTM then produces the step ahead forecast based on the context vector received. Using the Mean Squared Error (MSE) as the criterion, loss values are computed.
- **Backward Pass and Parameter Update:** The gradient of the loss with respect to model parameters are calculated by the backpropagation algorithm. Using the Adam optimiser, we then update parameter

values so as to optimise the loss function subject to regularisation constraints. After each update, the optimiser's gradients are reset to zero using `optimizer.zero_grad()` to prevent gradient accumulation.

- **Epoch Logging:** Over 500 epochs, the training loop is repeated with training loss outputted every 10 epochs to monitor the convergence rate and the overall performance.

The Bayesian LSTM is thus optimised iteratively via the training loop over 500 epochs, allowing it to capture the complex temporal dependencies present in the time-series.

5.2.2 Counterfactual Estimation and Inference

The counterfactual period will be estimated using the optimised LSTM network for periods beyond the treatment window. By extrapolating and generating predictions for the hypothetical 'untreated' period, the counterfactual values will deviate from the observed ground truth, thus revealing the causal impact of the shock or treatment on the variable of interest. In this paper, the counterfactual values represent how the Dutch economy would have evolved in the absence of a financial crisis from 2008 onwards, assuming that macroeconomic variables, such as Real GDP per capita, continued to grow at pre-crisis rates.

5.3 Model Evaluation

Model evaluation will be based on two dimensions: the accuracy of the time series forecasts and the quality of the estimated counterfactuals. Quantitative metrics such as Root Mean Squared Error (RMSE), and the Prediction Interval Coverage Probability (PICP) will assess the model's forecasting accuracy and the reliability of the probabilistic forecast. The quantitative metrics will be supported by qualitative evaluation, where through domain expertise, the feasibility of the counterfactual will be assessed to ensure it reflects reality and is consistent with economic theory or previously observed data.

5.3.1 Forecast accuracy

The forecasting accuracy of the LSTM model will be evaluated using traditional time series evaluation metrics such as Root Mean Squared Error alongside the Prediction Interval Coverage Probability, which assesses the quality of the probabilistic predictions.

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in [\hat{y}_i^L, \hat{y}_i^U]\} \tag{4}$$

The Prediction Interval Coverage Probability measures the proportion of the ground truth data which lies within the upper and lower bounds of the uncertainty interval. A high PICP value indicates that the uncertainty interval is reliable, and the majority of the ground truth is contained within the interval.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

The Root Mean Squared Error (RMSE) is given as the square root of the average of the squared differences between the predicted values ($\hat{y}_i$) and the actual values (y_i). The RMSE is the designated loss function due to its ability to penalise predictions that deviate significantly from the ground truth, making the loss function sensitive to outliers. Additionally, it also has the property of being differentiable, which facilitates tractable learning and enables efficient optimisation.

5.3.2 Counterfactual Evaluation

When estimating counterfactuals, the evaluation process will also assess the feasibility and consistency of the counterfactuals. It is crucial that quantitative metrics are not the sole method of evaluation as forecasts may deviate from economic intuition and reality, and may instead produce an infeasible counterfactual period. The qualitative judgement, derived from domain expertise, will ensure that the counterfactual values align with the pre-treatment trend and remains true to the model's perception of the data. For example, the growth of the Real GDP per capita during the counterfactual period has to remain consistent with the data observed previously, and it is highly unlikely that the counterfactual values will be less than the Real GDP per capita observed during the financial crisis.

6 Results

The counterfactual period is estimated by the model, and plotted along with the uncertainty interval produced through 50,000 samples of approximated posterior sampling through Monte Carlo dropout. The model accuracy is captured by the mean squared error loss of 0.000210 on normalised data from 1870 to 2008. The

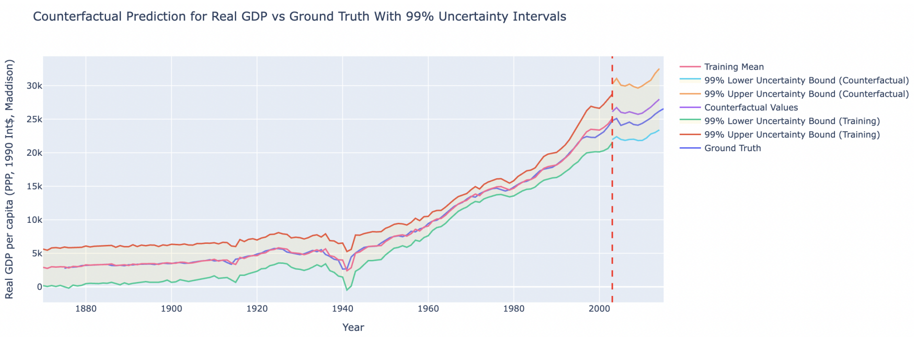

Figure 2: Counterfactual prediction for real GDP per capita (in 1990 International Dollars) compared with ground truth values, along with 99% uncertainty intervals. The solid lines represent the mean estimates for the training period and counterfactual values. The shaded regions indicate the 99% confidence bounds, where the red shaded area corresponds to the counterfactual period, and the green shaded area corresponds to the training period. The vertical dashed line at 2008 marks the intervention point after which the counterfactual is estimated.

model is intentionally overfitted so as to preserve the characteristics of the pretreatment data to extrapolate the counterfactual period.

The graph shows that the ground truth lies entirely within the uncertainty interval, yielding a PICP of 1, indicating complete coverage. This is a testament to the quality of the model's uncertainty interval. Despite there being complete coverage, the bandwidth of the uncertainty interval remains relatively narrow, mitigating the trade-off between coverage and precision, and this instills confidence in the probabilistic estimates.

Additionally, the point-wise and cumulative effects are also given. The point-wise causal effect reflects the difference between the observed value of real GDP per capita and the predicted counterfactual value for a given year, and the point-wise difference in 2008 is -$1,858, which indicates that the real GDP per capita was lowered by that exact amount as a direct result of the financial crisis. This negative point-wise effect accumulates each year following the crisis. By 2014, the point-wise difference grows to $3,227, meaning the economy was performing $3,227 below where it would have been had the crisis not occurred. The cumulative causal effect by 2014 is estimated to be around $15,692, which indicates that each Dutch citizen, on average, experienced a loss of $15,692 in real GDP per capita from 2008 to 2014. The data shows a significant and persisting impact on the Dutch economy, where the greatest point-wise causal effect occurred in 2014, which may indicate that the financial crisis had left economic scarring and this may have impacted the growth rate of the economy.

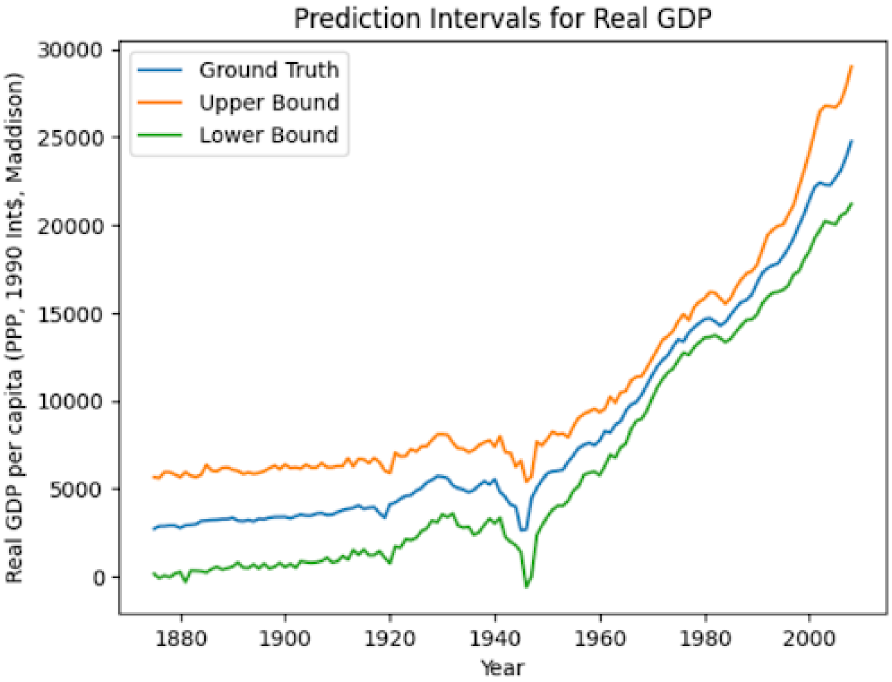

Figure 3: Prediction intervals for real GDP per capita (in 1990 International Dollars, Maddison database) over time. The blue line represents the ground truth values, while the orange and green lines indicate the upper and lower bounds of the predicted interval, respectively. These bounds provide a measure of uncertainty in the predicted values of real GDP per capita across different historical periods.

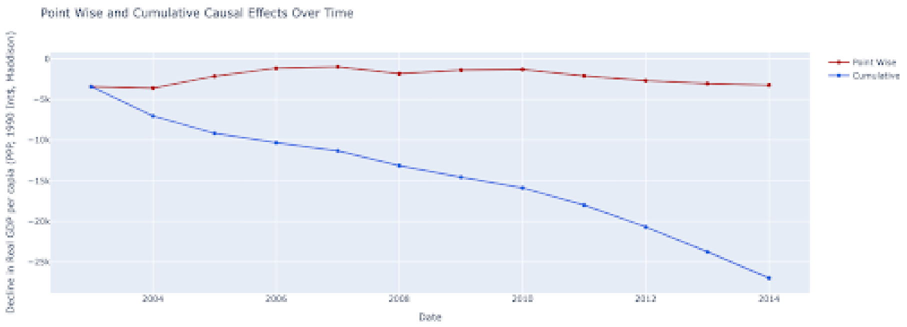

Figure 4: Point-wise and cumulative causal effects of the 2008 financial crisis on real GDP per capita (in 1990 International Dollars, Maddison) over time. The red line represents the point-wise effects, showing the yearly difference in real GDP per capita between the observed data and the counterfactual scenario without the crisis. The blue line shows the cumulative causal effects, illustrating the accumulated economic loss per capita from 2008 to 2014 as a result of the crisis.

In aggregate, the total loss in the Dutch economy's GDP from 2008 to 2014 amounts to approximately $261 billion, which translates to a loss of around 5% of GDP. This figure is consistent with empirical findings where the Dutch GDP declined by 3.7% in 2009, and this metric does not account for the hypothetical growth that would have occurred in the counterfactual period, as is provided in our metric [22]. Our metric also takes into account the compounding effects of the financial crisis, and this is consistent with empirical data for the Dutch economy.

7 Conclusion

This paper demonstrated the effectiveness of the Bayesian LSTM in estimating a counterfactual period and producing probabilistic estimates through the high accuracy of the model's predictive abilities, with a mean squared error of 0.000210 on normalized historical data, suggests a high fidelity to pre-treatment data characteristics, and the precise and complete coverage uncertainty intervals generated through Monte Carlo dropout achieving a PICP score of 1. The cumulative effect over the period from 2008 to 2014 suggests a significant economic loss, amounting to an average reduction of $15,692 in real GDP per capita per Dutch citizen. These findings indicate not only immediate economic setbacks but also long-term scarring effects on growth and other macroeconomic variables, potentially altering the economic trajectory of the Netherlands.

Despite our simple architecture, it shows that neural networks are highly efficient at capturing temporal dependencies, and should be utilised and incorporated more in economics and other disciplines. Future work could explore extending this model to other contexts and integrating additional covariates and introduce further model complexity to further enhance its predictive capabilities. This approach could offer policymakers and researchers a powerful method for understanding and mitigating the impacts of significant economic events.

References

[1] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters, 2017.

[2] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

[3] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2018.

[4] David Gruen, John Romalis, and Naveen Chandra. The lags of monetary policy. *Economic Record*, 75(3):280–294, September 1999.

[5] Anders Fredriksson and Gustavo Oliveira. Impact evaluation using difference-in-differences. *RAUSP Management Journal*, 54:519–532, 10 2019.

[6] J.M. Robins, M.A. Hernan, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[7] Gregory Wawro. Estimating dynamic panel data models in political science. *Political Analysis*, 10(1):25–48, 2002.

[8] Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

[9] T. Clark, M. Bradburn, S. Love, and D. G. Altman. Survival analysis part i: Basic concepts and first analyses. *British Journal of Cancer*, 89:232–238, 2003.

[10] Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980. Accessed: 2024-10-25.

[11] James H. Stock and Mark W. Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115, December 2001.

[12] José L. Montiel Olea, James H. Stock, and Mark W. Watson. Inference in structural vector autoregressions identified with an external instrument. *Journal of Econometrics*, 225(1):74–87, 2021.

[13] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274, 2015.

[14] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[15] Saeed Moshiri and Norman Cameron. Neural network vs econometric models in forecasting inflation. *Journal of Forecasting*, 19(3):201–217, 01 1998.

[16] W. Brian Arthur. Foundations of complexity economics. *Nature Reviews Physics*, 3:136–145, 2021.

[17] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[18] Andrea L. Schaffer, Timothy A. Dobbins, and Sallie-Anne Pearson. Interrupted time series analysis using autoregressive integrated moving average (arima) models: A guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*, 21:58, 2021.

[19] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Penguin Books, 2008.

[20] Sylvia Richardson and Louis Leblond. Some comments on misspecification of priors in bayesian modelling of measurement error problems. *Statistics in Medicine*, 16(1-3):203–213, Jan 15–Feb 15 1997.

[21] Òscar Jordà, Moritz Schularick, and Alan M. Taylor. Macrofinancial history and the new business cycle facts. In Martin Eichenbaum and Jonathan A. Parker, editors, *NBER Macroeconomics Annual 2016, Volume 31*, pages 213–263. University of Chicago Press, Chicago, 2017.

[22] Wiljan van den Berge, Hugo Erken, Marloes de Graaf-Zijl, and Eric van Loon. The dutch labour market during the great recession, 2014. Accessed: 2024-10-25.