# ERASMUS UNIVERSITY ROTTERDAM

## Erasmus School of Economics

## Master Thesis Health Economics

# Evaluating and improving the long-term care budget projections in The Netherlands

Name student: Rijk van Oostenbrugge
Student ID number: 432778
Supervisor: dr. P.L.H. Bakx
Second assessor: dr. T.M. Marreiros Bago d'Uva
Internship supervisor: G. Kaper (Nederlandse Zorgautoriteit)

Date: April 28, 2023



**Abstract**

In this paper we evaluate the forecasting models which recommend the Dutch Ministry of Health, Welfare, and Sport on whether the long-term budget (Wlz-kader) set for a given year will be sufficient. This recommendation is performed twice a year by the Dutch Healthcare Authority, in February and July. We restrict our analysis to a macro-level approach, using time series model types and levels of aggregation to seek better forecasting performance, compared to the current methodology, which uses ETS models with univariate data on aggregated on the level of the insurer to make forecasts for the coming year. The forecasts in this paper are performed with time series models, which use the amount of monthly declared budgets for a single care type or a set of care types as inputs, forecasting 10 to 15 months ahead to calculate whether the budget remains sufficient. The absolute sums of budget errors and absolute deviations of each model are calculated for six simulations of the recommendation, from 2019 to 2022. From our analyses, we do not find statistical significant evidence that any of the tested model types show a gain or loss in performance relative to the current method.

# Contents

# 1 Introduction

Long-term care (LTC) expenditures are set to rise with an aging population, changing societal models, increasing demands for better quality of care, and technological changes (Pavolini and Ranci, 2008). This can lead to increasing pressures on the governmental budget sheets, as LTC expenditures are relatively significant, amounting to 1.5% of the total GDP in OECD countries in 2018 (Muller and Morgan, 2020). Among these, The Netherlands can be seen as an outlier with almost 4% of Dutch GDP being spent on LTC in 2018, which is the highest share of all OECD countries. To manage LTC expenditures well, the Dutch government sets yearly budgets for LTC (Wlz-kader).

In 2015, The Netherlands implemented the Long Term Care Act (Wet langdurige zorg, Wlz). Wlz covers care for people who need permanent supervision or permanent access to care, due to severe somatic, psychogeriatric, mental, or sensory limitations (Ministerie van Volksgezondheid, Welzijn en Sport, 2022). The amount and types of care one is eligible for is determined by the type and severity of their disability, this is independently assessed by the Dutch Centre of Needs Assessment (Centrum Indicatiestelling Zorg, CIZ), where patients are given a care needs assessment code in the case that they are eligible for long-term care via Wlz. In the case of eligibility, one can choose for receiving in-kind care or receiving a budget that can be used to purchase care (persoonsgebonden budget, PGB).

Prior to 2015, LTC was governed by two acts: the Exceptional Medical Expenses Act (Algemene Wet Bijzondere Kosten, AWBZ) and the Social Support Act (Wet Maatschappelijke Ondersteuning, Wmo), where AWBZ covered a much larger share of the Dutch long-term care system, as Wmo only covered assistance programs (Maarse and Jeurissen, 2016). In 2015, the contents of the AWBZ has been transferred to four acts: The Long Term Care Act, the Health Insurance Act (Zorgverzekeringswet, Zvw), the Social Support Act (Wet Maatschappelijke Ondersteuning, Wmo), and the Youth Act (Jeugdwet). The latter three acts previously existed and received extra services in their respective domains. However, Wlz took over most of the budget and tasks that AWBZ had (Van Ginneken et al., 2015).

The goal of the implementation of Wlz is to contain costs in LTC. This led to a shift in the responsibility for social care from the national to the municipality level and an increased importance for care to take place at home, preferably via informal care, instead of at an institution (Kroneman et al., 2016).

The execution of Wlz is delegated to 7 legal entities of health insurers, each of which manage at least one of 31 regional purchasing offices. The purpose of a regional purchasing office is that it purchases in-kind care and allocates PGB's for those within the region who are eligible for and opt in to long-term care. Furthermore, as access to long-term care is a legal right in The Netherlands, the regional purchasing offices do not bear any financial risk, this is borne by the Dutch government. However, the regional purchasing offices are subject to budget constraints. The Dutch government sets a national budget, which is then divided over the 31 regional purchasing offices by the Dutch Healthcare Authority (Nederlandse Zorgautoriteit, NZa).

Prior to the start of each year $t$, the Dutch government publishes the long-term care budget for the next year through two letters, published in June and October, respectively containing a preliminary and a final budget. This is followed by negotiations between the regional purchasing offices and service providers about the price and quantity of in-kind long-term care for year $t$, leading to budget requests. Then NZa tests whether these requests fit within the regional budgets and applies restrictions when these exceed their respective budgets. During year $t$, the Dutch Ministry of Health, Welfare, and Sport (Ministerie van Volksgezondheid, Welzijn en Sport) requests NZa in the letter containing the final budget to assess whether the long-term care budget for year $t$ would still be sufficient to fund long-term care in year $t$ multiple times throughout the year. Previously, the timing of these assessments has not been formally set and has varied over the years. However, from 2022 onwards, these assessments are set to be performed in February and July. In the case that NZa finds that the current budget will not be sufficient, the Ministry of Health, Welfare, and Sport will request the Ministry of Finance for extra funding. Following these assessments, on the first of November, the service providers are allowed to request an adjustment of the budget for year $t$. These adjustments are then tested by NZa whether these will fit in the budget. Then, in June of year $t+1$ NZa calculates for each insurer whether their realized revenue is below the budget set for year $t$.

Currently NZa forecasts the expected in-kind LTC budget using univariate Error, Trend, Seasonality (ETS) models on the amount of declarations for each care type at the level of the insurers within Wlz. The choice for this model is mainly determined due to its ease of use. The PGB part of the LTC budget is calculated by linear extrapolation of the development of the PGB expenditure of the previous year.

In practice, there exists some strategic behaviour for insurers in picking the care type mix that they use, since the long-term care service providers have a financial incentive to maximize their revenues. For example, in 2019 prices were adjusted for some care types in the care of people with disabilities. This led to a different care mix, which shows some scope for substitution in long-term care. This could on its turn lead to difficulties in forecasting budgets using only univariate models.

In this paper, we will offer alternatives to the current model and assess the performances of these models. We do reduce the scope of this paper to focus on only forecasting the in-kind LTC budget of Wlz to keep the research feasible within the given time frame. We extend the analysis to multivariate models, with the idea of enlarging the information set that we can use in forecasting, as we expect that trends between care types may comove.

We will therefore answer the following research questions: Do inter-series dependencies exist between long-term care services? And if so, what are the relationships between types of services? As a secondary question we would like to answer whether the usage of methods which incorporate inter-series dependencies will lead to better model performance.

The social relevance of this subject is that we gain insight in how changes in certain long-term care services may affect other long-term care services, which can aid in further research in policy-making, e.g. price setting, which NZa will research in the future. Furthermore, forecasting

a too low budget might lead to problems due to the right to access long-term care. However, setting a too high budget might lead to a decrease in efficiency of long-term care (Bakx and Wouterse, 2021). We find academic relevance in analyzing substitution and complement effects in long-term health care, which allows us to gain insight in how the actors in long-term health care set their care mix. Relevance for the NZa is found in offering an evaluation of the current methodology, and finding relations between time series for an upcoming research on the costs in long-term care.

We continue this paper with the theoretical background in Section 2, followed by a description of the data in Section 3. Then we explain the methodology behind the paper in Section 4, after which we show the results in Section 5. We conclude and discuss the findings of the paper in Section 6.

# 2 Theoretical Background

In this section we aim to place this thesis in the context of the existing literature. First, we look at the international differences in views on LTC and drivers of LTC costs. Then, we look at the literature of forecasting LTC budgets in 2.3, which gives us the framework for the methodology. Furthermore, we discuss the theory behind budget forecasting and substitutions in LTC in 2.4 and 2.5, which seem to be the major economic themes behind this problem.

## 2.1 International overview of long-term care

Long-term care is gaining importance as populations across the OECD are ageing, with projections estimating a doubling of the population over 80-years old. At the same time costs of LTC are very high, as low income individuals, who exceed the retirement age, with relatively low needs for LTC at home spend more than half of their disposable income on LTC (Hashiguchi and Llena-Nozal, 2020). This is further researched by Scheil-Adlung et al. (2015), who reviewed the weaknesses of LTC protection in 46 countries. They find that spending on LTC is still not a trivial question, as the global average of LTC expenditure accounts for less than one percent of GDP, while most persons aged 65+ that need LTC are at great financial risk due to high out-of-pocket payments. Denmark, The Netherlands, and Norway are the biggest relative spenders with LTC expenditures that exceed two percent of GDP. Furthermore, this article acknowledges that demand for LTC is expected to increase significantly due to an ageing demographic and that there exists a worldwide formal workforce shortage in LTC, which leaves informal workers to replace these gaps. This may have different side effects, such as fiscal effects (Geyer et al., 2017), due to forgone wages, or growing gender inequalities (Scheil-Adlung et al., 2015), due to a greater expectation of female family members to deliver informal health care.

Currently, countries face a trilemma in long-term care, where the first corner entails the coverage of needs, the second the extent of reliance on informal care, and the third being rising public expenditure (Pavolini, 2021). Six different models of long-term care social protection are identified, where the extent of state intervention varies, and a distinction is made between whether protection was provided through cash or through benefits. The Netherlands is one of a few countries where participation in an insurance for LTC is mandatory, where the patient needs to pay a relatively small contribution to fund their care, while most of the financial cost is borne by the Dutch Government. Along with The Netherlands, the Dutch LTC system is clustered under the same model with Denmark and Sweden by Pavolini (2021). This makes finding literature relevant to this thesis hard to find, as there are few countries sharing similarities to the Dutch system, however this also allows us to study these systems in greater depth. Additionally, comparing Dutch LTC to other systems is challenging, due to the transferral of LTC to multiple acts, with different conditions for eligibility for each act.

In Denmark, LTC is generally provided free-of-charge, due to financing via general taxation at both the national and the local level, which takes away concerns of affordability. However, accessibility of LTC is a greater issue in Denmark, having average waiting times of half a year in

2016 and 2018 (Commission et al., 2021). The Swedish LTC system is mainly funded by regional and municipal taxation, and is decentralised where the municipality bears the main responsibility for the delivery of health care. Incentives are placed to place care closer to the patient in both a physical and personal manner. Concerning the definition of long-term care in both Denmark and Sweden we need to remark that both of these countries consider a wider definition of LTC in comparison to The Netherlands, where they do not require that LTC patients have permanent access to care or require permanent supervision. Both systems do not explicitly forecast LTC costs on a yearly basis, which unfortunately means that we can not test their models against the current model of NZa (Astolfi et al., 2012a). Sweden does use a microsimulation model of the life course of the population of Sweden, which allows for the calculation of LTC costs. However, this model is used for a longer forecasting window than the model of NZa Brouwers et al. (2016).

## 2.2 Drivers of long-term care costs

The drivers of long-term care utilization were studied in The Netherlands, where Wong et al. (2010) find that increasing age, absence of a spouse, and disease have a positive effect on long-term care use. Note that this study was done during the previous system of long-term care in The Netherlands, which handled a different definition of long-term care, which included other services in addition to what is currently treated as long-term care in The Netherlands (Ministerie van Volksgezondheid, Welzijn en Sport, 2019). Furthermore, Colombo et al. (2011) identify four reasons which will likely affect the growth of future LTC costs in OECD countries: demographic transformations caused by aging populations; changing societal models, as family sizes decline and female participation in the formal labor market lead to a decrease in the availability of family care, while increasing the demand for paid care; demand of better quality of care, which is caused by the wealth increase of societies, being able to afford more expensive forms of care; and technological changes, which could lead to methods for better prevention or for more care being delivered at the same cost. Additionally, they predict that LTC costs relative to GDP at least double by 2050 due to these factors, using 2007 as a base year.

## 2.3 Long-term care costs forecasting

There exist vastly different ways of forecasting the costs of LTC, both in methodology, as in length of the forecasts. Most of the literature focuses on longer term projections of LTC costs, for example, Fukawa and Sato (2009) use simulation models at the macro level to make projections of the long-term care expenditures in Japan, while Fukawa (2011) does this at the household level. European Commission et al. (2015) use dependency rates along with population forecasts to make budgetary projections for LTC for all EU countries from 2013 to 2060. Lagergren et al. (2018) adapt this methodology to use empirical dependency rates from epidemiological studies instead of assumed dependency rates to make projections from 2010 through 2040 for Japan and Sweden. Astolfi et al. (2012a) compares health forecasting methods used by government agencies in OECD countries. The authors identify four classes of forecasting models: Microsimulation

models, which use characteristics and behaviours of a current sample to simulate future costs; component-based models, which analyses expenditures from all relevant actors, e.g., providers, financing agents, and individuals; macro-level models, which analyse aggregate health expenditures. This is a preferred class of forecasting models for short-term analyses when structural breaks are absent (European Commission et al., 2010) and exploits inertia of health expenditures (Getzen and Poullier, 1992); and combined models, which combines these approaches for a more flexible modelling approach. The current model used by NZa would be classified as a macro-level model, as aggregated health expenditures are modelled. Shorter term forecasts similar to the forecasts by NZa, i.e., forecast horizons of about one year, are relatively rare, as Astolfi et al. (2012a) find one forecast with similar forecasting windows to the forecasting window of NZa, namely The Canadian Institute for Health Information, who use ETS models to forecast health expenditure data. Getzen (2000) finds that the growth rate of the future growth in health care costs is best explained by the growth rate of the prior year.

In The Netherlands, two institutions make forecasts on the costs of LTC. The Dutch Healthcare Authority (NZa) provides relatively shorter forecasts of LTC to accordingly budget for the current year. The budgeted amount for declarations of in-kind care is predicted using error, trend, seasonality (ETS) models, while the amount of the PGB is calculated using linear extrapolation of the previous year (Nederlandse Zorgautoriteit, 2022). Then these are combined to calculate the LTC budget for that year. Additionally, The Netherlands Bureau for Economic Policy Analysis (Centraal Planbureau, CPB) publishes longer term forecasts, looking multiple years ahead using macroeconomic forecasts and applying calculating the effects of policy adjustments (Zeilstra et al., 2019). These macroeconomic forecasts contain five main growth components: The price developments of the GDP, the additional increases of real wages and real prices, the demographic state, the growth in income per capita, and remaining growth components, which contain factors, such as technology improvements and cultural developments. Then the current LTC costs are multiplied with the forecasted growth rates, after which policy effects are applied.

## 2.4 Budget forecasting

There exist many difficulties in forecasting the demand of health care in OECD countries, where intergenerational conflicts play a large role, mainly due to rising costs from pensions and health care (Fogel, 2018). In health care it is found that there is an increase in severity of conditions over time, and in costs to prevent worsening of conditions with increasing age. Furthermore, using U.S. data, Fogel (2018) identifies issues in forecasting health care costs due to an increase of severity of conditions over time, and an increase in costs to prevent worsening of those conditions with increasing age. In public budget making, Williams and Calabrese (2016) review the literature of budget forecasting and state that forecast errors in revenue forecasting can not only be attributed to how forecasts are established, but can also reflect political decisions, and finds that state and local governments in the U.S. have an underestimation bias to protect themselves against uncertainty in future revenue or set budget constraints. Astolfi et al. (2012b)

review forecasting models for health care expenditure in OECD countries and identify drivers of health expenditure: Demographic factors, income, health-seeking behaviour, treatment practices, technological progress, health prices and productivity, and organisation of the health care system. Furthermore, they mention that the credibility of these models rely on validity, accuracy, tractability, and transparency. In this paper we use models that adhere to these four principles, as this is highly valued in the public sector.

### 2.5 Substitution and complement effects in long-term care

As we have previously noted, there are some side effects coming forth of the need for long-term care. In this subsection we will discuss the literature that evaluates the further side effects due to policy.

In the literature there exists a relationship regarding the substitution of long-term care spending and other types of health care, e.g., Lu et al. (2020) find that the introduction of a long-term care medical insurance pilot in China alleviates some overcrowding of top- and second-tier hospitals. Additionally, the pilot has led to lower expenditure of patients and an improvement of delivery of care. Forder (2009) analyses the substitution effects between hospital and long-term care services in the United Kingdom and finds that every extra £1 invested in care home services leads to a decrease of £0.35 in hospital services and explains that this is not fully substituted, since not all admissions to LTC come from the hospital. Kattenberg and Bakx (2021) analyse the substitution effects of a reform in the municipality budget for domestic help in the Netherlands. They find that grant increases for domestic help increase the use of domestic help, while other types of home care decrease. Additionally, the authors find that the increases in spending on domestic help are neutralized by the decreases in spending on other types of long-term home care.

Additionally, there is evidence that substitution effects between formal and informal long term care exist. Bonsang and Schoenmaeckers (2015) find that having children are a factor in the supply of long-term health care, with a larger effect if they have daughters. Additionally, the authors find that people with children nearby decreases their propensity to purchase private long-term care insurance. Bremer et al. (2017) research the relationship between formal and informal caregiving for people with dementia in eight European countries. They find that informal caregiving is a substitute for home help and nurse visits and that there is a weaker complementary relationship between informal caregiving and outpatient visits. These relationships are also found when single-living elderly in Europe are examined (Bolin et al., 2008). Van Houtven and Norton (2004) find a slightly different result when examining these relationships in older adults in the United States, namely that informal care by children is a substitute for long-term care, hospital care, and physician visits, while being a complement to outpatient surgery.

# 3 Data

In this thesis we use data that is not publicly available, but are made accessible via a thesis internship at NZa. NZa retrieves data from Vektis, which collects health care data of insurers and health care service providers in The Netherlands. Due to time constraints, the scope of this paper is focused on in-kind care.

Vektis delivers monthly data on the amount of declarations for each care type at the insurer level from January 2015 onwards. In this thesis we will use time series data up until December 2021. The data by Vektis has then been cleaned and transformed by NZa to account for the number of days in a month, since an extra day in a month likely means an extra day that the care type needs to be delivered, which causes unnecessary fluctuation in the data. The data on the price of care types is supplied by the service providers. We choose to apply the prices of 2022 in our analysis to all years to avoid issues with indexing, which would complicate the thesis further, as the model in that case would also need to forecast indexing. Furthermore, declaration data and prices have been harmonized such that their units are congruent, i.e., when a certain declaration series is set in minutes that the according price is then set in euros per minute. Then the declaration and price data are multiplied such that the realized costs for each month for each care type are calculated at the level of the insurer.

The monthly data on the amount of declarations for each care type contains 405 care types for each insurer. Some series have missing data, due to these belonging to a care type introduced after 2015 and some care types are not delivered by some insurers, as there was no demand for that care type, leading to an empty time series. Therefore, in order to allow for a proper fit of the models, we require that a series would have at least 12 data points in the train set for the simulation of a letter. For example, for the letter in February 2020, we train the models using data up until 5 months prior to February 2020 due to data quality issues, which improve significantly after the fifth month due to later deliveries of data by insurers. In the case that the data until that starting moment does not have 12 data points for that care type at the insurer level, that series will be disregarded for that simulation of the letter and will also not be included in further aggregation steps to ensure equal footing when comparing the performance of models at different levels of aggregation. Additionally, we needed to truncate some series in the training set when using multivariate models, as these models do not handle missing data points well, therefore we truncate the data within each set of series to the length of the smallest length series in that set. Following these cleaning steps, we find the total budgets which the models need to estimate in Table 3, where the budgets in the February letters contain the budget over the full year, while the budgets in the July letters contain the budgets from March onwards.

To motivate the choice that we opt to use the daily budget, we show the monthly costs of long-term in-kind care in Figure 1. We notice that these series are severely affected by the amount of days in the month, hence that these are standardized to daily costs in each month.

We show these daily costs of long-term in-kind care calculated for each month in Figure 2. In the beginning of the series, we notice a decrease in costs, due to the transitioning to the Long

| | Realized budget |
|---|---|
| February 2019 | 22.5 |
| July 2019 | 18.8 |
| February 2020 | 22.6 |
| July 2020 | 18.8 |
| February 2021 | 23.5 |
| July 2021 | 19.7 |

Table 1: Total realized budget contained in each letter (in 2022 indexed $\times$€1,000,000,000).

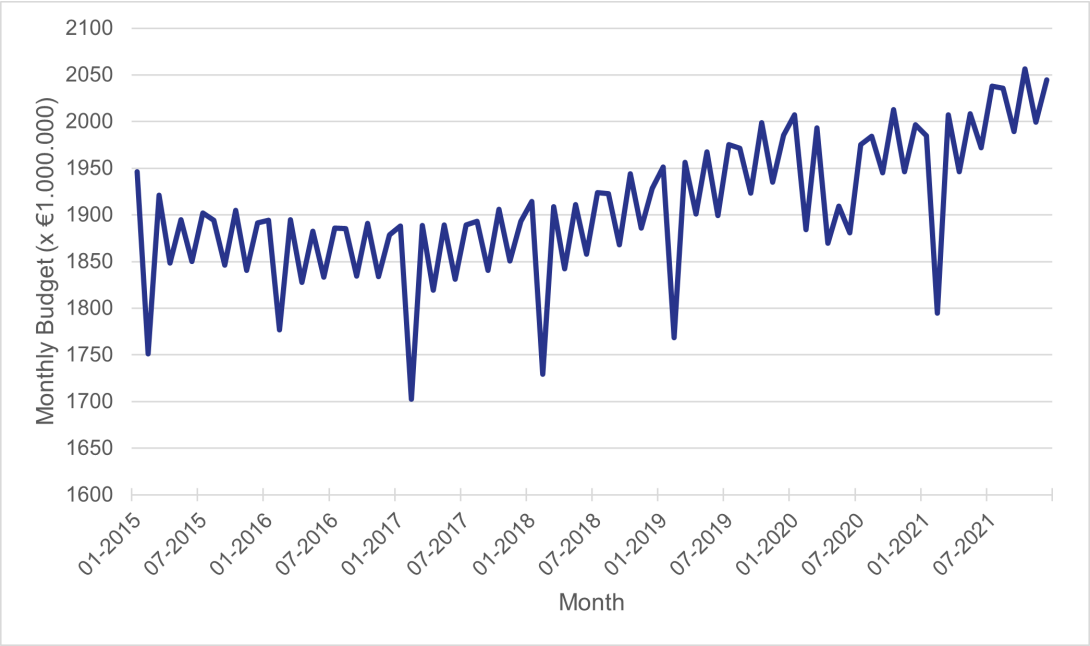


Figure 1: Monthly realized long-term in-kind care costs.

Term Care Act from the general law on exceptional medical expenses, which caused an outflow in nursing homes when the Wlz was introduced. This is followed by a steady fast increase in daily costs until spring of 2020, where a severe dip takes place due to SARS-Cov-2, with a fairly swift recovery afterwards.

In order to test the effects of the level of aggregation on the forecasts, we aggregate the data for each care type at the insurer level such that a time series is created, which contains the data for each care type at the national level.

Additionally, for the purpose of creating groupings based on the care needs assessments, we manipulated the data set such that each care type was assigned a single care needs assessment. This care needs assessment is required to have the largest share of the budget for this care type among all other care needs assessments. Then a time series is created, which aggregates the data of care types at the national level with the same assigned care needs assessment such that a time series is created, which contains the daily budget in each month for each care needs assessment at the national level. This requires that we use an auxiliary data set, which contains


Figure 2: Daily realized long-term in-kind care costs.

the value of declarations in 2019, grouped by care types and associated care needs assessments. Data regarding the care needs assessments are provided by CIZ, while the declaration data is provided by Vektis.

In Figure 3 we show the contents of the care needs assessment group VV7, which contains the care types that are assigned to VV7, as an example. A person who has been assigned the care needs assessment VV7 is someone who lives in a protected environment, where an intensive level of care is given, due to specific condition, with an emphasis on accompaniment. Examples of these specific conditions are dementia, Korsakov, and severe brain damage.[1] The care types V071, Z071, V073, and Z073 are assigned to VV7, as VV7 patients account for the largest amount of costs for these care types. This should not be a surprise as the third character contains a 7, which indicates for these specific care types that they are reserved for people with the seventh level of care needs. V and Z denote whether the care is given at home or at a care home, respectively. The 1 and 3 in the final character denote that a patient is receiving care with daytime activity and are distinctive in whether these patients do not or do receive treatment, respectively.

[1]More details about VV7 can be found in `https://www.ciz.nl/zorgprofessional/meer-informatie/factsheet-vv07` (in Dutch). Other descriptions of care groups assessments can be found in `https://wetten.overheid.nl/BWBR0036014/2022-04-15/0#BijlageA`.

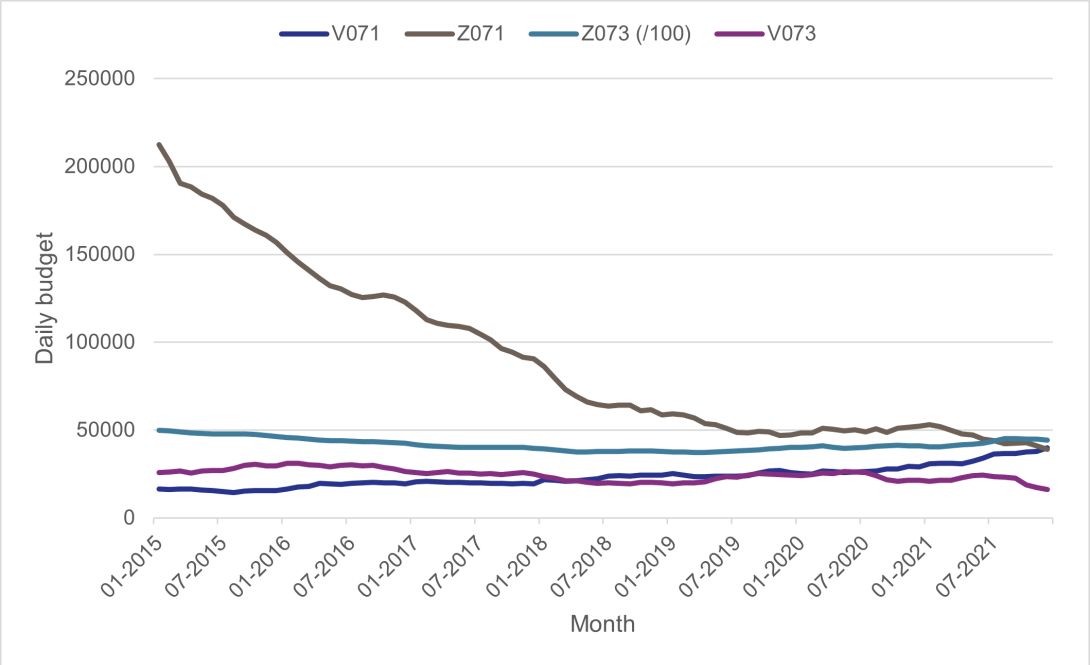


Figure 3: Daily realized long-term in-kind care costs in VV7.

# 4 Methodology

In this section we discuss the used econometric methods, where we start off with ETS, which is currently used by NZa. This is followed by its multivariate extension Vector ETS (VETS), which allows for commonalities between elements of the observations. Afterwards, We offer an alternative to the ETS class of models by discussing the ARIMA class of models. Then we conclude the econometric methods with the Vector Error Correction Model (VECM), which allows for the interpretation of short- and long-term effects of time series. We wrap up the chapter with discussing how we select model specifications in 4.5 and how we evaluate and test the forecasting performances of the models in 4.6 and 4.7. We apply the univariate models on three levels of aggregation: each series represents the costs of a care type at the insurer level (denoted without suffix); each series represents the costs of a care type at the national level (suffix *agg*); each series represents the costs of a group of care types at the national level (suffix *group*). The groups of care types are determined by the care needs assessments that are assigned to each care type, which is described in Section 3. We apply VETS to the first two types of series, while VECM is only applied on series at the insurer level, as the amount of series relative to the length of the data in a group would lead to us being unable to estimate some groups, which would account for a significant amount of the budget.

## 4.1 Error, Trend, Seasonality Models

Currently, the long-term care budget forecasts are predicted using Error, Trend, Seasonality (ETS) models. This class of models is able to capture trends and seasonality of a time series and creates forecasts as weighted averages of the previous level, trend, and seasonal effect and of the previous forecast of these factors. For some specifications of ETS models it has been shown to be equal to an ARIMA type model. However, since this is not the case for all ETS models, these are recognized as a different class of models. Many specifications are available, since we have different options to model the trend, seasonality and error of the time series. The taxonomy of ETS($\cdot$,$\cdot$,$\cdot$) models are described in Hyndman and Athanasopoulos (2018), among others. Here we will limit our discussion of different options to the options that are used in this paper. There are five options for the trend: No trend (N), additive (A), dampened additive (Ad), multiplicative (M), or dampened multiplicative (Md). Furthermore, there exist three options to deal with seasonality: No seasonality (N), additive seasonality (A), and multiplicative seasonality (M). Lastly, there are two ways to incorporate error terms: Additive (A), and multiplicative (M). We show an ETS(A,A,A) model as an example, being conceptually the simplest, but containing all factors in an additive manner:

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t, \tag{1}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t, \tag{2}$$

$$b_t = b_{t-1} + \beta\epsilon_t, \tag{3}$$

$$s_t = s_{t-m} + \gamma\epsilon_t, \tag{4}$$

where $y_t$ is the observed variable and $l_t$, $b_t$, and $s_t$ are respectively the level, trend and seasonal component of the time series $y$ at time $t$. Furthermore, $\epsilon_t$ is the error term at time $t$, $m$ is the period of the seasonality, and $\alpha$, $\beta$, and $\gamma$ are smoothing parameters for respectively the level, trend and seasonal component.

In contrast, we can also opt for a pure multiplicative ETS, containing only multiplicative factors, e.g., ETS(M,M,M):

$$\ln y_t = \ln l_{t-1} + \ln b_{t-1} + \ln s_{t-m} + \ln(1 + \epsilon_t), \tag{5}$$

$$\ln l_t = \ln l_{t-1} + \ln b_{t-1} + \ln(1 + \alpha\epsilon_t), \tag{6}$$

$$\ln b_t = \ln b_{t-1} + \ln(1 + \beta\epsilon_t), \tag{7}$$

$$\ln s_t = \ln s_{t-m} + \ln(1 + \gamma\epsilon_t). \tag{8}$$

This enforces positive values in the analysis, which is desirable from an interpretation standpoint in this case. However, this might lead to difficulties from a forecasting standpoint, as the derivation of the conditional expectations of $y_{t+h}$ is not trivial in the cases where $h$ is larger than 1 (Svetunkov, 2022). Thus, to work around this issue, we opt for using the `ets()` function from the `forecast` package, which uses the pure additive model in which it applies the log transformation, which gives rise to the following set of equations in the case of an ETS(A,A,A):

$$\ln y_t = \ln l_{t-1} + \ln b_{t-1} + \ln s_{t-m} + \epsilon_t, \tag{9}$$

$$\ln l_t = \ln l_{t-1} + \ln b_{t-1} + \alpha\epsilon_t, \tag{10}$$

$$\ln b_t = \ln b_{t-1} + \beta\epsilon_t, \tag{11}$$

$$\ln s_t = \ln s_{t-m} + \gamma\epsilon_t, \tag{12}$$

which has the advantage of having closed form expressions for the conditional mean and variance, and under some conditions show similar forecasts as the non-transformed pure multiplicative case (Svetunkov et al., 2022). In this case the trend and seasonal components can be interpreted as an average percentage increase or decrease, instead of the absolute increase or decrease that result from the additive models. An issue that arose in enforcing the log transformation (suffix *log*) is that it would estimate an additive ETS model when a zero value would be encountered in a series, which we could not separate from the successful estimations.

### 4.2 Vector ETS Model

The vector ETS (VETS) Model is a multivariate extension of the ETS model. Svetunkov et al. (2022) propose a taxonomy of VETS models, where this allows for commonality or individuality of trend or seasonal patterns in pure additive and multiplicative models. The authors devise a VETS(·,·,·)PIC(·,·,·) taxonomy, where the elements in VETS(·,·,·) are the same as in the univariate case, and the elements in PIC(·,·,·) correspond to commonalities in Parameters, Initial values, and Components. These three categories can contain Level, Trend, and Seasonality, while Parameters may also contain commonality for a Damping parameter. The pure additive model

can be formulated as:

$$\mathbf{y}_t = \mathbf{W}\mathbf{v}_{t-\mathbf{l}} + \epsilon_t, \tag{13}$$

$$\mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-\mathbf{l}} + \mathbf{G}\epsilon_t, \tag{14}$$

where $\mathbf{y}_t$ is a vector containing the time series of a group at time $t$, $\mathbf{v}_t$ contains the states of the time series, i.e., the error, trend, and seasonal components, $\mathbf{l}$ is a vector of lags of components, $\mathbf{W}$ is the measurement matrix, $\mathbf{F}$ is the transition matrix, $\mathbf{G}$ is the persistence matrix, $\epsilon_t$ is a vector containing the error terms, which we assume to follow the Multivariate Normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. This notation is general, since this allows for all forms of the model to be written in this form, at the cost of coming across as abstract. Therefore, we show a bivariate example using VETS(AAA)PIC(LTS,S,N):

$$y_{1,t} = l_{1,t-1} + b_{1,t-1} + s_{t-m} + \epsilon_{1,t}, \tag{15}$$

$$y_{2,t} = l_{2,t-1} + b_{2,t-1} + s_{t-m} + \epsilon_{2,t}, \tag{16}$$

$$l_{1,t} = l_{1,t-1} + b_{1,t-1} + \alpha\epsilon_{1,t}, \tag{17}$$

$$l_{2,t} = l_{2,t-1} + b_{2,t-1} + \alpha\epsilon_{2,t}, \tag{18}$$

$$b_{1,t} = b_{1,t-1} + \beta\epsilon_{1,t}, \tag{19}$$

$$b_{2,t} = b_{2,t-1} + \beta\epsilon_{2,t}, \tag{20}$$

$$s_{1,t} = s_{1,t-m} + \gamma\epsilon_{1,t}, \tag{21}$$

$$s_{2,t} = s_{2,t-m} + \gamma\epsilon_{2,t}. \tag{22}$$

For these two systems of equations to be equivalent, the following needs to hold true:

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathbf{v}_t = \begin{pmatrix} l_{1,t} \\ l_{2,t} \\ b_{1,t} \\ b_{2,t} \\ s_{1,t} \\ s_{2,t} \end{pmatrix}, \quad \mathbf{v}_{t-\mathbf{l}} = \begin{pmatrix} l_{1,t-1} \\ l_{2,t-1} \\ b_{1,t-1} \\ b_{2,t-1} \\ s_{1,t-m} \\ s_{2,t-m} \end{pmatrix},$$

Through the common elements in this framework, one can detect whether the series show commonalities, e.g., common trends or seasonality. When this is correctly specified, this will give an advantage in efficiency over ETS, as VETS can incorporate the data points of multiple series in the estimation of the parameters. This could for example lead to better estimations of trend or seasonality effects with a smaller length of the data series. Note that in VETS there is the underlying assumption that all series in the group would have the same ETS model specification. The `vets()` function also allows for the estimation of a block diagonal $\mathbf{\Sigma}$ using the parameter: `loss="diagonal"`. Svetunkov et al. (2022) found better performance in their simulations using the block diagonal $\mathbf{\Sigma}$, therefore we make the distinction between estimating a full $\mathbf{\Sigma}$ (no suffix), as opposed to a block diagonal $\mathbf{\Sigma}$ (suffix *diag*).

The pure multiplicative model may show some practical advantages, as this guarantees positive values for all variables and is formulated as:

$$\ln \mathbf{y}_t = \mathbf{W} \ln \mathbf{v}_{t-\mathbf{l}} + \ln \epsilon_t, \tag{23}$$

$$\ln \mathbf{v}_t = \mathbf{F} \ln \mathbf{v}_{t-\mathbf{l}} + \mathbf{G} \ln \epsilon_t, \tag{24}$$

This has the same advantages and issues as the univariate case in 4.1. In this case, there are only packages programmed for the pure additive and pure multiplicative cases for the `vets()` function in the `legion` package in R, described in Svetunkov et al. (2022). The pure multiplicative case is implemented as taking the logarithm over the data and then applying a pure additive model to the data. For example, VETS(MMdM)PIC(LTSD,S,N) is theoretically specified by (25)-(28), while this is implemented as (29)-(32) in `vets()`.

$$y_{i,t} = l_{i,t-1} b_{i,t-1}^{\phi} s_{t-m} \epsilon_{i,t}, \tag{25}$$

$$l_{i,t} = l_{i,t-1} b_{i,t-1}^{\phi} \epsilon_{i,t}^{\alpha}, \tag{26}$$

$$b_{i,t} = b_{i,t-1}^{\phi} \epsilon_{i,t}^{\beta}, \tag{27}$$

$$s_{i,t} = s_{i,t-m} \epsilon_{i,t}^{\gamma}. \tag{28}$$

$$\ln y_{i,t} = \ln l_{t-1} + \phi \ln b_{t-1} + \ln s_{t-m} + \epsilon_{i,t}, \tag{29}$$

$$\ln l_{i,t} = \ln l_{t-1} + \phi \ln b_{t-1} + \alpha \epsilon_{i,t}, \tag{30}$$

$$\ln b_{i,t} = \phi \ln b_{t-1} + \beta \epsilon_{i,t}, \tag{31}$$

$$\ln s_{i,t} = \ln s_{t-m} + \gamma \epsilon_{i,t}. \tag{32}$$

VETS shows similar issues in estimation when a zero is in a time series as ETS, however this would affect the estimation of the whole group, therefore we do not consider a VETSlog approach unlike for ETS, as this would affect most groups in the simulations of some letters. During the implementation of VETS we have come across some practical issues, as `vets()` throws errors when there are series that are perfectly correlated, and with series where the value of declarations does not change over time, i.e., a fixed amount of declarations over time. In the case of perfect correlation, we aggregate all the series in the group on which we apply `auto.arima()` to still achieve a forecast for this set of care types. When some series do not contain any variation, these series are forecast individually, for which we forecast the single amount that occurred in that series for the whole forecast horizon. Another issue that we came across was that certain groups contained only one series, i.e., there exist some care types that were only performed by one insurer, which also caused `vets()` to run into an error. These series were then estimated with `auto.arima()` due to its ease of use. On a budget level, the choice of this method should not have a significant effect as single series generally have a relatively small impact.

### 4.3 Autoregressive Integrated Moving Average Model

Another class of models that we can consider is the Autoregressive integrated moving average (ARIMA) models, which is one of the most widely used methods for time series forecasting,

together with the previously discussed exponential smoothing method (Hyndman and Athanasopoulos, 2018). ARIMA models can be split in three components: The autoregressive (AR) component, the integration (I) component, and the moving average (MA) component, which leads to the ARIMA($p$,$d$,$q$) notation. The first and third elements denote the amount of lags that are incorporated in the autoregressive and moving average components respectively, while the second element denotes the order of integration. Note that $p$, $d$, and $q$ are integers. In this class of models we are able to account for integration, while modeling mean reversion and shocks in a univariate time series, using the I, AR, and MA components, respectively. In this problem it will be likely to encounter some series with integration, as a person with a certain care type assessment which receives some set of care types in one month is likely to receive the same set of care types in the next month. This then carries over to the series on the levels of the insurer and the national level, where additional smoothing effects take place due to aggregating over multiple people. Furthermore, the model is relatively transparent and tractable, giving some indication for why this model would be a good candidate.

The ARIMA($p$,$d$,$q$) model with drift takes the following form:

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t, \tag{33}$$

where $y'$ is the $d$ differenced series of $y$. The $\phi$'s and $\theta$'s are the parameters of the model, and $c$ is a constant (Hyndman and Athanasopoulos, 2018).

This methodology can be extended to contain seasonal components, which would lead to an ARIMA($p$,$d$,$q$)($P$,$D$,$Q$)$_m$ framework. In this case we keep the notation from ARIMA($p$,$d$,$q$) but add $P$, $D$, and $Q$, which are the seasonal counterparts to $p$, $d$, and $q$ in ARIMA, and $m$, which is the period of the seasonality. From this follows the that arise from this take the following shape in the case of an ARIMA($p$,$d$,$q$)($P$,$D$,$Q$)$_m$:

$$\left(1 - \left(\sum_{i=1}^{p} \phi_i L^i\right)\right)\left(1 - \sum_{i=1}^{P} \Phi_i L^{im}\right)(1-L)^d \left(1-L^m\right)^D y_t = c + \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\left(1 + \sum_{i=1}^{Q} \Theta_i L^{im}\right)\epsilon_t, \tag{34}$$

where the terms keep the same interpretation as in (33), the capitalized parameters denote the seasonal counterparts of the ARIMA parameters, and $L$ denotes the lag operator.

### 4.4 Vector Error Correction Model

The Vector Error Correction Model (VECM) is a multivariate extension of the Error Correction Model and can be rewritten to both a Vector Autoregressive (VAR) Model and to a vector moving average representation. Therefore, with some hand-waving, VECM could be seen as a multivariate extension of ARIMA. VECM has the advantage of allowing for a relatively easy way to deal with cointegration in multivariate time series. This means that a linear combination of non-stationary series can be constructed such that this linear combination is stationary. In our case this can occur when the increase in two series are caused by a common driver. For example, one additional person with a certain care needs assessment could lead to an increase in the declarations of a set of care types. As previously mentioned, a time series of a single care

type may not be stationary due to persistence in LTC, however this series might be cointegrated with another series where, e.g., the increase in one series can be subtracted by a (scaled) increase or decrease in the other series to retrieve a stationary series. This could lead to a potentially better modelling of the dynamics in these series (Meuriot, 2015).

The VECM takes the following form:

$$\Delta \mathbf{y}_t = \mu_t + \Pi \mathbf{y}_{t-1} + \Gamma_1 \Delta \mathbf{y}_{t-1} + ... + \Gamma_{p-1} \Delta \mathbf{y}_{t-p+1} + \mathbf{u}_t, \tag{35}$$

where subscript $t$ denotes the time, $\mathbf{y}_t$ is a vector of observations, $\mathbf{u}_t$ is an error term, $\Delta$ is the lag operator, $\mu_t$ contains a deterministic term, $\Gamma$'s are autoregressive coefficient matrices, and $\Pi$ is a matrix containing error correction parameters. In the case that we know some long-term equilibrium relationships between some time series, $\Pi$ can be decomposed into two matrices $\alpha$ and $\beta$, where $\beta$ is the cointegration matrix, which contains information on the long term equilibria, while $\alpha$ is the loading matrix, which represent how fast the time series converge towards their long-term equilibrium (Lütkepohl, 2005). Therefore, this notation has the advantage that it allows for the derivation of long- and short term effects between different time series.

In specifying a VECM, we require knowledge of the lag order and the cointegration rank. Since lag order tests do not necessarily require a cointegration rank, while cointegration rank tests generally require a lag order, we will first determine the lag order of the system. Since a VECM with lag order $p-1$ corresponds to a VAR with lag order $p$, we can use tests that are devised for lag order selection in VAR. Since $\Gamma_p$ is not restricted to be non-zero, one can view the lag order $p$ as an upper bound for the true lag order. However as $p$ increases, the variance of the mean squared error of the one-step ahead forecast increases as well (Lütkepohl, 2005). This effect however diminishes with longer time series. Since the time series in our data set are relatively short, it would be prudent to not overestimate the lag order $p$, as this may lead to not having enough observations to estimate the parameters. Then one can choose from different criteria to minimize with respect to the lag order $p$, which will be discussed in Section 4.5.

Cointegration rank can be determined with several statistical tests, where we opt for the trace test by Johansen (1991), where a specification for $\mu_t$ needs to be chosen. We can specify $\mu_t = \mu_0$ in the case of a constant, $\mu_t = \mu_0 + \mu_1 t$ in the case of a trend, and $\mu_t = 0$ when neither is the case. The implications of choosing different model specifications can be found in Johansen (1994). In short, opting for a constant, non-zero deterministic term implies an affine trend in the series, and opting for an affine trend in the deterministic term implies a quadratic trend in the series.

A number of steps need to be taken to apply a VECM in this problem. First, we find the lag order by minimizing the AIC in VAR models in levels, using the `VAR()` function from the `vars` package. This function automatically selects the optimal amount of lags, using a information criterion that the user can specify. Then we test for cointegration in each group with the amount of lags using the Johansen test, which is performed in R by the `ca.jo()` function in the `urca` package. This calculates the Johansen trace statistics with which we can determine the amount

of cointegration relations and is also the function in which we need to input the form of $\mu_t$. Finally, we can use the found lag order and number of cointegration relations to specify the VECM, with which we can make forecasts. This is done using the `vec2var` function from the `vars` package, which uses the object resulting from `ca.jo()`.

During our implementation we encountered that it is not trivial to choose the form of the deterministic terms, i.e., a constant, a trend, or neither, to each group. An issue here is that we come across exploding forecasts in some variables (i.e., when the absolute eigenvalues of $\Pi$ are greater than 1, leading to a forecast with values that are increasing or decreasing very fast) and find that setting the deterministic terms to `'trend'` for all groups relieves this to some degree. However, it might not be realistic to assume that all care types contain a deterministic trend, which might decrease the efficiency of the models. Therefore, we intend to apply as few restrictions as possible, as long as the results stay within a certain threshold. We incorporate this for each series by first specifying the model without a constant or trend for the error to not exceed the threshold. This process is repeated by the model including a constant, after which we also allow for a trend. This approach should lead to non-exploding budgets, as we already know that the model specifications including a trend are bounded. However, this approach is sensitive to misspecifying the model when, for example, a constant is contained in the true model, while the model without deterministic term could meet the condition to be chosen. Multiple options for thresholds are possible, e.g., relative or absolute thresholds. We opt for a relative threshold, which is set at half of the realized budget for a given care type. This guarantees that the values are not exploding, however this comes at the cost of contaminating the forecast as this tunes the model towards the result.

### 4.5 Selection of model specifications

As we have seen, every model has many different specifications which we need to compare. To compare the different specifications of a certain model, we can use an information criterion to decide the model specification that will estimate a time series, or a group of time series in the case of multivariate models. There are two information criteria that can be calculated in all models, these are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), devised by Akaike (1974) and Schwarz (1978), respectively. There exist extensions which correct for sample size, however we will not discuss these in this paper, as it is unclear whether this will provide performance benefits and falls outside the scope of this paper. The AIC tends to be used when model fit is tested with the goal of forecasting, as this criterion is constructed to minimize the variance of the forecasting error. This does come with the theoretical drawback that the use of this criterion asymptotically leads to a positive probability of overestimating the amount of parameters. When the goal is to explain the data using a model, one usually prefers the BIC, this is because the BIC is a consistent estimator of the amount of parameters (Lütkepohl, 2005). These are defined below:

$$AIC = 2k - 2\ln(L), \tag{36}$$

$$BIC = k\ln(n) - 2\ln(L), \tag{37}$$

where $k$ is the amount of parameters estimated in the model, $n$ is the length of the time series, and $\ln(L)$ is the log-likelihood of a model. In both metrics, it is the objective to choose a model specification which minimizes the AIC or BIC. Then, the model specification with the lowest AIC or BIC is selected and will be fit on the data, after which the forecast is made. Since the aim of this paper is to test the forecasting power of the models, we opt for the AIC as the information criterion that we need to minimize to select between model specifications.

### 4.6 Model Evaluation

When assessing different forecasting models, it can be beneficial to calculate the forecasting errors that the models produce. The two letters that NZa writes to the Ministry of Health, Welfare, and Sport are written in February and July from 2022 onwards, so we will be reconstructing letters written in February and July from 2019 to 2021. We start at 2019, as this allows for sufficiently long time series to estimate a significant share of the available time series. A challenge in this task is that NZa makes forecasts using data from 5 months prior, because of data quality issues. This leads to making 15- and 10-month ahead forecasts for the respective letters. We focus on evaluating errors on the whole budget, rather than on some individual care types, as this approach then aligns with the objective of NZa to forecast the total LTC budget for the whole year. This leads to the following calculation of the budget error:

$$\text{Budget error} = \sum_i \sum_t \hat{y}_{i,T+t} - y_{i,T+t}, \tag{38}$$

where $y_{i,T+t}$ and $\hat{y}_{i,T+t}$ denote the realized and estimated value of declarations of care type $i$ at time $T+t$, respectively. Here $T$ denotes the time at which the forecasting takes place and $h$ denotes the amount of months that we forecast ahead. Note that the scope of $t$ is not always starting at 1. In the case of the February letter, we forecast from October of the previous year onwards, but our focus lies on the budget error in the current year, hence we discard the forecast errors made in the previous year, which leads $t$ to range from 4 to 15. In the July letter, we forecast from March onwards, leading to a range of 1 to 10 for $t$.

However, NZa also finds it of importance to select a model where the forecasts do not cancel each other out, e.g., a model with large errors on the series level that cancel each other out would lead to a small budget error, but this is deemed as undesirable by NZa. Therefore, we also calculate the sum of absolute deviations ($SAD$) on the care type level, we calculate this as:

$$SAD = \sum_i \left| \sum_t \hat{y}_{i,T+t} - y_{i,T+t} \right|. \tag{39}$$

This measure captures the absolute errors within series and allows us to observe whether the errors in the estimated budget cancel each other out.

To condense these measures to a single statistic, we calculate a Mean Absolute Error (MAE)

type statistic over the obtained budget errors as follows:

$$MAE_{\text{Budget error}} = \sum_i \frac{|\text{Budget error}_i|}{n}, \tag{40}$$

where $\text{Budget error}_i$ is the budget error obtained for the letter written at moment $i$, and $n$ is the count of $i$, which is equal to the amount of simulated letters, equal to 6 in this case. For the SAD we can calculate the mean, as the SAD is non-negative by definition.

### 4.7 Statistical testing of evaluation metrics

Now that evaluation metrics can be obtained of several models, we may apply tests to these metrics to see whether the observed differences between models show statistical significance. As it is our objective to minimize the error on the budget level, we use the series of budget errors as the statistic on which we apply the tests. Furthermore, in this thesis we mainly want to evaluate the performance of the current ETS model against the alternative models presented in this thesis, leading to paired tests. It is also desirable to find whether there is a model which performs best when compared to all other models. Therefore we also test each model pairwise against the other models. We choose the 10% significance level, due to the small amount of budget errors that we obtain. We will mainly discuss two types of tests, one non-parametric test, which will test two samples for whether they are drawn from differing distributions and whether one of these distributions is greater than the other. The other will test whether the variances from two samples differ in a statistically significant manner, which will allow us to test whether a method would have a greater dispersion of the budget errors, as a smaller variance of the error may be preferred when the means are equal.

Firstly, we can apply a Mann-Whitney U test (Mann and Whitney, 1947), which statistically tests from two samples, $X$ and $Y$, which in our case can be interpreted as the budget errors resulting from two methods, from respective distributions $f$ and $g$ whether these distributions $f$ and $g$ are equal. This is a useful test, as this is a non-parametric test and this makes that we do not have to make assumptions on the distribution of the errors on the budget level. It is calculated as follows:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j), \tag{41}$$

where $S(X_i, Y_j)$ is a function which assigns values $S(X_i, Y_j) = 1$ if $X_i > Y_j$, $S(X_i, Y_j) = 0$ if $X_i < Y_j$, and $S(X_i, Y_j) = 0.5$ if $X_i = Y_j$, where $X_i$ and $Y_j$ are sample points from groups $X$ and $Y$, respectively. Under the null hypothesis $U$ is distributed with the use of combinatorics, where we can use a reference table to look up critical values (Wackerly et al., 2012). In this procedure we do need to ensure that we take the absolute values of the errors on the budget level, as the U test is a test, which uses order statistics. This may lead to some confusing results if the absolute values are not taken, for example suppose distributions $f$ and $g$. Let the sample taken from $g$ contain strictly positive values, while the sample taken from $f$ contains strictly positive values except for one sample point $i \in \mathbb{R}$. In this case when $i$ would be a large negative value, it would

be treated the same as when $i$ would be the smallest value of both the samples, due to the U test using order statistics. However, we would like to prevent the possibility of such a case occurring and therefore we would use the absolute value of the obtained budget errors.

Secondly, we may add an assumption of the models being unbiased. This leads to the distribution of the budget errors to have mean zero. Then we can compare whether the variances differ significantly. Several tests are able to test this, e.g. the F-test of equality of variances (Snedecor and Cochran, 1992), Bartlett's test (Bartlett, 1937), and Levene's test (Olkin and Levene, 1960). We prefer to use Levene's test in this case, as the F-test and Bartlett's test heavily depend on normality of the distributions, while Levene's test is more robust to non-normality. Note that in the default Python implementation of Levene's test in SciPy (Virtanen et al., 2020) the median is used in the calculation of $Z_{..}$, which was proposed by Brown and Forsythe (1974), which we will call the Brown-Forsythe test from here on, but it is possible to change the default implementation to use the original Levene's test, which uses the mean in the calculation of $Z_{..}$. The advantage of the Brown-Forsythe test is that this test is generally better able in handling skewed distributions, while Levene's test generally is preferred when the distributions are symmetric and moderate tailed. For completeness, we include all four methods of testing the variance of the SAE in our analysis and report whether major differences occur, however as we prefer the Levene's type tests, the results for the F- and Bartlett's test will be reported in the Appendix. It is however impossible to calculate Levene's/Brown-Forsythe test with a user specified mean or median for both distributions, as the numerator of the test would then always be equal to zero. This can be seen in the calculation of the test statistic:

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} N_i (Z_{ij} - Z_{i\cdot})^2}, \tag{42}$$

where $W \sim F(k-1, N-k)$ is the test-statistic for the Levene's and Brown-Forsythe tests, $k$ is equal to the number of groups that are included in the test, which in our case is always equal to two, as we only perform paired tests. $N_i$ is equal to the number of data points in group $i$, which is fixed at six. $N = \sum_i N_i$ is the total amount of data points included in the test, in our case equal to twelve. $Z_{\cdot\cdot}$ is set to be the mean of all data points, while $Z_{i\cdot}$ is equal to the mean of the members in group $i$. $Z_{ij}$ is the distance of the $j$-th data point to the specified metric of group $i$, i.e., mean or median. In the case that the mean is specified, $Z_{ij} = |Y_{ij} - \bar{Y}_i|$, where $Y_{ij}$ is the $j$-th data point of group $i$, and $\bar{Y}_i$ is the mean of group $i$. We see that the numerator calculates the sum of the differences between the means of each group and the total mean. However, when we define the means ourselves and set these all equal to zero, this would be a summation of zeroes for any input of groups, therefore we are unfortunately not fully able to make the assumption of zero mean in this test.

When we make an additional assumption that the budget errors are distributed normally, we can use the previously mentioned F-test of equality of variances and Bartlett's test. We do not show any preference to either test, as the F-test is designed to only test for equality in variance of two samples, while Bartlett's test is able to generalize the test of equality of variance

across more than two samples, however both operate under similar assumptions relevant for our problem, namely normality of the distribution of the budget errors. The F-test is calculated as follows:

$$F = \frac{S_X^2}{S_Y^2}, \tag{43}$$

where $S_X^2$ and $S_Y^2$ are the sample variances for $X$ and $Y$, respectively, and $F$ is the test statistic which follows the $F(N_X - 1, N_Y - 1)$ distribution. Bartlett's test is calculated via

$$\chi^2 = \frac{(N-k)\ln\left(S_p^2\right) - \sum_{i=1}^{k}(N_i - 1)\ln\left(S_i^2\right)}{1 + \frac{1}{3(k+1)}\left(\sum_{i=1}^{k}\left(\frac{1}{N_i - 1}\right) - \frac{1}{N-k}\right)}, \tag{44}$$

where $\chi^2$ is Bartlett's test statistic which approximately follows the $\chi^2(k-1)$ distribution, $S_p^2$ denotes the pooled sample variance, while $S_i^2$ is the sample variance of group $i$.

Lastly, it would be possible to relax the assumption that the budget errors have a distribution of mean zero and perform a t-test (Student, 1908), which requires other assumptions such as homogeneity of variance and an approximate normal distribution. This allows us to compare directly whether one model type has a statistically significant lower MAE value. However, due to the small sample size, the results of these tests will likely contain a lot of uncertainty, and therefore we choose not to apply the t-test in this thesis.

In the case of the variance tests it could also be beneficial to analyse whether the absolute values of the budget errors lead to different results when compared to the realized budget errors, as the gap between positive and negative values will inflate the variances, which may bias these results towards non-rejection of the null hypothesis. In this case we definitely prefer the Levene's and Brown-Forsythe test statistics, as these are more robust against non-normality, as normality will most likely not hold due to the absolute transformation. However, we also report the results of the two other test statistics in the Appendix for completeness.

# 5 Results

In this section we discuss the results of the paper. We have simulated the February and July letters from 2019 until 2021. The series of care types that we include in our analysis are required to have at least 12 months of data for the model to train on and we require a full test set, i.e., the year of the letter that we simulate needs to have no missing data points.

## 5.1 ETS

We apply the ETS framework at three levels of aggregation: data on the insurer, national, and care needs assessment level (denoted as ETS, ETSagg, ETSgroup, respectively).

We do not research the parameter estimates, but only consider the model specifications, due to the amount of parameters nested in these models, which would not be trivial to visualize and deduce useful insights from. We show the count of model specifications in Appendices A to F. This gives us some indication of the stability of estimations over time. In general, we can see that the amount of series increase over time, due to more and longer time series of data being available. Furthermore, the specifications seem to show a shock leading to a shift towards ETS(A,N,N) specifications in 2021, which may be explained due to the effects of SARS-Cov-2, as this would likely lead to a decrease in fit of all model specifications for certain series, after which a simpler model would be preferred due to our penalty for more parameters in the calculation of the AIC.

## 5.2 Vector ETS

Following ETS, we now treat its multivariate approach. Here, the main challenge in applying this method lies in finding useful groupings. In this paper we treat two of those groupings. The first grouping collects the series of a single care type of all insurers. The second grouping is described in Section 3 and collects the aggregated care types with the same allocated care needs assessment.

We find the distributions of specifications of VETS models in Appendices G to J. In this case it is much less clear to notice the effects of a disturbance due to SARS-Cov-2 when compared to ETS. Additionally, one may notice the differences in the amount of estimated models of VETSagg, as opposed to VETSaggdiag. This is caused due to a combination of relatively small series lengths in some cases with groups being filled with a greater amount of series. Then in the case of diagonal $\Sigma$, there are less parameters needing to be estimated, which causes less groups to be underidentified.

Finally, we show an example of VETS using the group VV7, which is mentioned previously in the paper. Regular VETS estimates VETS(MMdM)PIC(LTSD,S,N), which means that the model contains multiplicative error, trend, and seasonality, where the trend is dampened. The model has the same parameters for level, trend, seasonality, and dampening, but only share a common initialization of the seasonality level. The VETS model with diagonal $\Sigma$ estimating VV7 is a VETS(MMM)PIC(LTS,S,N). This model specification is equal to the previous example

without the dampened trend. In (25) to (28) the specification form is shown, however due to the implementation of `vets()` this will in reality follow (29) to (32). Note in the case without dampened trend $\phi$ is equal to 1.

Table 5.2 shows the parameter estimations of the forecasts of VETSagg and VETSaggdiag for the care types attributed to VV7 in the simulated letter for February 2020. Note that February is the reference month for seasonality, being denoted as the twelfth month, which leads to month 1 corresponding to March, month 2 corresponding to April, etc. We can interpret this seasonal parameter as $\exp(s_i)$ times the baseline, e.g., assuming a scenario in which zero model error occurs, due to updating of the seasonal parameters by $\gamma\epsilon_{i,t}$, we find using the model considering a full $\Sigma$ that on average the estimated values in March is due to seasonality $\exp(-0.0130) \approx 0.987$ times the baseline value in February, indicating a decrease of about 1.3% due to seasonality in March.

Furthermore, note that only $\alpha, \beta, \gamma$, and $\phi$ remain constant in these estimations and that the other values are estimates of the initial values, which are allowed to move independent of each other, influenced by their errors in $\epsilon_{i,t}$. We observe high values of $\alpha$ in both cases, implying that the trend is easily influenced by $\epsilon$, while relatively smaller values for $\beta$ and $\gamma$ are estimated, which implies that these values are relatively more persistent. In the case of full $\Sigma$, we find $\phi$ to be within the plausible range of values between 0.8 and 0.98, stated in Hyndman and Athanasopoulos (2018), meaning that some damping effect of the trend is captured in this specification. The interplay between $\beta$ and $\phi$ also explains why the estimated values for $\beta$ differ.

| | Full | Diagonal | | Full | Diagonal |
|---|---|---|---|---|---|
| $\alpha$ | 0.9931 | 0.9942 | $s_1$ | -0.0130 | -0.0126 |
| $\beta$ | 0.1832 | 0.2413 | $s_2$ | -0.0109 | -0.0103 |
| $\gamma$ | 0.1375 | 0.1344 | $s_3$ | -0.0048 | -0.0042 |
| $\phi$ | 0.9346 | 1 | $s_4$ | -0.0005 | 0.0004 |
| $l_1$ | 9.7880 | 9.8010 | $s_5$ | 0.0028 | 0.0033 |
| $l_2$ | 11.9139 | 11.8983 | $s_6$ | 0.0101 | 0.0097 |
| $l_3$ | 15.3368 | 15.3330 | $s_7$ | 0.0128 | 0.0119 |
| $l_4$ | 10.3543 | 10.3577 | $s_8$ | 0.0124 | 0.0120 |
| $b_1$ | 0.0212 | 0.0094 | $s_9$ | 0.0066 | 0.0057 |
| $b_2$ | -0.0434 | -0.0048 | $s_{10}$ | 0.0007 | 0.0004 |
| $b_3$ | -0.0141 | -0.0076 | $s_{11}$ | -0.0042 | -0.0047 |
| $b_4$ | -0.0247 | -0.0321 | | | |

Table 2: VETS estimation results for VV7 in February 2020.

## 5.3 ARIMA

We apply ARIMA in the same way as in the case of ETS, i.e., estimating individual time series aggregated at the insurer level, the national level, and care assessment needs level. We show

the number of parameters in our ARIMA specifications, together with the specific estimated models, in Appendices K to M. We find that the proportion of ARIMA(0,0,0) specifications shows a decreasing trend, which may be attributed to more data being available in some time series. However, we do not find clear shifts in the amount of estimated parameters. It must also be noted that none of the specifications contain seasonal components.

### 5.4 VECM

The first and only type of grouping that we apply for VECM is the same as the first grouping in VETS, where we group the same care types at the level of the insurer.

As an example, we show the process behind modelling the budget for care type Z073. Care type Z073 entails patients with a VV7 care needs assessment, receiving care and treatment in combination with daytime activity. The series can be seen in Figure 3 in Section 3. Using the AIC, we obtain that $p = 4$ lags is optimal. Then using 4 lags, we perform a Johansen test with a constant term. We then find that 3 cointegration relations are detected at the 5%-level, as this is the highest amount of cointegration relations in which the test statistic exceeds the critical value at the 95% level. These test statistics with its associated critical values are shown in Table 40 of Appendix N. Now we can estimate the VECM, shown in (45) in Appendix N. The used implementation of VECM decomposes $\Pi$ in $\alpha$ and $\beta'$, which allows us to see the cointegration relations in $\beta'$. Note that there are three lags shown in (45), while we found that $p = 4$. This is due to differencing, which includes an extra lag in the estimation, as $\Delta y_{t-3} = y_{t-3} - y_{t-4}$.

We tried to apply a second grouping with time series of care types at the national level similar to the grouping performed in VETS. However, we find that the model for a significant amount of groups can not be estimated, due to a too small data length relative to the amount of series that are contained in a group. This would lead to omission of relatively large parts of the budget. Therefore, we did not pursue further research of this model at this level of aggregation.

### 5.5 Comparison of forecast errors

The SAE and SAD values of the reconstructed February and July letters with their subsequent MAE values are shown in Tables 3 and 4, respectively. These tables contain vertical lines to demarcate the level of data aggregation for each model, where the first block of models uses data for each performance of each insurer. This is followed by models using data for each performance on the national level and concluded by models which are aggregated at the level of each care assessment indication. When solely considering the SAE values, we find that the VECM shows the best performance, followed by ETS aggregated at the group level, ETS aggregated at the national level, and VETS grouping care types within a care needs assessment.

When considering the SAD values, we can clearly see that the level of aggregation affects the level of the SAD values, with higher levels of aggregation leading to lower values of the SAD. This makes intuitive sense, as a higher level of aggregation means that there exist fewer opportunities for errors to cancel each other out, due to less series being estimated.

In Table 5 we can see the p-values of whether two paired methods show equal distributions according to the Mann-Whitney U test. From this table we can not conclude that any implementations of models show a statistically significant difference in distribution of the SAE values. From Tables 6 and 7 we see that there is no statistical significant difference of the variance of the SAE values between all methods when using the Levene's and Brown-Forsythe tests. However, when considering the Levene's and Brown-Forsythe tests of equal variances over the absolute values of budget errors in Tables 8 and 9, we see that the VECM has a statistically significant difference in variance when compared to most alternative methods with only VETS, VETSdiag, and ETSgroup being the only tests which do not reject the null hypothesis at the 10% level in the case of the Brown-Forsythe test, while in the case of Levene's test for VECM, we find that the null hypothesis is rejected at the 10% level for all methods except for VETS and VETSdiag as well as the test between VETS and VETSgroup. In Table 10 we show the standard deviations for the realized and absolute values of the SAE. When considering the rejections of both the Brown-Forsythe and the Levene's test, this implies that the standard deviation of the absolute values of the SAE by VECM is statistically significant smaller than the standard deviations of the absolute values of the SAE by all other tested methods, except for VETS, VETSdiag, and ETSgroup. The results of the F- and Bartlett's tests are reported in Appendix P, where we find that none of the paired tests show a statistically significant difference in variance of the budget errors.

When taking all results of the statistical tests into account, we may not conclude that there is a single method showing better performance in calculating the total budget required for the rest of the year. VECM does show some improvement in performance when compared to the current methodology in the variance of the absolute values of the SAE, however the Mann-Whitney U test between VECM and ETS is not rejected at the 10% level and the average SAD values of VECM are the highest of all model types, which therefore does not give convincing evidence that VECM outperforms ETS. As this is the case for all methods, we can not point towards one best performing method as a result from these simulations.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| February 2019 | -148.00 | -49.45 | 87.53 | -366.84 | -312.93 | -247.09 | -148.11 | -125.76 | -118.24 | -137.51 | -64.73 | -39.18 | -221.12 | -125.63 |
| July 2019 | -63.12 | -57.97 | -85.76 | -100.79 | -51.03 | -243.32 | -46.35 | -86.38 | -64.77 | -62.93 | -45.75 | -66.09 | -47.46 | -102.17 |
| February 2020 | 561.72 | 549.67 | 493.41 | 367.81 | 382.94 | 340.11 | 500.04 | 550.50 | 621.17 | 530.72 | 528.16 | 422.24 | 428.18 | 671.09 |
| July 2020 | 540.29 | 547.14 | 585.60 | 441.32 | 541.17 | 458.54 | 527.47 | 556.21 | 601.88 | 513.50 | 575.86 | 507.52 | 552.96 | 530.78 |
| February 2021 | -106.89 | -188.48 | -131.91 | -22.52 | 35.38 | -63.11 | 68.13 | 148.43 | 37.26 | -102.33 | 62.68 | 212.12 | -102.74 | -143.72 |
| July 2021 | -414.33 | -366.45 | -388.75 | -478.54 | -478.49 | -295.77 | -385.81 | -365.60 | -399.72 | -571.11 | -416.35 | -425.84 | -515.31 | -394.04 |
| MAE | 305.73 | 293.19 | 295.49 | 296.30 | 300.32 | 274.66 | 279.32 | 305.48 | 307.17 | 319.68 | 282.25 | 278.83 | 311.30 | 327.91 |

Table 3: Budget errors of the estimated models (× €1,000,000)

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| February 2019 | 1162.64 | 1255.62 | 1210.21 | 1146.02 | 1087.05 | 1317.56 | 731.34 | 815.85 | 867.92 | 547.54 | 678.79 | 333.95 | 362.43 | 335.54 |
| July 2019 | 606.71 | 616.69 | 640.50 | 450.79 | 453.65 | 1402.82 | 362.89 | 373.17 | 472.83 | 310.29 | 325.76 | 220.31 | 245.77 | 250.36 |
| February 2020 | 1445.13 | 1452.25 | 1558.33 | 1275.41 | 1244.55 | 1513.36 | 1027.61 | 1172.34 | 1202.42 | 1002.58 | 1045.92 | 579.51 | 576.12 | 796.46 |
| July 2020 | 922.52 | 910.83 | 977.48 | 811.41 | 842.74 | 1053.03 | 691.80 | 725.33 | 781.35 | 706.33 | 797.07 | 573.59 | 591.27 | 583.55 |
| February 2021 | 1219.62 | 1297.11 | 1350.75 | 1032.33 | 1054.01 | 1247.27 | 796.98 | 839.27 | 889.15 | 1079.51 | 999.10 | 757.66 | 490.73 | 596.32 |
| July 2021 | 827.91 | 922.59 | 800.08 | 777.03 | 778.13 | 821.19 | 586.71 | 674.93 | 602.58 | 998.79 | 665.54 | 536.82 | 622.67 | 454.63 |
| Mean | 1030.75 | 1075.85 | 1089.56 | 915.50 | 910.02 | 1225.87 | 699.56 | 766.82 | 802.71 | 774.17 | 752.03 | 500.31 | 481.50 | 502.81 |

Table 4: SAD values of the estimated models (× €1,000,000)

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.82 | 0.94 | 0.70 | 0.70 | 0.82 | 0.70 | 0.94 | 1.00 | 0.94 | 0.70 | 0.82 | 0.94 | 1.00 |
| ETSlog | 0.82 | 1.00 | 0.82 | 1.00 | 0.82 | 1.00 | 0.82 | 0.70 | 0.82 | 0.82 | 1.00 | 0.94 | 0.94 | 0.82 |
| ARIMA | 0.94 | 0.82 | 1.00 | 0.82 | 0.70 | 0.82 | 0.82 | 0.94 | 1.00 | 0.82 | 0.59 | 0.82 | 0.94 | 0.59 |
| VETS | 0.70 | 1.00 | 0.82 | 1.00 | 0.94 | 0.59 | 0.82 | 0.94 | 0.82 | 0.48 | 0.94 | 1.00 | 0.70 | 0.59 |
| VETSdiag | 0.70 | 0.82 | 0.70 | 0.94 | 1.00 | 0.70 | 0.94 | 0.70 | 0.70 | 0.59 | 0.82 | 1.00 | 0.82 | 0.70 |
| VECM | 0.82 | 1.00 | 0.82 | 0.59 | 0.70 | 1.00 | 0.94 | 0.82 | 0.94 | 0.82 | 1.00 | 1.00 | 0.94 | 0.82 |
| ETSagg | 0.70 | 0.82 | 0.82 | 0.82 | 0.94 | 0.94 | 1.00 | 0.59 | 0.94 | 0.59 | 1.00 | 0.94 | 0.70 | 0.59 |
| ETSagglog | 0.94 | 0.70 | 0.94 | 0.94 | 0.70 | 0.82 | 0.59 | 1.00 | 0.94 | 0.94 | 0.59 | 0.70 | 0.94 | 1.00 |
| ARIMAagg | 1.00 | 0.82 | 1.00 | 0.82 | 0.70 | 0.94 | 0.94 | 0.94 | 1.00 | 1.00 | 0.70 | 1.00 | 1.00 | 0.70 |
| VETSagg | 0.94 | 0.82 | 0.82 | 0.48 | 0.59 | 0.82 | 0.59 | 0.94 | 1.00 | 1.00 | 0.59 | 0.48 | 0.94 | 0.82 |
| VETSaggdiag | 0.70 | 1.00 | 0.59 | 0.94 | 0.82 | 1.00 | 1.00 | 0.59 | 0.70 | 0.59 | 1.00 | 1.00 | 0.82 | 0.48 |
| ETSgroup | 0.82 | 0.94 | 0.82 | 1.00 | 1.00 | 1.00 | 0.94 | 0.70 | 1.00 | 0.48 | 1.00 | 1.00 | 0.48 | 0.70 |
| ETSgrouplog | 0.94 | 0.94 | 0.94 | 0.70 | 0.82 | 0.94 | 0.70 | 0.94 | 1.00 | 0.94 | 0.82 | 0.48 | 1.00 | 0.94 |
| ARIMAgroup | 1.00 | 0.82 | 0.59 | 0.59 | 0.70 | 0.82 | 0.59 | 1.00 | 0.70 | 0.82 | 0.48 | 0.70 | 0.94 | 1.00 |

Table 5: p-values of Mann-Whitney U tests

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.96 | 0.94 | 0.98 | 0.92 | 0.84 | 0.97 | 0.89 | 0.89 | 0.94 | 0.99 | 0.97 | 0.95 | 0.96 |
| ETSlog | 0.96 | 1.00 | 0.90 | 0.93 | 0.87 | 0.88 | 0.99 | 0.84 | 0.85 | 0.90 | 0.97 | 0.99 | 0.91 | 0.93 |
| ARIMA | 0.94 | 0.90 | 1.00 | 0.96 | 0.97 | 0.76 | 0.90 | 0.94 | 0.94 | 0.98 | 0.92 | 0.89 | 1.00 | 0.99 |
| VETS | 0.98 | 0.93 | 0.96 | 1.00 | 0.93 | 0.79 | 0.94 | 0.89 | 0.90 | 0.95 | 0.96 | 0.93 | 0.96 | 0.98 |
| VETSdiag | 0.92 | 0.87 | 0.97 | 0.93 | 1.00 | 0.73 | 0.87 | 0.97 | 0.96 | 0.99 | 0.90 | 0.85 | 0.97 | 0.97 |
| VECM | 0.84 | 0.88 | 0.76 | 0.79 | 0.73 | 1.00 | 0.85 | 0.69 | 0.73 | 0.78 | 0.84 | 0.84 | 0.78 | 0.82 |
| ETSagg | 0.97 | 0.99 | 0.90 | 0.94 | 0.87 | 0.85 | 1.00 | 0.83 | 0.85 | 0.90 | 0.98 | 1.00 | 0.91 | 0.93 |
| ETSagglog | 0.89 | 0.84 | 0.94 | 0.89 | 0.97 | 0.69 | 0.83 | 1.00 | 0.99 | 0.96 | 0.86 | 0.81 | 0.94 | 0.94 |
| ARIMAagg | 0.89 | 0.85 | 0.94 | 0.90 | 0.96 | 0.73 | 0.85 | 0.99 | 1.00 | 0.96 | 0.87 | 0.83 | 0.94 | 0.94 |
| VETSagg | 0.94 | 0.90 | 0.98 | 0.95 | 0.99 | 0.78 | 0.90 | 0.96 | 0.96 | 1.00 | 0.92 | 0.89 | 0.98 | 0.98 |
| VETSaggdiag | 0.99 | 0.97 | 0.92 | 0.96 | 0.90 | 0.84 | 0.98 | 0.86 | 0.87 | 0.92 | 1.00 | 0.98 | 0.93 | 0.95 |
| ETSgroup | 0.97 | 0.99 | 0.89 | 0.93 | 0.85 | 0.84 | 1.00 | 0.81 | 0.83 | 0.89 | 0.98 | 1.00 | 0.90 | 0.93 |
| ETSgrouplog | 0.95 | 0.91 | 1.00 | 0.96 | 0.97 | 0.78 | 0.91 | 0.94 | 0.94 | 0.98 | 0.93 | 0.90 | 1.00 | 0.99 |
| ARIMAgroup | 0.96 | 0.93 | 0.99 | 0.98 | 0.97 | 0.82 | 0.93 | 0.94 | 0.94 | 0.98 | 0.95 | 0.93 | 0.99 | 1.00 |

Table 6: p-values of the paired Brown-Forsythe tests.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.93 | 0.79 | 0.74 | 0.82 | 0.56 | 0.70 | 0.84 | 0.96 | 0.98 | 0.79 | 0.64 | 0.93 | 0.81 |
| ETSlog | 0.93 | 1.00 | 0.85 | 0.79 | 0.88 | 0.62 | 0.76 | 0.91 | 0.89 | 0.92 | 0.85 | 0.70 | 0.99 | 0.73 |
| ARIMA | 0.79 | 0.85 | 1.00 | 0.94 | 0.98 | 0.80 | 0.91 | 0.94 | 0.76 | 0.80 | 1.00 | 0.86 | 0.87 | 0.62 |
| VETS | 0.74 | 0.79 | 0.94 | 1.00 | 0.92 | 0.87 | 0.98 | 0.88 | 0.71 | 0.75 | 0.95 | 0.93 | 0.82 | 0.57 |
| VETSdiag | 0.82 | 0.88 | 0.98 | 0.92 | 1.00 | 0.79 | 0.90 | 0.96 | 0.79 | 0.82 | 0.98 | 0.85 | 0.90 | 0.66 |
| VECM | 0.56 | 0.62 | 0.80 | 0.87 | 0.79 | 1.00 | 0.89 | 0.72 | 0.55 | 0.60 | 0.80 | 0.94 | 0.67 | 0.40 |
| ETSagg | 0.70 | 0.76 | 0.91 | 0.98 | 0.90 | 0.89 | 1.00 | 0.85 | 0.68 | 0.72 | 0.92 | 0.95 | 0.79 | 0.54 |
| ETSagglog | 0.84 | 0.91 | 0.94 | 0.88 | 0.96 | 0.72 | 0.85 | 1.00 | 0.81 | 0.84 | 0.94 | 0.79 | 0.92 | 0.66 |
| ARIMAagg | 0.96 | 0.89 | 0.76 | 0.71 | 0.79 | 0.55 | 0.68 | 0.81 | 1.00 | 0.98 | 0.76 | 0.62 | 0.89 | 0.86 |
| VETSagg | 0.98 | 0.92 | 0.80 | 0.75 | 0.82 | 0.60 | 0.72 | 0.84 | 0.98 | 1.00 | 0.80 | 0.67 | 0.92 | 0.85 |
| VETSaggdiag | 0.79 | 0.85 | 1.00 | 0.95 | 0.98 | 0.80 | 0.92 | 0.94 | 0.76 | 0.80 | 1.00 | 0.87 | 0.87 | 0.62 |
| ETSgroup | 0.64 | 0.70 | 0.86 | 0.93 | 0.85 | 0.94 | 0.95 | 0.79 | 0.62 | 0.67 | 0.87 | 1.00 | 0.74 | 0.48 |
| ETSgrouplog | 0.93 | 0.99 | 0.87 | 0.82 | 0.90 | 0.67 | 0.79 | 0.92 | 0.89 | 0.92 | 0.87 | 0.74 | 1.00 | 0.76 |
| ARIMAgroup | 0.81 | 0.73 | 0.62 | 0.57 | 0.66 | 0.40 | 0.54 | 0.66 | 0.86 | 0.85 | 0.62 | 0.48 | 0.76 | 1.00 |

Table 7: p-values of the paired Levene's tests.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSgroup | VETSgroupdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.91 | 0.88 | 0.32 | 0.58 | 0.03* | 0.83 | 0.75 | 0.49 | 0.52 | 0.53 | 0.53 | 0.76 | 0.93 |
| ETSlog | 0.91 | 1.00 | 0.99 | 0.38 | 0.66 | 0.06* | 0.95 | 0.86 | 0.47 | 0.52 | 0.52 | 0.66 | 0.88 | 0.87 |
| ARIMA | 0.88 | 0.99 | 1.00 | 0.38 | 0.66 | 0.05* | 0.96 | 0.86 | 0.44 | 0.47 | 0.48 | 0.65 | 0.89 | 0.85 |
| VETS | 0.32 | 0.38 | 0.38 | 1.00 | 0.66 | 0.54 | 0.37 | 0.46 | 0.18 | 0.18 | 0.19 | 0.56 | 0.42 | 0.36 |
| VETSdiag | 0.58 | 0.66 | 0.66 | 0.66 | 1.00 | 0.23 | 0.67 | 0.78 | 0.32 | 0.34 | 0.34 | 0.92 | 0.73 | 0.59 |
| VECM | 0.03* | 0.06* | 0.05* | 0.54 | 0.23 | 1.00 | 0.04* | 0.09* | 0.02* | 0.01* | 0.01* | 0.11 | 0.06* | 0.07* |
| ETSagg | 0.83 | 0.95 | 0.96 | 0.37 | 0.67 | 0.04* | 1.00 | 0.88 | 0.39 | 0.35 | 0.40 | 0.65 | 0.91 | 0.81 |
| ETSagglog | 0.75 | 0.86 | 0.86 | 0.46 | 0.78 | 0.09* | 0.88 | 1.00 | 0.39 | 0.40 | 0.41 | 0.81 | 0.96 | 0.75 |
| ARIMAagg | 0.49 | 0.47 | 0.44 | 0.18 | 0.32 | 0.02* | 0.39 | 0.39 | 1.00 | 0.73 | 0.85 | 0.26 | 0.37 | 0.63 |
| VETSgroup | 0.52 | 0.52 | 0.47 | 0.18 | 0.34 | 0.01* | 0.35 | 0.40 | 0.73 | 1.00 | 0.86 | 0.21 | 0.34 | 0.76 |
| VETSgroupdiag | 0.53 | 0.52 | 0.48 | 0.19 | 0.34 | 0.01* | 0.40 | 0.41 | 0.85 | 0.86 | 1.00 | 0.26 | 0.38 | 0.71 |
| ETSgroup | 0.53 | 0.66 | 0.65 | 0.56 | 0.92 | 0.11 | 0.65 | 0.81 | 0.26 | 0.21 | 0.26 | 1.00 | 0.74 | 0.59 |
| ETSgrouplog | 0.76 | 0.88 | 0.89 | 0.42 | 0.73 | 0.06* | 0.91 | 0.96 | 0.37 | 0.34 | 0.38 | 0.74 | 1.00 | 0.76 |
| ARIMAgroup | 0.93 | 0.87 | 0.85 | 0.36 | 0.59 | 0.07* | 0.81 | 0.75 | 0.63 | 0.76 | 0.71 | 0.59 | 0.76 | 1.00 |

Table 8: p-values of the absolute paired Brown-Forsythe tests. *: $p < 0.10$.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSgroup | VETSgroupdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.90 | 0.87 | 0.30 | 0.57 | 0.03* | 0.82 | 0.71 | 0.42 | 0.49 | 0.45 | 0.47 | 0.75 | 0.92 |
| ETSlog | 0.90 | 1.00 | 0.99 | 0.43 | 0.68 | 0.06* | 0.95 | 0.84 | 0.43 | 0.52 | 0.47 | 0.63 | 0.88 | 0.85 |
| ARIMA | 0.87 | 0.99 | 1.00 | 0.39 | 0.66 | 0.04* | 0.96 | 0.83 | 0.37 | 0.41 | 0.38 | 0.59 | 0.87 | 0.82 |
| VETS | 0.30 | 0.43 | 0.39 | 1.00 | 0.78 | 0.19 | 0.40 | 0.52 | 0.13 | 0.10* | 0.11 | 0.70 | 0.48 | 0.35 |
| VETSdiag | 0.57 | 0.68 | 0.66 | 0.78 | 1.00 | 0.17 | 0.68 | 0.79 | 0.27 | 0.30 | 0.29 | 0.97 | 0.76 | 0.57 |
| VECM | 0.03* | 0.06* | 0.04* | 0.19 | 0.17 | 1.00 | 0.04* | 0.06* | 0.01* | 0.01* | 0.01* | 0.09* | 0.05* | 0.05* |
| ETSagg | 0.82 | 0.95 | 0.96 | 0.40 | 0.68 | 0.04* | 1.00 | 0.87 | 0.33 | 0.34 | 0.33 | 0.61 | 0.91 | 0.78 |
| ETSagglog | 0.71 | 0.84 | 0.83 | 0.52 | 0.79 | 0.06* | 0.87 | 1.00 | 0.30 | 0.31 | 0.30 | 0.76 | 0.96 | 0.69 |
| ARIMAagg | 0.42 | 0.43 | 0.37 | 0.13 | 0.27 | 0.01* | 0.33 | 0.30 | 1.00 | 0.68 | 0.82 | 0.19 | 0.32 | 0.56 |
| VETSgroup | 0.49 | 0.52 | 0.41 | 0.10* | 0.30 | 0.01* | 0.34 | 0.31 | 0.68 | 1.00 | 0.82 | 0.16 | 0.33 | 0.71 |
| VETSgroupdiag | 0.45 | 0.47 | 0.38 | 0.11 | 0.29 | 0.01* | 0.33 | 0.30 | 0.82 | 0.82 | 1.00 | 0.17 | 0.32 | 0.64 |
| ETSgroup | 0.47 | 0.63 | 0.59 | 0.70 | 0.97 | 0.09* | 0.61 | 0.76 | 0.19 | 0.16 | 0.17 | 1.00 | 0.72 | 0.51 |
| ETSgrouplog | 0.75 | 0.88 | 0.87 | 0.48 | 0.76 | 0.05* | 0.91 | 0.96 | 0.32 | 0.33 | 0.32 | 0.72 | 1.00 | 0.73 |
| ARIMAgroup | 0.92 | 0.85 | 0.82 | 0.35 | 0.57 | 0.05* | 0.78 | 0.69 | 0.56 | 0.71 | 0.64 | 0.51 | 0.73 | 1.00 |

Table 9: p-values of the absolute paired Levene's tests. *: $p < 0.10$.

| | Realized standard deviation | Absolute standard deviation |
|---|---|---|
| ETS | 398.41 | 226.11 |
| ETSlog | 386.18 | 228.62 |
| ARIMA | 378.85 | 221.82 |
| VETS | 374.18 | 188.42 |
| VETSdiag | 391.91 | 214.05 |
| VECM | 327.79 | 130.41 |
| ETSagg | 363.75 | 218.05 |
| ETSagglog | 378.13 | 215.16 |
| ARIMAagg | 412.54 | 268.82 |
| VETSgroup | 424.27 | 241.52 |
| VETSgroupdiag | 381.02 | 251.44 |
| ETSgroup | 348.12 | 200.81 |
| ETSgrouplog | 403.70 | 216.76 |
| ARIMAgroup | 424.85 | 240.46 |

Table 10: Standard deviations of the realized and absolute SAE values.

# 6 Discussion

In this paper we tried to identify whether inter-series dependencies exist and we evaluated models with the objective of forecasting the declaration budgets in Dutch long-term care as precise as possible. Using historical data, we trained and tested several time series models. We identify some inter-series dependencies between some care types, where we see cointegrations arise in VECM, and trends and seasonalities in VETS. However, we do not find a statistically significant forecasting gain when incorporating multivariate models.

Considering our results as a whole, we can not point towards one method that performs best. We do find some signs of performance gain in VECM, due to a smaller variance of the absolute budget error, however there exist some practical issues in our implementation of VECM. One difficulty we found in estimating the VECM models is that the true optimal amount of lags in the model might be infeasible at the moment of writing due to the relatively small length of the time series relative to the amount of series in the vector. This could also lead to an incorrect amount of cointegration relations that are observed, which both may result in a misspecified model. Moreover, due to the relatively small sample size, the interpretation of the long-run cointegration relations may not be reliable. Furthermore, we made a strong assumption in the estimation of VECM, which was that we opted for a relative threshold depending on the realized budget. This could contaminate the results in the sense that the forecasts may be tuned by the realized values, which is not desirable in general. In practice we also find that estimating a VECM may lead to other issues, as NZa is also required to estimate the budgets for newly introduced care types, which means that the model in that case needs to be fit with relatively few data points, while requiring to estimate relatively many parameters. This could make the use of VECM infeasible in these cases, as NZa prefers a uniform method. In contrast, we find ETS and VETS to perform relatively well in terms of the budget error, which both require less parameters and in the case of VETS also has the ability to increase model complexity given a certain data length with a greater group size. This allows the VETS to fit more complex models with a smaller data length when compared to the univariate ETS. Lastly, VECM has the weakest performance in terms of the SAD, which is a measure that is valued by the NZa.

Following these results we recommend NZa to continue the use of their current methodology, as the infrastructure to forecast the budgets and to process these are already in place. In the case that another methodology needs to be added, we would recommend to use ETS with care types aggregated on the national level, as this is a relatively simple adjustment of the current methodology and will likely be an easy decision to explain towards the stakeholders, while giving similar performance. This level of aggregation still allows for a decomposition of the budgets for each care type to each insurer, using relatively simple assumptions, which is somewhat more complicated to do when the data is aggregated at the care needs assessment level. NZa could also consider the use of VETS, grouping by care needs assessment, due to the need for less data to fit relatively more complex models.

A practical issue in the models that we have used is that we do not strictly enforce non-

negative values, which is not desirable in this case due to the impossibility of negative budgets for a certain care type. We tried to remedy this by estimating logarithmic ETS models. However, when a zero value occurred in a series, this would result in the estimation of the model with a regular ETS, which then does not guarantee a non-negative value in the forecasts. This issue is somewhat remedied at higher levels of aggregation, as this decreases the likelihood of a zero value occurring in the series, but does not completely solve this issue.

Another issue regarding this topic would be the possible existence of a self-fulfilling prophecy, as the LTC-budget can be adjusted according to the results of the model, after which the insurers can optimize their quantity and mix of care types accordingly such that their costs would fit within the budget. However, we speculate that it is likely that this does not occur, as a self-fulfilling prophecy would tend to favor the results of the current methodology, i.e., ETS using data aggregated on the insurer level. However, since we have not found significantly different performing models, it can be assumed that we do not encounter a self-fulfilling prophecy in this case.

This still leaves the issue of perfectly collinear data in some of the care types. This could potentially be solved by using a dimension reduction method technique, such as principal component analysis. By definition this creates uncorrelated principal components, but may lead to a discussion in how many principal components need to be retained. This can for example be done by fixing a minimum percentage of the explained variance being retained, which is a parameter which could also be studied if this avenue of research is taken. A drawback of this approach is that this would add another step to the process, making the model more complex, especially as dimension reduction might not be trivial knowledge to the stakeholders.

Another possible point of improvement could be to follow the approach of Svetunkov et al. (2022) more closely, by splitting the current groups up, such that group care types with the same ETS model specifications are grouped together. This might decrease the probability of misspecifying the model for some series, however this also requires that the estimated ETS model would be correctly specified and dulls the point that VETS makes by requiring less data points to fit more complex model specifications. This could be used in combination with robust ETS estimations such that the possibility of a misspecification would be minimized, however this would lead to more labor for NZa, while performance gain is not guaranteed.

An oversight made in this paper is not applying a naive forecast, which would forecast the budget for each performance or group in each month, using only the last observation. However, it must be noted that we partly applied this approach indirectly, as an ETS(A,N,N) model forecasts in the same way as the naive forecast, which projects the latest value in the observed series to the whole length of the forecast. This would also open the discussion in how the naive forecast would be constructed, for example by using the average of multiple previous months. This would open new possibilities of research in examining the use of rolling windows, i.e., using a fixed amount of prior observations to estimate a time series model. This may mitigate the effects of policy changes in some care types, as the potential break in a time series will be excluded out of the training sample, when enough time passes. Additionally, this would

lead to the phasing out of the data points acquired during the SARS-Cov-2 outbreak, which may have affected the parameters and/or model specifications for some care types. This would require a similar study as this paper for determining a suitable length of the rolling window, however the amount of data points might be too small to properly execute this at the time of writing. Restricting this study to the use of VETS, which performed relatively well, could mitigate this issue somewhat, as VETS requires less data points when compared to the other models. VETS also has the advantage of being able to fit more complicated model specifications when having the same amount of data points, compared to the other models discussed in this paper, and therefore could have better performance. This also raises the question whether more statistically robust model specifications would be preferred, because a minor finding in this paper is that the model specifications of some models are not robust to the shock in budget that was induced by SARS-Cov-2. However, this point is not formally tested and therefore remains somewhat speculative. This would also raise the question whether a robustified model would show increased performance, see for example Crevits and Croux (2017) which proposes a robust alternative to ETS, but where the authors mention the caveat that robust ETS is not expected to consistently outperform ETS. For robust estimation of ARIMA and VECM one can use Chen and Liu (1993) and Zhao and Palomar (2017), respectively, as starting points. To our knowledge at the time of writing, there is no robust VETS available. Additionally, it might be useful to research whether the decrease of the budgets of in-kind care, due to SARS-Cov-2, would be paired with an increase in the budgets of PGB, as the literature suggests some substitution effects, which could lead to an increase in model performance.

Future research may extend the current model to a hierarchical model, which could incorporate variables such as demographics to achieve higher robustness to, e.g., demographic shifts. Another way to implement a hierarchical model would be to use the amount of people with care needs assessments for long term care, assess probabilities of using certain care types for these care needs assessments and then predict the number of care needs assessments as a proxy for use of the long term care budget. This would likely reduce numerical issues that certain packages find with, e.g., perfect correlation of some care types due to the demand of these care types being directly driven by the amount of people with the same care needs assessments. It could also be beneficial to include the data of PGB to examine whether substitution effects between in-kind care and PGB hold during the time when SARS-Cov-2 led to a decrease in daily in-kind care budgets. Furthermore, the study of the ETS models can be continued further if we would estimate the models using rolling averages, especially as the forecasts of the ETS(A,N,N) and ETS(M,N,N) models use the last observation to forecast over the whole forecast window. This means that these forecasts are highly dependent on this last observation, which may be more sensitive to shocks.

Another avenue of research could examine and quantify the effect of the groupings of care types that we found in this paper. This could for example be done by randomizing the groupings by fixing the sizes of the found groups and then assign the care types randomly without replacement among the groups.

Lastly, the use of different models for some subsets of care types could be researched in an attempt to find an increase in model performance. However, this would not be a practical solution, as NZa prefers to apply a single type of models to increase transparency towards the other agents affected by these forecasted budgets.

# References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Astolfi, R., Lorenzoni, L., and Oderkirk, J. (2012a). A comparative analysis of health forecasting methods. *OECD*.

Astolfi, R., Lorenzoni, L., and Oderkirk, J. (2012b). Informing policy makers about future health spending: a comparative analysis of forecasting methods in oecd countries. *Health Policy*, 107(1):1–10.

Bakx, P. L. H. and Wouterse, B. (2021). Review NZa-prognosemodel Wlz-kader.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

Bolin, K., Lindgren, B., and Lundborg, P. (2008). Informal and formal care among single-living elderly in europe. *Health economics*, 17(3):393–409.

Bonsang, E. and Schoenmaeckers, J. (2015). Long-term care insurance and the family: does the availability of potential caregivers substitute for long-term care insurance. *Ageing in Europe-Supporting Policies for an Inclusive Society. De Gruyter*, pages 369–80.

Bremer, P., Challis, D., Hallberg, I. R., Leino-Kilpi, H., Saks, K., Vellas, B., Zwakhalen, S. M., Sauerland, D., Consortium, R., et al. (2017). Informal and formal care: Substitutes or complements in care for people with dementia? empirical evidence for 8 european countries. *Health Policy*, 121(6):613–622.

Brouwers, L., Ellegård, L. M., Janlöv, N., Johansson, P., Mossler, K., and Ekholm, A. (2016). Simulating the need for health-and elderly care in sweden–a model description of sesim-lev. In *New pathways in microsimulation*, pages 57–76. Routledge.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297.

Colombo, F., Llena-Nozal, A., Mercier, J., and Tjadens, F. (2011). Help wanted. *Ageing and long-term care*, 17(2-3):3.

Commission, E., Directorate-General for Employment, S. A., and Inclusion (2021). *Long-term care report : trends, challenges and opportunities in an ageing society. Volume II, Country profiles*. Publications Office.

Crevits, R. and Croux, C. (2017). Forecasting using robust exponential smoothing with damped trend and seasonal components. *KBI_1741*.

European Commission, Council of the European Union, Directorate-General for Economic and Financial Affairs, and Economic Policy Committee (2015). *The 2015 ageing report : economic and budgetary projections for the 28 EU Member States (2013-2060)*. Publications Office.

European Commission, Directorate-General for Economic and Financial Affairs, and Przywara, B. (2010). *Projecting future health care expenditure at European level drivers, methodology and main results*. European Commission.

Fogel, R. W. (2018). Forecasting the demand for health care in OECD nations and China. In *Urbanization and Social Welfare in China*, pages 23–37. Routledge.

Forder, J. (2009). Long-term care and hospital utilisation by older people: An analysis of substitution rates. *Health economics*, 18(11):1322–1338.

Fukawa, T. (2011). Household projection and its application to health/long-term care expenditures in japan using inahsim-ii. *Social Science Computer Review*, 29(1):52–66.

Fukawa, T. and Sato, I. (2009). Projection of pension, health and long-term care expenditures in japan through macro simulation. *The Japanese Journal of Social Security Policy*, 8(1):33–42.

Getzen, T. E. (2000). Forecasting health expenditures: short, medium and long (long) term. *Journal of Health Care Finance*, 26(3):56–72.

Getzen, T. E. and Poullier, J.-P. (1992). International health spending forecasts: concepts and evaluation. *Social Science & Medicine*, 34(9):1057–1068.

Geyer, J., Haan, P., and Korfhage, T. (2017). Indirect fiscal effects of long-term care insurance. *Fiscal studies*, 38(3):393–415.

Hashiguchi, T. C. O. and Llena-Nozal, A. (2020). The effectiveness of social protection for long-term care in old age. *OECD*, (117).

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pages 1551–1580.

Johansen, S. (1994). The role of the constant and linear terms in cointegration analysis of nonstationary variables. *Econometric reviews*, 13(2):205–229.

Kattenberg, M. and Bakx, P. (2021). Substitute services: a barrier to controlling long-term care expenditures. *European Journal of Ageing*, 18(1):85–97.

Kroneman, M., Boerma, W., van den Berg, M., Groenewegen, P., de Jong, J., van Ginneken, E., Organization, W. H., et al. (2016). Netherlands: health system review.

Lagergren, M., Kurube, N., and Saito, Y. (2018). Future costs of long-term care in japan and sweden. *International Journal of Health Services*, 48(1):128–147.

Lu, B., Mi, H., Yan, G., Lim, J. K., and Feng, G. (2020). Substitutional effect of long-term care to hospital inpatient care? *China Economic Review*, 62:101466.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maarse, J. H. and Jeurissen, P. P. (2016). The policy and politics of the 2015 long-term care reform in the netherlands. *Health Policy*, 120(3):241–245.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Meuriot, V. (2015). The concept of cointégration: the decisive meeting between hendry and granger (1975). *Cahiers d'economie Politique*, 68(1):91–118.

Ministerie van Volksgezondheid, Welzijn en Sport (2019). Hervorming langdurige zorg.

Ministerie van Volksgezondheid, Welzijn en Sport (2022). Toegang tot Wlz-zorg.

Muller, M. and Morgan, D. (2020). Spending on long-term care.

Nederlandse Zorgautoriteit (2022). Februaribrief benutting Budgettair kader Wlz 2022.

Olkin, I. and Levene, H. (1960). *Robust tests for equality of variances*, page 278–292. Stanford University Press.

Pavolini, E. (2021). Long-term care social protection models in the EU.

Pavolini, E. and Ranci, C. (2008). Restructuring the welfare state: reforms in long-term care in Western European countries. *Journal of European Social Policy*, 18(3):246–259.

Scheil-Adlung, X. et al. (2015). *Long-term care protection for older persons: a review of coverage deficits in 46 countries*. ILO Geneva, Switzerland.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Snedecor, G. W. and Cochran, W. G. (1992). *Statistical methods*. Iowa State Univ. Press.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Svetunkov, I. (2022). Forecasting and analytics with ADAM. OpenForecast. (version: 29-03-2022).

Svetunkov, I., Chen, H., and Boylan, J. E. (2022). A new taxonomy for Vector Exponential Smoothing and Its Application to Seasonal Time Series. *European Journal of Operational Research*.

Van Ginneken, E., Kroneman, M., et al. (2015). Long-term care reform in the Netherlands: too large to handle? *Eurohealth*, 21(3):47–50.

Van Houtven, C. H. and Norton, E. C. (2004). Informal care and health care use of older adults. *Journal of health economics*, 23(6):1159–1180.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wackerly, D. D., Mendenhall, W., and Scheaffer, R. L. (2012). *Mathematical statistics with applications*. Brooks/Cole.

Williams, D. W. and Calabrese, T. D. (2016). The status of budget forecasting. *Journal of Public and Nonprofit Affairs*, 2(2):127–160.

Wong, A., Elderkamp-de Groot, R., Polder, J., and Van Exel, J. (2010). Predictors of long-term care utilization by Dutch hospital patients aged 65+. *BMC Health Services Research*, 10(1):1–14.

Zeilstra, A., den Ouden, A., and Vermeulen, W. (2019). Middellangetermijnverkenning zorg 2022-2025.

Zhao, Z. and Palomar, D. P. (2017). Robust maximum likelihood estimation of sparse vector error correction model. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 913–917. IEEE.

# A Distribution of ETS model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 141 | 149 | 144 | 145 | 112 | 118 |
| ETS(A,Ad,N) | 71 | 73 | 93 | 103 | 110 | 115 |
| ETS(A,N,N) | 679 | 754 | 768 | 787 | 990 | 1039 |
| ETS(M,A,N) | 180 | 189 | 166 | 169 | 130 | 120 |
| ETS(M,Ad,N) | 61 | 63 | 73 | 74 | 64 | 71 |
| ETS(M,N,N) | 554 | 514 | 538 | 528 | 400 | 381 |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 11: Count of ETS models, aggregated at the insurer level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 8% | 9% | 8% | 8% | 6% | 6% |
| ETS(A,Ad,N) | 4% | 4% | 5% | 6% | 6% | 6% |
| ETS(A,N,N) | 40% | 43% | 43% | 44% | 55% | 56% |
| ETS(M,A,N) | 11% | 11% | 9% | 9% | 7% | 7% |
| ETS(M,Ad,N) | 4% | 4% | 4% | 4% | 4% | 4% |
| ETS(M,N,N) | 33% | 30% | 30% | 29% | 22% | 21% |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 12: Percentage distribution of ETS models, aggregated at the insurer level.

# B Distribution of logETS model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 313 | 290 | 281 | 280 | 202 | 202 |
| ETS(A,Ad,N) | 183 | 178 | 219 | 226 | 204 | 201 |
| ETS(A,N,N) | 1190 | 1274 | 1282 | 1300 | 1400 | 1441 |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 13: Count of log ETS models, aggregated at the insurer level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 19% | 17% | 16% | 16% | 11% | 11% |
| ETS(A,Ad,N) | 11% | 10% | 12% | 13% | 11% | 11% |
| ETS(A,N,N) | 71% | 73% | 72% | 72% | 78% | 78% |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 14: Percentage distribution of log ETS models, aggregated at the insurer level.

# C Distribution of ETSagg model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 45 | 36 | 34 | 40 | 26 | 31 |
| ETS(A,Ad,N) | 32 | 33 | 35 | 34 | 31 | 26 |
| ETS(A,N,N) | 97 | 120 | 100 | 108 | 164 | 172 |
| ETS(M,A,N) | 48 | 57 | 64 | 53 | 28 | 31 |
| ETS(M,Ad,N) | 26 | 26 | 33 | 27 | 30 | 27 |
| ETS(M,N,N) | 108 | 93 | 102 | 107 | 90 | 97 |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 15: Count of ETS models, aggregated at the national level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 13% | 10% | 9% | 11% | 7% | 8% |
| ETS(A,Ad,N) | 9% | 9% | 10% | 9% | 8% | 7% |
| ETS(A,N,N) | 27% | 33% | 27% | 29% | 44% | 45% |
| ETS(M,A,N) | 13% | 16% | 17% | 14% | 8% | 8% |
| ETS(M,Ad,N) | 7% | 7% | 9% | 7% | 8% | 7% |
| ETS(M,N,N) | 30% | 25% | 28% | 29% | 24% | 25% |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 16: Percentage distribution of ETS models, aggregated at the national level.

# D Distribution of logETSagg model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 93 | 78 | 80 | 89 | 56 | 61 |
| ETS(A,Ad,N) | 75 | 78 | 84 | 73 | 60 | 55 |
| ETS(A,N,N) | 188 | 209 | 204 | 207 | 253 | 268 |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 17: Count of log ETS models, aggregated at the national level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 26% | 21% | 22% | 24% | 15% | 16% |
| ETS(A,Ad,N) | 21% | 21% | 23% | 20% | 16% | 14% |
| ETS(A,N,N) | 53% | 57% | 55% | 56% | 69% | 70% |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 18: Percentage distribution of log ETS models, aggregated at the national level.

# E Distribution of ETSgroup model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 12 | 10 | 11 | 10 | 10 | 9 |
| ETS(A,Ad,N) | 4 | 6 | 6 | 3 | 3 | 1 |
| ETS(A,N,N) | 5 | 11 | 8 | 8 | 16 | 18 |
| ETS(M,A,N) | 11 | 6 | 9 | 10 | 5 | 5 |
| ETS(M,Ad,N) | 5 | 6 | 6 | 5 | 6 | 7 |
| ETS(M,N,N) | 10 | 8 | 7 | 11 | 7 | 7 |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 19: Count of ETS models, aggregated at the care needs assessment level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 26% | 21% | 23% | 21% | 21% | 19% |
| ETS(A,Ad,N) | 9% | 13% | 13% | 6% | 6% | 2% |
| ETS(A,N,N) | 11% | 23% | 17% | 17% | 34% | 38% |
| ETS(M,A,N) | 23% | 13% | 19% | 21% | 11% | 11% |
| ETS(M,Ad,N) | 11% | 13% | 13% | 11% | 13% | 15% |
| ETS(M,N,N) | 21% | 17% | 15% | 23% | 15% | 15% |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 20: Percentage distribution of ETS models, aggregated at the care needs assessment level.

# F Distribution of logETSgroup model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 18 | 21 | 15 | 14 | 17 | 13 |
| ETS(A,Ad,N) | 18 | 10 | 17 | 17 | 7 | 9 |
| ETS(A,N,N) | 11 | 16 | 15 | 16 | 23 | 25 |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 21: Count of log ETS models, aggregated at the care needs assessment level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ETS(A,A,N) | 38% | 45% | 32% | 30% | 36% | 28% |
| ETS(A,Ad,N) | 38% | 21% | 36% | 36% | 15% | 19% |
| ETS(A,N,N) | 23% | 34% | 32% | 34% | 49% | 53% |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 22: Percentage distribution of log ETS models, aggregated at the care needs assessment level.

# G Distribution of VETS model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 20 | 16 | 20 | 23 | 15 | 13 |
| VETS(AAdA)PIC(LTSD,S,N) | 16 | 18 | 19 | 25 | 17 | 13 |
| VETS(AAdN)PIC(LTD,N,N) | 51 | 52 | 58 | 62 | 65 | 68 |
| VETS(AAN)PIC(LT,N,N) | 31 | 16 | 20 | 21 | 28 | 20 |
| VETS(ANA)PIC(LS,S,N) | 31 | 34 | 34 | 34 | 55 | 35 |
| VETS(ANN)PIC(L,N,N) | 71 | 92 | 78 | 80 | 97 | 134 |
| VETS(MMdM)PIC(LTSD,S,N) | 7 | 5 | 11 | 10 | 2 | 4 |
| VETS(MMdN)PIC(LTD,N,N) | 17 | 10 | 10 | 7 | 10 | 13 |
| VETS(MMM)PIC(LTS,S,N) | 9 | 14 | 7 | 6 | 6 | 4 |
| VETS(MMN)PIC(LT,N,N) | 8 | 9 | 7 | 6 | 5 | 4 |
| VETS(MNM)PIC(LS,S,N) | 20 | 25 | 27 | 34 | 5 | 10 |
| VETS(MNN)PIC(L,N,N) | 27 | 27 | 31 | 25 | 29 | 20 |
| Total | 308 | 318 | 322 | 333 | 334 | 338 |

Table 23: Count of VETS models, grouped at the insurer level by care type.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 6% | 5% | 6% | 7% | 4% | 4% |
| VETS(AAdA)PIC(LTSD,S,N) | 5% | 6% | 6% | 8% | 5% | 4% |
| VETS(AAdN)PIC(LTD,N,N) | 17% | 16% | 18% | 19% | 19% | 20% |
| VETS(AAN)PIC(LT,N,N) | 10% | 5% | 6% | 6% | 8% | 6% |
| VETS(ANA)PIC(LS,S,N) | 10% | 11% | 11% | 10% | 16% | 10% |
| VETS(ANN)PIC(L,N,N) | 23% | 29% | 24% | 24% | 29% | 40% |
| VETS(MMdM)PIC(LTSD,S,N) | 2% | 2% | 3% | 3% | 1% | 1% |
| VETS(MMdN)PIC(LTD,N,N) | 6% | 3% | 3% | 2% | 3% | 4% |
| VETS(MMM)PIC(LTS,S,N) | 3% | 4% | 2% | 2% | 2% | 1% |
| VETS(MMN)PIC(LT,N,N) | 3% | 3% | 2% | 2% | 1% | 1% |
| VETS(MNM)PIC(LS,S,N) | 6% | 8% | 8% | 10% | 1% | 3% |
| VETS(MNN)PIC(L,N,N) | 9% | 8% | 10% | 8% | 9% | 6% |
| Total | 308 | 318 | 322 | 333 | 334 | 338 |

Table 24: Percentage distribution of VETS models, grouped at the insurer level by care type.

# H Distribution of VETSdiag model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 16 | 10 | 14 | 14 | 9 | 13 |
| VETS(AAdA)PIC(LTSD,S,N) | 10 | 14 | 12 | 17 | 16 | 13 |
| VETS(AAdN)PIC(LTD,N,N) | 55 | 63 | 59 | 64 | 65 | 62 |
| VETS(AAN)PIC(LT,N,N) | 31 | 15 | 21 | 23 | 24 | 23 |
| VETS(ANA)PIC(LS,S,N) | 16 | 26 | 25 | 27 | 71 | 73 |
| VETS(ANN)PIC(L,N,N) | 74 | 83 | 79 | 82 | 90 | 97 |
| VETS(MMdM)PIC(LTSD,S,N) | 25 | 15 | 23 | 30 | 8 | 8 |
| VETS(MMdN)PIC(LTD,N,N) | 8 | 8 | 12 | 6 | 5 | 7 |
| VETS(MMM)PIC(LTS,S,N) | 28 | 22 | 17 | 15 | 5 | 4 |
| VETS(MMN)PIC(LT,N,N) | 7 | 8 | 6 | 7 | 6 | 5 |
| VETS(MNM)PIC(LS,S,N) | 7 | 25 | 25 | 23 | 10 | 10 |
| VETS(MNN)PIC(L,N,N) | 31 | 29 | 29 | 25 | 25 | 23 |
| Total | 308 | 318 | 322 | 333 | 334 | 338 |

Table 25: Count of VETS models, with diagonal $\Sigma$, grouped at the insurer level by care type.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 5% | 3% | 4% | 4% | 3% | 4% |
| VETS(AAdA)PIC(LTSD,S,N) | 3% | 4% | 4% | 5% | 5% | 4% |
| VETS(AAdN)PIC(LTD,N,N) | 18% | 20% | 18% | 19% | 19% | 18% |
| VETS(AAN)PIC(LT,N,N) | 10% | 5% | 7% | 7% | 7% | 7% |
| VETS(ANA)PIC(LS,S,N) | 5% | 8% | 8% | 8% | 21% | 22% |
| VETS(ANN)PIC(L,N,N) | 24% | 26% | 25% | 25% | 27% | 29% |
| VETS(MMdM)PIC(LTSD,S,N) | 8% | 5% | 7% | 9% | 2% | 2% |
| VETS(MMdN)PIC(LTD,N,N) | 3% | 3% | 4% | 2% | 1% | 2% |
| VETS(MMM)PIC(LTS,S,N) | 9% | 7% | 5% | 5% | 1% | 1% |
| VETS(MMN)PIC(LT,N,N) | 2% | 3% | 2% | 2% | 2% | 1% |
| VETS(MNM)PIC(LS,S,N) | 2% | 8% | 8% | 7% | 3% | 3% |
| VETS(MNN)PIC(L,N,N) | 10% | 9% | 9% | 8% | 7% | 7% |
| Total | 308 | 318 | 322 | 333 | 334 | 338 |

Table 26: Percentage distribution of VETS models, with diagonal $\Sigma$, grouped at the insurer level by care type.

# I Distribution of VETSagg model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 5 | 5 | 5 | 6 | 4 | 1 |
| VETS(AAdA)PIC(LTSD,S,N) | 1 | 6 | 3 | 3 | 2 | 3 |
| VETS(AAdN)PIC(LTD,N,N) | 7 | 6 | 6 | 4 | 5 | 5 |
| VETS(AAN)PIC(LT,N,N) | 4 | 1 | 1 | 5 | 6 | 6 |
| VETS(ANA)PIC(LS,S,N) | 3 | 0 | 2 | 2 | 5 | 3 |
| VETS(ANN)PIC(L,N,N) | 5 | 6 | 3 | 4 | 3 | 3 |
| VETS(MMdM)PIC(LTSD,S,N) | 1 | 4 | 2 | 4 | 1 | 6 |
| VETS(MMdN)PIC(LTD,N,N) | 1 | 2 | 2 | 2 | 1 | 2 |
| VETS(MMM)PIC(LTS,S,N) | 5 | 1 | 4 | 1 | 3 | 0 |
| VETS(MMN)PIC(LT,N,N) | 3 | 3 | 5 | 4 | 5 | 7 |
| VETS(MNM)PIC(LS,S,N) | 0 | 0 | 0 | 1 | 0 | 0 |
| VETS(MNN)PIC(L,N,N) | 3 | 2 | 2 | 2 | 3 | 2 |
| Total | 38 | 36 | 35 | 38 | 38 | 38 |

Table 27: Count of VETS models, grouped at the national level by care needs assessment.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 13% | 14% | 14% | 16% | 11% | 3% |
| VETS(AAdA)PIC(LTSD,S,N) | 3% | 17% | 9% | 8% | 5% | 8% |
| VETS(AAdN)PIC(LTD,N,N) | 18% | 17% | 17% | 11% | 13% | 13% |
| VETS(AAN)PIC(LT,N,N) | 11% | 3% | 3% | 13% | 16% | 16% |
| VETS(ANA)PIC(LS,S,N) | 8% | 0% | 6% | 5% | 13% | 8% |
| VETS(ANN)PIC(L,N,N) | 13% | 17% | 9% | 11% | 8% | 8% |
| VETS(MMdM)PIC(LTSD,S,N) | 3% | 11% | 6% | 11% | 3% | 16% |
| VETS(MMdN)PIC(LTD,N,N) | 3% | 6% | 6% | 5% | 3% | 5% |
| VETS(MMM)PIC(LTS,S,N) | 13% | 3% | 11% | 3% | 8% | 0% |
| VETS(MMN)PIC(LT,N,N) | 8% | 8% | 14% | 11% | 13% | 18% |
| VETS(MNM)PIC(LS,S,N) | 0% | 0% | 0% | 3% | 0% | 0% |
| VETS(MNN)PIC(L,N,N) | 8% | 6% | 6% | 5% | 8% | 5% |
| Total | 38 | 36 | 35 | 38 | 38 | 38 |

Table 28: Percentage distribution of VETS models, grouped at the national level by care needs assessment.

# J Distribution of VETSaggdiag model specifications

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 3 | 2 | 0 | 5 | 2 | 2 |
| VETS(AAdA)PIC(LTSD,S,N) | 5 | 6 | 3 | 4 | 1 | 4 |
| VETS(AAdN)PIC(LTD,N,N) | 8 | 8 | 10 | 9 | 6 | 4 |
| VETS(AAN)PIC(LT,N,N) | 10 | 6 | 5 | 6 | 4 | 4 |
| VETS(ANA)PIC(LS,S,N) | 1 | 3 | 2 | 4 | 4 | 2 |
| VETS(ANN)PIC(L,N,N) | 5 | 6 | 6 | 4 | 14 | 12 |
| VETS(MMdM)PIC(LTSD,S,N) | 0 | 1 | 0 | 1 | 0 | 0 |
| VETS(MMdN)PIC(LTD,N,N) | 2 | 3 | 3 | 1 | 1 | 2 |
| VETS(MMM)PIC(LTS,S,N) | 0 | 0 | 1 | 1 | 0 | 1 |
| VETS(MMN)PIC(LT,N,N) | 4 | 3 | 3 | 5 | 4 | 5 |
| VETS(MNM)PIC(LS,S,N) | 1 | 1 | 1 | 1 | 1 | 1 |
| VETS(MNN)PIC(L,N,N) | 4 | 2 | 3 | 2 | 4 | 4 |
| Total | 43 | 41 | 37 | 43 | 41 | 41 |

Table 29: Count of VETS models, with diagonal $\Sigma$, grouped at the national level by care needs assessment.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| VETS(AAA)PIC(LTS,S,N) | 7% | 5% | 0% | 12% | 5% | 5% |
| VETS(AAdA)PIC(LTSD,S,N) | 12% | 15% | 8% | 9% | 2% | 10% |
| VETS(AAdN)PIC(LTD,N,N) | 19% | 20% | 27% | 21% | 15% | 10% |
| VETS(AAN)PIC(LT,N,N) | 23% | 15% | 14% | 14% | 10% | 10% |
| VETS(ANA)PIC(LS,S,N) | 2% | 7% | 5% | 9% | 10% | 5% |
| VETS(ANN)PIC(L,N,N) | 12% | 15% | 16% | 9% | 34% | 29% |
| VETS(MMdM)PIC(LTSD,S,N) | 0% | 2% | 0% | 2% | 0% | 0% |
| VETS(MMdN)PIC(LTD,N,N) | 5% | 7% | 8% | 2% | 2% | 5% |
| VETS(MMM)PIC(LTS,S,N) | 0% | 0% | 3% | 2% | 0% | 2% |
| VETS(MMN)PIC(LT,N,N) | 9% | 7% | 8% | 12% | 10% | 12% |
| VETS(MNM)PIC(LS,S,N) | 2% | 2% | 3% | 2% | 2% | 2% |
| VETS(MNN)PIC(L,N,N) | 9% | 5% | 8% | 5% | 10% | 10% |
| Total | 43 | 41 | 37 | 43 | 41 | 41 |

Table 30: Percentage distribution of VETS models, with diagonal $\Sigma$, grouped at the national level by care needs assessment.

# K Distribution of ARIMA model specifications

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 106 | 106 | 83 | 78 | 35 | 41 |
| 1 | 848 | 883 | 852 | 844 | 947 | 948 |
| 2 | 420 | 401 | 469 | 490 | 386 | 391 |
| 3 | 187 | 181 | 210 | 202 | 245 | 249 |
| 4 | 68 | 111 | 104 | 115 | 115 | 118 |
| 5 | 48 | 54 | 54 | 69 | 65 | 77 |
| 6 | 8 | 2 | 9 | 7 | 12 | 18 |
| 7 | 1 | 3 | 1 | 1 | 1 | 2 |
| 8 | 0 | 1 | 0 | 0 | 0 | 0 |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 31: Count of amount of parameters in ARIMA models, aggregated at the insurer level.

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 6% | 6% | 5% | 4% | 2% | 2% |
| 1 | 50% | 51% | 48% | 47% | 52% | 51% |
| 2 | 25% | 23% | 26% | 27% | 21% | 21% |
| 3 | 11% | 10% | 12% | 11% | 14% | 14% |
| 4 | 4% | 6% | 6% | 6% | 6% | 6% |
| 5 | 3% | 3% | 3% | 4% | 4% | 4% |
| 6 | 0% | 0% | 1% | 0% | 1% | 1% |
| 7 | 0% | 0% | 0% | 0% | 0% | 0% |
| 8 | 0% | 0% | 0% | 0% | 0% | 0% |
| Total | 1686 | 1742 | 1782 | 1806 | 1844 | 1844 |

Table 32: Percentage distribution of amount of parameters in ARIMA models, aggregated at the insurer level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ARIMA(0,0,0) | 106 | 106 | 83 | 78 | 35 | 41 |
| ARIMA(0,1,0) | 720 | 743 | 751 | 751 | 804 | 812 |
| ARIMA(0,2,0) | 3 | 3 | 3 | 3 | 3 | 2 |
| ARIMA(0,0,1) | 25 | 37 | 21 | 15 | 22 | 26 |
| ARIMA(0,1,1) | 277 | 267 | 325 | 345 | 224 | 218 |
| ARIMA(0,2,1) | 110 | 118 | 120 | 112 | 93 | 89 |
| ARIMA(0,0,2) | 9 | 11 | 6 | 4 | 12 | 11 |
| ARIMA(0,1,2) | 39 | 20 | 29 | 27 | 64 | 60 |
| ARIMA(0,2,2) | 12 | 16 | 22 | 20 | 35 | 38 |
| ARIMA(0,0,3) | 3 | 1 | 0 | 0 | 2 | 2 |
| ARIMA(0,1,3) | 6 | 24 | 18 | 17 | 16 | 23 |
| ARIMA(0,2,3) | 5 | 2 | 6 | 8 | 7 | 8 |
| ARIMA(0,1,4) | 0 | 0 | 2 | 78 | 2 | 2 |
| ARIMA(0,2,4) | 0 | 0 | 0 | 0 | 0 | 1 |
| ARIMA(1,0,0) | 103 | 103 | 80 | 0 | 121 | 110 |
| ARIMA(1,1,0) | 77 | 74 | 88 | 83 | 77 | 87 |
| ARIMA(1,2,0) | 6 | 6 | 1 | 1 | 1 | 0 |
| ARIMA(1,0,1) | 26 | 16 | 27 | 28 | 37 | 30 |
| ARIMA(1,1,1) | 8 | 10 | 22 | 21 | 52 | 76 |
| ARIMA(1,2,1) | 9 | 17 | 18 | 18 | 19 | 15 |
| ARIMA(1,0,2) | 0 | 0 | 6 | 4 | 1 | 2 |
| ARIMA(1,1,2) | 11 | 4 | 7 | 10 | 6 | 7 |
| ARIMA(1,2,2) | 1 | 0 | 3 | 3 | 4 | 5 |
| ARIMA(1,0,3) | 1 | 1 | 2 | 2 | 0 | 0 |
| ARIMA(1,1,3) | 1 | 1 | 1 | 1 | 3 | 6 |
| ARIMA(1,0,4) | 0 | 1 | 0 | 0 | 0 | 0 |
| ARIMA(1,2,4) | 0 | 1 | 0 | 0 | 0 | 1 |
| ARIMA(2,0,0) | 28 | 30 | 20 | 27 | 33 | 43 |
| ARIMA(2,1,0) | 11 | 15 | 22 | 27 | 20 | 11 |
| ARIMA(2,2,0) | 0 | 3 | 2 | 2 | 0 | 1 |
| ARIMA(2,0,1) | 7 | 5 | 8 | 9 | 10 | 8 |
| ARIMA(2,1,1) | 10 | 23 | 17 | 27 | 23 | 22 |
| ARIMA(2,2,1) | 8 | 13 | 7 | 10 | 7 | 8 |
| ARIMA(2,0,2) | 1 | 5 | 5 | 8 | 1 | 3 |
| ARIMA(2,1,2) | 16 | 20 | 24 | 37 | 35 | 44 |
| ARIMA(2,2,2) | 1 | 1 | 3 | 1 | 6 | 7 |
| ARIMA(2,0,3) | 0 | 0 | 0 | 1 | 0 | 1 |
| ARIMA(2,1,3) | 0 | 0 | 1 | 1 | 2 | 2 |
| ARIMA(2,2,3) | 1 | 1 | 1 | 0 | 1 | 1 |
| ARIMA(2,0,4) | 0 | 0 | 0 | 0 | 1 | 1 |
| ARIMA(2,2,4) | 0 | 1 | 0 | 0 | 0 | 0 |
| ARIMA(2,0,5) | 0 | 0 | 0 | 0 | 0 | 0 |
| ARIMA(3,0,0) | 3 | 6 | 2 | 1 | 2 | 1 |
| ARIMA(3,1,0) | 15 | 15 | 11 | 9 | 14 | 8 |
| ARIMA(3,2,0) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(3,0,1) | 1 | 1 | 0 | 1 | 1 | 1 |
| ARIMA(3,1,1) | 12 | 11 | 5 | 7 | 2 | 2 |
| ARIMA(3,2,1) | 1 | 1 | 3 | 2 | 2 | 3 |
| ARIMA(3,0,2) | 0 | 0 | 1 | 0 | 0 | 0 |
| ARIMA(3,1,2) | 5 | 0 | 2 | 2 | 1 | 3 |
| ARIMA(3,2,2) | 0 | 0 | 0 | 1 | 0 | 0 |
| ARIMA(4,0,0) | 2 | 2 | 2 | 1 | 0 | 0 |
| ARIMA(4,1,0) | 4 | 6 | 5 | 2 | 3 | 0 |
| ARIMA(4,2,0) | 0 | 0 | 0 | 1 | 0 | 0 |
| ARIMA(4,1,1) | 1 | 0 | 0 | 0 | 0 | 0 |
| ARIMA(4,2,1) | 0 | 1 | 0 | 0 | 0 | 0 |
| ARIMA(5,0,0) | 1 | 0 | 0 | 0 | 1 | 1 |
| ARIMA(5,1,0) | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 1686 | 1742 | 1782 | 1806 | 1806 | 1844 |

Table 33: Count of ARIMA models, aggregated at the insurer level.

# L Distribution of ARIMAagg model specifications

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 23 | 24 | 23 | 24 | 12 | 19 |
| 1 | 164 | 157 | 146 | 145 | 165 | 180 |
| 2 | 83 | 94 | 95 | 96 | 105 | 96 |
| 3 | 37 | 37 | 63 | 55 | 51 | 44 |
| 4 | 31 | 27 | 20 | 26 | 20 | 24 |
| 5 | 17 | 23 | 17 | 19 | 9 | 14 |
| 6 | 1 | 3 | 3 | 3 | 5 | 6 |
| 7 | 0 | 0 | 0 | 1 | 2 | 1 |
| 8 | 0 | 0 | 1 | 0 | 0 | 0 |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 34: Count of amount of parameters in ARIMA models, aggregated at the national level.

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 6% | 7% | 6% | 7% | 3% | 5% |
| 1 | 46% | 43% | 40% | 39% | 45% | 47% |
| 2 | 23% | 26% | 26% | 26% | 28% | 25% |
| 3 | 10% | 10% | 17% | 15% | 14% | 11% |
| 4 | 9% | 7% | 5% | 7% | 5% | 6% |
| 5 | 5% | 6% | 5% | 5% | 2% | 4% |
| 6 | 0% | 1% | 1% | 1% | 1% | 2% |
| 7 | 0% | 0% | 0% | 0% | 1% | 0% |
| 8 | 0% | 0% | 0% | 0% | 0% | 0% |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 35: Percentage distribution of amount of parameters in ARIMA models, aggregated at the national level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ARIMA(0,0,0) | 23 | 24 | 23 | 24 | 12 | 19 |
| ARIMA(0,1,0) | 142 | 141 | 129 | 128 | 139 | 149 |
| ARIMA(0,2,0) | 4 | 3 | 4 | 2 | 1 | 1 |
| ARIMA(0,0,1) | 1 | 4 | 6 | 5 | 6 | 9 |
| ARIMA(0,1,1) | 52 | 52 | 64 | 65 | 45 | 37 |
| ARIMA(0,2,1) | 25 | 26 | 37 | 36 | 24 | 20 |
| ARIMA(0,0,2) | 4 | 9 | 4 | 2 | 9 | 10 |
| ARIMA(0,1,2) | 2 | 3 | 10 | 3 | 6 | 3 |
| ARIMA(0,2,2) | 5 | 8 | 5 | 3 | 9 | 7 |
| ARIMA(0,0,3) | 0 | 1 | 0 | 1 | 0 | 0 |
| ARIMA(0,1,3) | 5 | 2 | 1 | 5 | 2 | 4 |
| ARIMA(0,2,3) | 0 | 4 | 1 | 1 | 0 | 1 |
| ARIMA(0,1,4) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(0,1,5) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(1,0,0) | 21 | 12 | 11 | 12 | 20 | 22 |
| ARIMA(1,1,0) | 13 | 15 | 16 | 20 | 21 | 19 |
| ARIMA(1,2,0) | 1 | 0 | 1 | 2 | 0 | 1 |
| ARIMA(1,0,1) | 6 | 6 | 3 | 3 | 15 | 9 |
| ARIMA(1,1,1) | 1 | 1 | 7 | 5 | 13 | 14 |
| ARIMA(1,2,1) | 5 | 6 | 5 | 8 | 5 | 5 |
| ARIMA(1,0,2) | 1 | 0 | 0 | 0 | 1 | 2 |
| ARIMA(1,1,2) | 4 | 4 | 5 | 3 | 1 | 2 |
| ARIMA(1,2,2) | 0 | 2 | 1 | 1 | 1 | 2 |
| ARIMA(1,0,3) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(1,1,3) | 0 | 0 | 0 | 1 | 1 | 1 |
| ARIMA(1,2,3) | 0 | 1 | 0 | 0 | 1 | 1 |
| ARIMA(1,1,4) | 0 | 1 | 1 | 0 | 0 | 0 |
| ARIMA(2,0,0) | 4 | 9 | 4 | 4 | 14 | 20 |
| ARIMA(2,1,0) | 4 | 5 | 7 | 5 | 3 | 3 |
| ARIMA(2,2,0) | 0 | 1 | 0 | 0 | 0 | 0 |
| ARIMA(2,0,1) | 1 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(2,1,1) | 10 | 3 | 3 | 3 | 0 | 4 |
| ARIMA(2,2,1) | 0 | 2 | 3 | 2 | 2 | 3 |
| ARIMA(2,0,2) | 0 | 0 | 0 | 1 | 0 | 0 |
| ARIMA(2,1,2) | 5 | 10 | 7 | 11 | 1 | 7 |
| ARIMA(2,2,2) | 1 | 1 | 1 | 2 | 2 | 4 |
| ARIMA(3,0,0) | 2 | 1 | 1 | 3 | 3 | 1 |
| ARIMA(3,1,0) | 2 | 2 | 0 | 2 | 2 | 2 |
| ARIMA(3,0,1) | 0 | 0 | 0 | 1 | 0 | 0 |
| ARIMA(3,1,1) | 6 | 4 | 3 | 2 | 1 | 0 |
| ARIMA(3,2,1) | 0 | 0 | 0 | 1 | 1 | 1 |
| ARIMA(3,0,2) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(3,1,2) | 0 | 0 | 1 | 0 | 0 | 0 |
| ARIMA(3,2,2) | 0 | 0 | 0 | 1 | 1 | 1 |
| ARIMA(3,1,3) | 0 | 0 | 0 | 0 | 0 | 0 |
| ARIMA(3,1,4) | 0 | 0 | 1 | 0 | 0 | 0 |
| ARIMA(4,0,0) | 0 | 1 | 1 | 0 | 0 | 0 |
| ARIMA(4,1,0) | 6 | 1 | 1 | 1 | 1 | 0 |
| ARIMA(4,0,1) | 0 | 0 | 1 | 0 | 0 | 0 |
| ARIMA(4,1,2) | 0 | 0 | 0 | 0 | 1 | 0 |
| Total | 356 | 365 | 368 | 369 | 369 | 384 |

Table 36: Count of ARIMA models, aggregated at the national level.

# M Distribution of ARIMAgroup model specifications

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 2 | 4 | 3 | 6 | 0 | 1 |
| 1 | 27 | 25 | 19 | 16 | 21 | 23 |
| 2 | 9 | 10 | 8 | 10 | 15 | 11 |
| 3 | 7 | 8 | 15 | 14 | 8 | 8 |
| 4 | 2 | 0 | 2 | 1 | 2 | 3 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 37: Count of amount of parameters in ARIMA models, aggregated at the care needs assessment level.

| Number of parameters | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| 0 | 4% | 9% | 6% | 13% | 0% | 2% |
| 1 | 57% | 53% | 40% | 34% | 45% | 49% |
| 2 | 19% | 21% | 17% | 21% | 32% | 23% |
| 3 | 15% | 17% | 32% | 30% | 17% | 17% |
| 4 | 4% | 0% | 4% | 2% | 4% | 6% |
| 5 | 0% | 0% | 0% | 0% | 0% | 0% |
| 6 | 0% | 0% | 0% | 0% | 2% | 0% |
| 7 | 0% | 0% | 0% | 0% | 0% | 2% |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 38: Percentage distribution of amount of parameters in ARIMA models, aggregated at the care needs assessment level.

| | February 2019 | July 2019 | February 2020 | July 2020 | February 2021 | July 2021 |
|---|---|---|---|---|---|---|
| ARIMA(0,0,0) | 1 | 4 | 3 | 6 | 0 | 1 |
| ARIMA(0,1,0) | 25 | 23 | 18 | 16 | 17 | 18 |
| ARIMA(0,2,0) | 1 | 2 | 3 | 3 | 1 | 1 |
| ARIMA(0,0,1) | 0 | 1 | 1 | 0 | 3 | 3 |
| ARIMA(0,1,1) | 3 | 3 | 2 | 6 | 11 | 4 |
| ARIMA(0,2,1) | 5 | 6 | 11 | 9 | 6 | 4 |
| ARIMA(0,1,2) | 1 | 1 | 2 | 1 | 0 | 1 |
| ARIMA(0,2,2) | 1 | 0 | 0 | 1 | 1 | 1 |
| ARIMA(0,1,3) | 0 | 0 | 0 | 0 | 0 | 1 |
| ARIMA(1,0,0) | 2 | 1 | 0 | 0 | 1 | 2 |
| ARIMA(1,1,0) | 4 | 3 | 2 | 0 | 2 | 2 |
| ARIMA(1,1,1) | 1 | 0 | 1 | 1 | 2 | 2 |
| ARIMA(1,1,2) | 0 | 0 | 1 | 0 | 1 | 1 |
| ARIMA(2,0,0) | 1 | 2 | 1 | 1 | 1 | 4 |
| ARIMA(2,1,0) | 1 | 1 | 1 | 1 | 0 | 1 |
| ARIMA(2,0,1) | 0 | 0 | 1 | 1 | 0 | 0 |
| ARIMA(2,1,1) | 1 | 0 | 0 | 0 | 0 | 0 |
| ARIMA(2,2,2) | 0 | 0 | 0 | 0 | 1 | 0 |
| ARIMA(3,0,0) | 0 | 0 | 0 | 1 | 0 | 0 |
| ARIMA(3,2,2) | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 47 | 47 | 47 | 47 | 47 | 47 |

Table 39: Count of ARIMA models, aggregated at the care needs assessment level.

# N VECM example

| | Test statistic | 90% | 95% | 99% |
|---|---|---|---|---|
| r $\leq 6$ | 7.03 | 6.50 | 8.18 | 11.65 |
| r $\leq 5$ | 9.49 | 12.91 | 14.90 | 19.19 |
| r $\leq 4$ | 17.02 | 18.90 | 21.07 | 25.75 |
| r $\leq 3$ | 28.82 | 24.78 | 27.14 | 32.14 |
| r $\leq 2$ | 34.19 | 30.84 | 33.32 | 38.78 |
| r $\leq 1$ | 65.47 | 36.25 | 39.43 | 44.59 |
| r $= 0$ | 73.93 | 42.06 | 44.91 | 51.3 |

Table 40: Johansen test statistics against the $90^{\text{th}}$, $95^{\text{th}}$, and $99^{\text{th}}$ percentiles.

$$
\begin{pmatrix} \Delta y_t^1 \\ \Delta y_t^2 \\ \Delta y_t^3 \\ \Delta y_t^4 \\ \Delta y_t^5 \\ \Delta y_t^6 \\ \Delta y_t^7 \end{pmatrix} = \begin{pmatrix} -54022.2 \\ 81135.6 \\ 89045.9 \\ 36903.7 \\ 316145.5 \\ -214469.3 \\ -97227.4 \end{pmatrix} + \begin{pmatrix} -0.898 & 2.240 & -0.148 \\ 0.043 & -0.497 & 0.051 \\ -0.238 & 0.887 & -0.038 \\ -0.098 & 0.412 & -0.019 \\ -0.092 & 2.058 & -0.076 \\ -1.397 & 5.643 & -0.410 \\ -0.210 & 0.635 & -0.060 \end{pmatrix} \begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \\ -2.134 & -1.383 & -12.740 \\ -1.510 & -0.638 & -6.822 \\ 0.353 & 0.068 & -0.569 \\ -1.861 & 0.459 & 15.740 \end{pmatrix}' \mathbf{y}_{t-1}
$$

$$
+ \begin{pmatrix} -0.442 & 1.308 & 0.243 & 0.157 & 0.576 & 0.063 & 0.837 \\ 0.068 & -0.212 & 0.047 & -0.050 & 0.042 & 0.026 & -0.009 \\ -0.007 & 0.493 & 0.499 & -1.536 & 0.048 & -0.001 & 0.829 \\ 0.005 & 0.183 & -0.032 & -0.253 & -0.137 & 0.018 & -0.060 \\ -0.227 & 0.672 & 0.376 & -1.724 & -0.303 & 0.098 & 1.787 \\ -1.000 & 3.586 & -0.087 & -0.784 & -0.095 & 0.443 & 2.408 \\ -0.093 & 0.324 & 0.036 & 0.045 & 0.020 & 0.013 & 0.434 \end{pmatrix} \begin{pmatrix} \Delta y_{t-1}^1 \\ \Delta y_{t-1}^2 \\ \Delta y_{t-1}^3 \\ \Delta y_{t-1}^4 \\ \Delta y_{t-1}^5 \\ \Delta y_{t-1}^6 \\ \Delta y_{t-1}^7 \end{pmatrix}
$$

$$
+ \begin{pmatrix} -0.557 & 0.267 & 0.081 & 0.405 & 0.582 & -0.141 & -0.186 \\ 0.049 & -0.564 & -0.019 & 0.148 & -0.064 & 0.029 & -0.193 \\ 0.043 & -0.628 & 0.126 & -0.199 & -0.289 & -0.026 & -0.636 \\ 0.017 & 0.060 & 0.064 & -0.432 & -0.037 & -0.027 & 0.327 \\ -0.283 & 0.474 & 0.131 & -1.066 & -0.457 & 0.066 & 0.774 \\ -1.403 & 4.194 & 0.201 & 1.776 & 0.654 & -0.268 & 1.284 \\ -0.181 & 0.255 & 0.038 & 0.136 & 0.130 & -0.012 & -0.293 \end{pmatrix} \begin{pmatrix} \Delta y_{t-2}^1 \\ \Delta y_{t-2}^2 \\ \Delta y_{t-2}^3 \\ \Delta y_{t-2}^4 \\ \Delta y_{t-2}^5 \\ \Delta y_{t-2}^6 \\ \Delta y_{t-2}^7 \end{pmatrix} \tag{45}
$$

$$
+ \begin{pmatrix} -0.541 & 2.148 & -0.491 & 0.826 & 0.628 & -0.064 & 1.134 \\ 0.093 & -0.483 & -0.062 & -0.061 & -0.096 & -0.036 & 0.486 \\ -0.217 & 0.229 & -0.191 & -0.348 & 0.086 & 0.014 & 1.497 \\ 0.002 & 0.168 & -0.029 & -0.119 & -0.036 & -0.009 & 0.035 \\ -0.377 & 0.951 & 0.340 & -2.787 & -0.444 & 0.277 & 1.223 \\ -1.326 & 1.625 & 0.906 & -3.151 & 1.002 & -0.134 & 2.670 \\ -0.145 & 0.485 & 0.002 & 0.204 & 0.149 & -0.010 & 0.306 \end{pmatrix} \begin{pmatrix} \Delta y_{t-3}^1 \\ \Delta y_{t-3}^2 \\ \Delta y_{t-3}^3 \\ \Delta y_{t-3}^4 \\ \Delta y_{t-3}^5 \\ \Delta y_{t-3}^6 \\ \Delta y_{t-3}^7 \end{pmatrix}
$$

# O Groups of care types

We show the different groups of variables that we estimate together. Firstly, we group each care type at the level of the insurer, this is rather trivial as each group only contains the amount of declarations by that care type, delivered by the different insurers. Secondly, we perform grouping at the care needs assessment level, containing the care types where patients with that care needs assessment is the largest source of costs. These groupings are shown in Table 41.[2]

[2]More information about these care needs assessments can be found in `https://wetten.overheid.nl/BWBR0036014/2022-04-15/0#BijlageA` (in Dutch). More information about the care types can be found in `https://puc.overheid.nl/nza/doc/PUC_646976_22/1/`, for codes starting with F, H, and M, and `https://puc.overheid.nl/nza/doc/PUC_658250_22/1/`, for codes starting with V and Z (both in Dutch).

| Care needs Assessment | Care type |
|---|---|
| 5VV | H104, H106, H117, H120, H126, H128, H321, H335, H531, H533, H802, V051, V053, V101, Z051, Z053, Z1003, Z103, Z110, V103 |
| 4VV | H127, V041, V043, Z041, Z043 |
| 4LG | H152, H832, H913, V640, V641, V642, Z1000, Z640, Z641, Z642, Z643, V643 |
| 3VG | F125, H150, H153, H300, H811, H900, V430, V431, V433, V941, V942, V980, Z430, Z431, Z432, Z433, V432 |
| 2ZGaud | H304, H921, V720, Z720, Z721, Z722, Z723, H303, H852, V721, V723 |
| 6VG | H325, H329, H334, H336, H812, H815, H891, H904, H941, V460, V461, V462, V463, V943, V945, V979, Z460, Z461, Z462, Z463, Z914, Z942, Z943, Z945, Z978, Z979, Z980, Z981, Z999, Z912, V914, V978, V981 |
| 8VG | H330, H332, H813, H816, H817, H819, H884, H885, H906, V481, V483, Z480, Z481, Z482, Z483, Z919, Z977, H942, V480, V482 |
| 6VV | H800, V061, V063, Z061, Z063, Z101 |
| 6LG | H833, H835, H914, V660, V661, Z660, Z661, Z662, Z663, V663, V662 |
| 4VG | H814, H881, V440, V441, V442, V443, Z440, Z441, Z442, Z443, Z913, V913 |
| 5VG | H820, H821, H882, H883, H903, V454, V455, V457, Z454, Z455, Z456, Z457, Z976, H818, V456 |
| 7VG | H822, H902, V472, V473, V944, Z470, Z471, Z472, Z473, Z911, Z915, Z941, Z944, H943, V471, Z983, V470 |
| 7VV | V071, Z071, Z073, V073 |
| 8VV | V081, Z081, Z083, Z920, V083, V920 |
| 9VV B | Z095, Z097, Z910, V097, V095 |
| 5LG | H916, V650, Z650, Z651, Z652, Z653, V651, V652, V653 |
| 2VV | V025, Z025 |
| 3VV | V031, V033, Z031, Z033 |
| 4GGZ B | Z242, Z243 |
| 7GGZ B | Z272, Z273, Z280, Z902, Z922 |
| 2LG | V624, V625, V977, Z624, Z625 |
| 7LG | H831, H910, V671, Z670, Z671, Z672, Z673, H836, H950, V672, V670, V673 |
| 3ZGvis | H871, Z830, Z831, Z833, H301, H930, V831, Z832, V832, V833, V830 |
| 4ZGvis | Z840, Z841, Z842, Z843, H302, V843 |
| ZZP0 | Z995, Z997, Z998 |
| 1VV | Z015, V015 |
| 5ZGvis | H873, Z850, Z851, Z852, Z853, H934, V853 |
| 6GGZ B | Z262, Z263 |
| 1LG | Z614, Z615, V614, V615 |
| 3LG | H915, V630, V631, Z630, Z631, Z632, Z633, V632, V633 |
| 5GGZ B | Z252, Z253, Z982 |
| 1LVG | Z513 |
| 2LVG | Z523 |
| 3LVG | Z533, Z560, V533 |
| 4LVG | Z543, V543 |
| 5LVG | Z553 |
| 1SGLVG | Z573 |
| 2VG | V424, Z424, Z425, V425 |
| 3ZGaud | H922, V730, V731, Z1002, Z730, Z731, Z732, Z733, H333, H337, H854, V733, H853, H856 |
| 1VG | Z414, Z415, V414, V415 |
| 3GGZ B | Z232, Z233 |
| 4ZGaud | H851, H920, V740, V741, Z740, Z741, Z743, Z742, V742 |
| 1ZGvis | Z814, Z815 |
| 2ZGvis | Z824, Z825, V824, V825 |
| 1ZGaud | Z710, Z711, V710, Z713 |
| Remaining care types | H138, H139, H306, H834, H840, H886, H887, H963, H964, H965, H966, H967, H968, H969, V940, Z492, Z493, Z494, Z918B, Z921B, Z923B, Z940, Z946, ZMZTO, H338, V921B, V923B, V9011, V841, V9010, V918B, V946, Z9010, Z9011 |

Table 41: Grouping of care types by care needs assessment.

# P F-test and Bartlett's tests

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.95 | 0.91 | 0.89 | 0.97 | 0.68 | 0.85 | 0.91 | 0.94 | 0.89 | 0.92 | 0.77 | 0.98 | 0.89 |
| ETSlog | 0.95 | 1.00 | 0.97 | 0.95 | 0.97 | 0.73 | 0.90 | 0.96 | 0.89 | 0.84 | 0.98 | 0.83 | 0.92 | 0.84 |
| ARIMA | 0.91 | 0.97 | 1.00 | 0.98 | 0.94 | 0.76 | 0.93 | 1.00 | 0.86 | 0.81 | 0.99 | 0.86 | 0.89 | 0.81 |
| VETS | 0.89 | 0.95 | 0.98 | 1.00 | 0.92 | 0.78 | 0.95 | 0.98 | 0.84 | 0.79 | 0.97 | 0.88 | 0.87 | 0.79 |
| VETSdiag | 0.97 | 0.97 | 0.94 | 0.92 | 1.00 | 0.70 | 0.87 | 0.94 | 0.91 | 0.87 | 0.95 | 0.80 | 0.95 | 0.86 |
| VECM | 0.68 | 0.73 | 0.76 | 0.78 | 0.70 | 1.00 | 0.82 | 0.76 | 0.63 | 0.58 | 0.75 | 0.90 | 0.66 | 0.58 |
| ETSagg | 0.85 | 0.90 | 0.93 | 0.95 | 0.87 | 0.82 | 1.00 | 0.93 | 0.79 | 0.74 | 0.92 | 0.93 | 0.82 | 0.74 |
| ETSagglog | 0.91 | 0.96 | 1.00 | 0.98 | 0.94 | 0.76 | 0.93 | 1.00 | 0.85 | 0.81 | 0.99 | 0.86 | 0.89 | 0.80 |
| ARIMAagg | 0.94 | 0.89 | 0.86 | 0.84 | 0.91 | 0.63 | 0.79 | 0.85 | 1.00 | 0.95 | 0.87 | 0.72 | 0.96 | 0.95 |
| VETSagg | 0.89 | 0.84 | 0.81 | 0.79 | 0.87 | 0.58 | 0.74 | 0.81 | 0.95 | 1.00 | 0.82 | 0.67 | 0.92 | 1.00 |
| VETSaggdiag | 0.92 | 0.98 | 0.99 | 0.97 | 0.95 | 0.75 | 0.92 | 0.99 | 0.87 | 0.82 | 1.00 | 0.85 | 0.90 | 0.82 |
| ETSgroup | 0.77 | 0.83 | 0.86 | 0.88 | 0.80 | 0.90 | 0.93 | 0.86 | 0.72 | 0.67 | 0.85 | 1.00 | 0.75 | 0.67 |
| ETSgrouplog | 0.98 | 0.92 | 0.89 | 0.87 | 0.95 | 0.66 | 0.82 | 0.89 | 0.96 | 0.92 | 0.90 | 0.75 | 1.00 | 0.91 |
| ARIMAgroup | 0.89 | 0.84 | 0.81 | 0.79 | 0.86 | 0.58 | 0.74 | 0.80 | 0.95 | 1.00 | 0.82 | 0.67 | 0.91 | 1.00 |

Table 42: p-values of the paired F-tests.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSagg | VETSaggdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.95 | 0.91 | 0.89 | 0.97 | 0.68 | 0.85 | 0.91 | 0.94 | 0.89 | 0.92 | 0.77 | 0.98 | 0.89 |
| ETSlog | 0.95 | 1.00 | 0.97 | 0.95 | 0.97 | 0.73 | 0.90 | 0.96 | 0.89 | 0.84 | 0.98 | 0.83 | 0.92 | 0.84 |
| ARIMA | 0.91 | 0.97 | 1.00 | 0.98 | 0.94 | 0.76 | 0.93 | 1.00 | 0.86 | 0.81 | 0.99 | 0.86 | 0.89 | 0.81 |
| VETS | 0.89 | 0.95 | 0.98 | 1.00 | 0.92 | 0.78 | 0.95 | 0.98 | 0.84 | 0.79 | 0.97 | 0.88 | 0.87 | 0.79 |
| VETSdiag | 0.97 | 0.97 | 0.94 | 0.92 | 1.00 | 0.70 | 0.87 | 0.94 | 0.91 | 0.87 | 0.95 | 0.80 | 0.95 | 0.86 |
| VECM | 0.68 | 0.73 | 0.76 | 0.78 | 0.70 | 1.00 | 0.82 | 0.76 | 0.63 | 0.58 | 0.75 | 0.90 | 0.66 | 0.58 |
| ETSagg | 0.85 | 0.90 | 0.93 | 0.95 | 0.87 | 0.82 | 1.00 | 0.93 | 0.79 | 0.74 | 0.92 | 0.93 | 0.82 | 0.74 |
| ETSagglog | 0.91 | 0.96 | 1.00 | 0.98 | 0.94 | 0.76 | 0.93 | 1.00 | 0.85 | 0.81 | 0.99 | 0.86 | 0.89 | 0.80 |
| ARIMAagg | 0.94 | 0.89 | 0.86 | 0.84 | 0.91 | 0.63 | 0.79 | 0.85 | 1.00 | 0.95 | 0.87 | 0.72 | 0.96 | 0.95 |
| VETSagg | 0.89 | 0.84 | 0.81 | 0.79 | 0.87 | 0.58 | 0.74 | 0.81 | 0.95 | 1.00 | 0.82 | 0.67 | 0.92 | 1.00 |
| VETSaggdiag | 0.92 | 0.98 | 0.99 | 0.97 | 0.95 | 0.75 | 0.92 | 0.99 | 0.87 | 0.82 | 1.00 | 0.85 | 0.90 | 0.82 |
| ETSgroup | 0.77 | 0.83 | 0.86 | 0.88 | 0.80 | 0.90 | 0.93 | 0.86 | 0.72 | 0.67 | 0.85 | 1.00 | 0.75 | 0.67 |
| ETSgrouplog | 0.98 | 0.92 | 0.89 | 0.87 | 0.95 | 0.66 | 0.82 | 0.89 | 0.96 | 0.92 | 0.90 | 0.75 | 1.00 | 0.91 |
| ARIMAgroup | 0.89 | 0.84 | 0.81 | 0.79 | 0.86 | 0.58 | 0.74 | 0.80 | 0.95 | 1.00 | 0.82 | 0.67 | 0.91 | 1.00 |

Table 43: p-values of the paired Bartlett tests.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSgroup | VETSgroupdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.98 | 0.97 | 0.70 | 0.91 | 0.25 | 0.94 | 0.92 | 0.71 | 0.89 | 0.82 | 0.80 | 0.93 | 0.90 |
| ETSlog | 0.98 | 1.00 | 0.95 | 0.68 | 0.89 | 0.24 | 0.92 | 0.90 | 0.73 | 0.91 | 0.84 | 0.78 | 0.91 | 0.91 |
| ARIMA | 0.97 | 0.95 | 1.00 | 0.73 | 0.94 | 0.27 | 0.97 | 0.95 | 0.68 | 0.86 | 0.79 | 0.83 | 0.96 | 0.86 |
| VETS | 0.70 | 0.68 | 0.73 | 1.00 | 0.79 | 0.44 | 0.76 | 0.78 | 0.45 | 0.60 | 0.54 | 0.89 | 0.77 | 0.61 |
| VETSdiag | 0.91 | 0.89 | 0.94 | 0.79 | 1.00 | 0.30 | 0.97 | 0.99 | 0.63 | 0.80 | 0.73 | 0.89 | 0.98 | 0.80 |
| VECM | 0.25 | 0.24 | 0.27 | 0.44 | 0.30 | 1.00 | 0.28 | 0.30 | 0.14 | 0.20 | 0.18 | 0.37 | 0.29 | 0.21 |
| ETSagg | 0.94 | 0.92 | 0.97 | 0.76 | 0.97 | 0.28 | 1.00 | 0.98 | 0.66 | 0.83 | 0.76 | 0.86 | 0.99 | 0.84 |
| ETSagglog | 0.92 | 0.90 | 0.95 | 0.78 | 0.99 | 0.30 | 0.98 | 1.00 | 0.64 | 0.81 | 0.74 | 0.88 | 0.99 | 0.81 |
| ARIMAagg | 0.71 | 0.73 | 0.68 | 0.45 | 0.63 | 0.14 | 0.66 | 0.64 | 1.00 | 0.82 | 0.89 | 0.54 | 0.65 | 0.81 |
| VETSgroup | 0.89 | 0.91 | 0.86 | 0.60 | 0.80 | 0.20 | 0.83 | 0.81 | 0.82 | 1.00 | 0.93 | 0.70 | 0.82 | 0.99 |
| VETSgroupdiag | 0.82 | 0.84 | 0.79 | 0.54 | 0.73 | 0.18 | 0.76 | 0.74 | 0.89 | 0.93 | 1.00 | 0.63 | 0.75 | 0.92 |
| ETSgroup | 0.80 | 0.78 | 0.83 | 0.89 | 0.89 | 0.37 | 0.86 | 0.88 | 0.54 | 0.70 | 0.63 | 1.00 | 0.87 | 0.70 |
| ETSgrouplog | 0.93 | 0.91 | 0.96 | 0.77 | 0.98 | 0.29 | 0.99 | 0.99 | 0.65 | 0.82 | 0.75 | 0.87 | 1.00 | 0.83 |
| ARIMAgroup | 0.90 | 0.91 | 0.86 | 0.61 | 0.80 | 0.21 | 0.84 | 0.81 | 0.81 | 0.99 | 0.92 | 0.70 | 0.83 | 1.00 |

Table 44: p-values of the absolute paired F-tests.

| | ETS | ETSlog | ARIMA | VETS | VETSdiag | VECM | ETSagg | ETSagglog | ARIMAagg | VETSgroup | VETSgroupdiag | ETSgroup | ETSgrouplog | ARIMAgroup |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETS | 1.00 | 0.98 | 0.97 | 0.70 | 0.91 | 0.25 | 0.94 | 0.92 | 0.71 | 0.89 | 0.82 | 0.80 | 0.93 | 0.90 |
| ETSlog | 0.98 | 1.00 | 0.95 | 0.68 | 0.89 | 0.24 | 0.92 | 0.90 | 0.73 | 0.91 | 0.84 | 0.78 | 0.91 | 0.91 |
| ARIMA | 0.97 | 0.95 | 1.00 | 0.73 | 0.94 | 0.27 | 0.97 | 0.95 | 0.68 | 0.86 | 0.79 | 0.83 | 0.96 | 0.86 |
| VETS | 0.70 | 0.68 | 0.73 | 1.00 | 0.79 | 0.44 | 0.76 | 0.78 | 0.45 | 0.60 | 0.54 | 0.89 | 0.77 | 0.60 |
| VETSdiag | 0.91 | 0.89 | 0.94 | 0.79 | 1.00 | 0.30 | 0.97 | 0.99 | 0.63 | 0.80 | 0.73 | 0.89 | 0.98 | 0.80 |
| VECM | 0.25 | 0.24 | 0.27 | 0.44 | 0.30 | 1.00 | 0.28 | 0.30 | 0.14 | 0.20 | 0.18 | 0.36 | 0.29 | 0.21 |
| ETSagg | 0.94 | 0.92 | 0.97 | 0.76 | 0.97 | 0.28 | 1.00 | 0.98 | 0.66 | 0.83 | 0.76 | 0.86 | 0.99 | 0.83 |
| ETSagglog | 0.92 | 0.90 | 0.95 | 0.78 | 0.99 | 0.30 | 0.98 | 1.00 | 0.64 | 0.81 | 0.74 | 0.88 | 0.99 | 0.81 |
| ARIMAagg | 0.71 | 0.73 | 0.68 | 0.45 | 0.63 | 0.14 | 0.66 | 0.64 | 1.00 | 0.82 | 0.89 | 0.54 | 0.65 | 0.81 |
| VETSgroup | 0.89 | 0.91 | 0.86 | 0.60 | 0.80 | 0.20 | 0.83 | 0.81 | 0.82 | 1.00 | 0.93 | 0.69 | 0.82 | 0.99 |
| VETSgroupdiag | 0.82 | 0.84 | 0.79 | 0.54 | 0.73 | 0.18 | 0.76 | 0.74 | 0.89 | 0.93 | 1.00 | 0.63 | 0.75 | 0.92 |
| ETSgroup | 0.80 | 0.78 | 0.83 | 0.89 | 0.89 | 0.36 | 0.86 | 0.88 | 0.54 | 0.69 | 0.63 | 1.00 | 0.87 | 0.70 |
| ETSgrouplog | 0.93 | 0.91 | 0.96 | 0.77 | 0.98 | 0.29 | 0.99 | 0.99 | 0.65 | 0.82 | 0.75 | 0.87 | 1.00 | 0.83 |
| ARIMAgroup | 0.90 | 0.91 | 0.86 | 0.60 | 0.80 | 0.21 | 0.83 | 0.81 | 0.81 | 0.99 | 0.92 | 0.70 | 0.83 | 1.00 |

Table 45: p-values of the absolute paired Bartlett's tests.