

ECONOMETRIC MODEL OF GENERALIZED LINEAR REGRESSION FOR PREDICTING FUTURE NUMBERS OF INSURED PEOPLE BY GENERAL HEALTH INSURANCE COMPANY OF THE CZECH REPUBLIC – COMPARISON OF RESULTS AFTER FIVE YEARS

Jaromír Běláček¹

¹Department of Statistics and Operation Analysis, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/ecos2026_003

ABSTRACT

This paper compares the results of predicting the number of insured persons of the General Health Insurance Company of the Czech Republic in 2018, based on a generalized-discounted linear regression model (GLM), which formalizes the projected number of insured persons in age groups through exogenous demographic, re-reregistration, and formal linear and quadratic trend components, with reality observed in 2018–2022. Compared to standard in-sample methods, derived population projections, or the traditional indexing method, the 2018 method, where the calibration of discount factors α meets the criteria for maximizing the determination coefficient $R^2(\alpha)$, overall yields better results. Accurate predictions of future numbers of insured individuals have the potential to better calculate and plan future expenditures of the health insurance funds during current period of population aging, and for several years ahead, too.

Keywords: public health insurance, time series forecasting, derived population projection, exponential smoothing, coefficient of determination, weighted-discounted least squares

JEL Code: C53, I11, J11



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license.
<https://creativecommons.org/licenses/by-sa/4.0/deed.en>

1 DEVELOPMENT OF THE ISSUE

In 2016–17, the task of predicting the future number of its insured individuals was addressed at the General Health Insurance Company of the Czech Republic (hereinafter referred to as VZP ČR). In the paper (Běláček, 2018), the deterministic model for predicting the future number of VZP insured was presented, based on the upper variant of the demographic forecast by the Czech Statistical Office for the Czech Republic from 2013 and on additive adjustments of extrapolated quarterly re-registration balances (between VZP and other health insurance companies in the Czech Republic) – see the model (1) in Section 2.1 below. In the subsequent paper (Běláček, 2019), a model of generalized discounted linear regression was described, where the weights of retrospective observations might be directly proportional to exponentially decreasing discounts α ($\alpha < 1$) assigned formally to time indices. In our review article, a universal algorithm described in (Cipra, 1986, p. 233) was used to estimate parameters, which also allows the choice of discounts $\alpha > 1$ – this model is presented in section 2.2.

In this paper, we compare the results of the prediction published in (Běláček, 2019) from the threshold date of 12/31/2017 to the actual numbers of insured persons obtained for the years 2018–22 retrospectively from the VZP yearbooks as of 12/31 of each calendar year (Yearbooks of the VZP of the Czech Republic for the years 2018–22, 2023). The aim of this paper is to compare the numbers of VZP insured persons determined after five years in total and in age groups – a) with the values predicted from the threshold year 2017; – b) with the predictions based on in-sample minimization of squared deviations in the last five years of the pre-threshold period; – c) compared to the best fit with real data from 2018–22 found ad poste – see paragraphs 2.3 and 3. The discussion of the presented results will provide general conclusions and recommendations for the possibility of using this technologies in comparable applications.

2 DATA SOURCES AND METHODOLOGY

2.1 Data and deterministic models

To project the future number in VZP insured persons, the final quarterly time series in total and in five-year age groups and the similarly structured balances of re-registrations of insured persons between VZP and rest health insurance companies in the Czech Republic, taken from the internal databases of VZP, were used. The primary time series (in thousands of persons) are visualized by black dots in Figures 1–5 for the total of all insured persons ($x = '0+'$) and five-year age groups ($x = '0-4', \dots, '85+'$). An important part of these figures are three projection deterministic models generated in paper (Běláček, 2018) using the formula

$$\widehat{P}_x^t = \left(\widehat{P}_x^{t-1} + S_x^t \right) \cdot (D_x^t / D_x^{t-1}), \quad t = t_0, t_0 + 1, \dots, \quad (1)$$

where $\widehat{P}_x^{t_0} \equiv P_x^{t_0}$ denotes the known entry numbers of insured persons at age x in the threshold year t_0 (at the end of the entry quarter), D_x^t is the mean number of living persons or interpolated per quarter from data officially projected by the Czech Statistical Office at time t (see *Projekce obyvatelstva ČR do r. 2100*, 2013) and S_x^t denotes the balance of VZP re-registrations assumed (or extrapolated) for the period between years $t - 1$ a t (see also in Běláček, 2019 for more details).

With the special choice $S_x^t = 0$ for all future times t model (1) is reduced to the common model of the so-called derived population projection (in Figures 1–5 marked as ‘D-model’ shown by light diamonds). But we can also consider the alternative of pure projection of cumulative re-registrations (in Figures 1–5 marked as ‘E-model’ represented by dashed connecting lines), which we obtain by choosing $D_x^t/D_x^{t-1} = 1$. This second development scenario is not realistic for most age groups, but it can be well used for the purposes of assessing the possible mutual synergistic influence of both explanatory components (demographic and re-registrations) on the future development of the number of insured persons. This is synthesized within the framework of the third deterministic model (in legends of Figures 1–5 marked as ‘D+E-model’ depicted by dark diamonds) created by directly applying formula (1).

2.2 Generalized linear models

All three deterministic models above were calculated in relation to the threshold date of 12/31/2017 toward the future (prospectively), but corresponding time series can also be calculated toward the past (retrospectively), as shown in Figures 1–5. These representations are important from the point of view in comparison with results for the following generalized linear regression models and with the actual numbers of insured persons determined later. They point out in what form mechanically generated demographic or reregistration variables can or must enter the subsequent weighted linear regression models as exogenous variables. Their formal significance in regression models can then be assessed using methods of statistic inference.

The most general linear regression model used later to predict the future number of VZP insured persons in each age group (see Běláček, 2019, p. 5) is based on a four-parameter linear regression model for difference

$$\Delta(P_x^t) = b_x + c_x \cdot t + d_x \cdot \Delta(D_x^t) + e_x \cdot S_x^t + \varepsilon(0, w_t \cdot \sigma_x^2), \quad t = 1, \dots, t_0, \quad (2)$$

where $\Delta(P_x^t) = (P_x^t - P_x^{t-1})$ denotes the retrospective difference in the number of insured persons in group x between the ends of adjacent quarters, t is a time variable (with a projection threshold of $t_0 = 12/31/2017$), $\Delta(D_x^t)$ denotes the interpolated quarterly increments of the demographic variable (derived from historical data for the Czech Republic) and S_x^t is the balance of re-registrations from formula (1); further, let us denote by the symbol $\theta_x = (b_x; c_x; d_x; e_x)$ the vector of linear regression parameters from (2) and $\varepsilon(0, w_t \cdot \sigma_x^2)$ the error term – with a zero mean value and a variance in the form of the product $w_t \cdot \sigma_x^2$ for the weight factor w_t functionally independent of the system parameter σ_x^2 . We consider important that all terms on the right-hand side in (2) are now part of the standard additive regression model, which includes the earlier deterministic models from (1) be his special cases (for fixed choices: $b_x = c_x = 0$ and $d_x = 1$ or $e_x = 1$).

We obtained the parameter estimates $\hat{\theta}_x = (\hat{b}_x; \hat{c}_x; \hat{d}_x; \hat{e}_x)$ in model (2) by minimizing (by $\alpha_x \in [0.65; 1.0]$) the weighted sum of squares

$$\sum_{k=1}^{t_0} w_k \cdot [\Delta(P_x^k) - b_x - c_x \cdot k - d_x \cdot \Delta(D_x^k) - e_x \cdot S_x^k]^2, \quad (3)$$

namely through a matrix formula

$$\widehat{\boldsymbol{\theta}}_x = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\Delta\left(P_x\right) = \left(\left(\mathbf{X}\mathbf{W}^{\frac{1}{2}}\right)' \left(\mathbf{W}^{\frac{1}{2}}\mathbf{X}\right)\right)^{-1} \left(\mathbf{X}\mathbf{W}^{\frac{1}{2}}\right)' \left(\mathbf{W}^{\frac{1}{2}}\Delta\left(P_x\right)\right), \quad (4)$$

where $\mathbf{X}$ is the matrix of the system of linear equations from (2), $\Delta(P_x)$ its corresponding vector with components $\Delta(P_x^t)$ and $\mathbf{W}$ a square diagonal matrix with diagonal elements $w_t = \alpha_x^{-t}$ for the discount factor $\alpha_x > 0$. This means that for $\alpha_x < 1$ we assign an exponentially decreasing influence to the historical data (with gradual distance from the threshold value t_0), for $\alpha_x = 1$ the estimates $\widehat{\boldsymbol{\theta}}_x$ of the parameter vector $\boldsymbol{\theta}_x$ correspond to the classical unweighted linear regression (all past values of the time series have the same weight) and for $\alpha_x > 1$ the historical values of the time series increase their weight exponentially. Substituting the optimized estimates $\widehat{\boldsymbol{\theta}}_x$ from (4) into formula (3) for $\boldsymbol{\theta}_x$ we obtain estimates $\widehat{\sigma}_x^2$ for the variance parameters σ_x^2 and we are counting the coefficients of determination $\widehat{R}^2(\alpha_x)$, depending on the selected discount level α_x (otherwise also the percentages of variance of the dependent difference variable explained by each of the selected regression models in age group). The final prediction of the number of insured persons for times $t = t_0 + 1, \dots, T$, where T denotes an unspecified projection horizon, was finally obtained by cumulative calculation

$$\begin{aligned} \widehat{P}_x^t(\alpha_x) &= \widehat{P}_x^{t-1}(\alpha_x) + \Delta\left(\widehat{P}_x^t(\alpha_x)\right) = P_x^{t_0} + \sum_{k=t_0+1}^t \left[\widehat{b}_x + \widehat{c}_x \cdot k + \widehat{d}_x \cdot \Delta\left(D_x^k\right) + \widehat{e}_x \cdot S_x^k \right] \\ &= \widehat{A}_x + \widehat{B}_x \cdot (t - t_0) + \widehat{C}_x \cdot (t - t_0)^2 + \widehat{D}_x \cdot \sum_{k=t_0+1}^t \Delta\left(D_x^k\right) + \widehat{E}_x \cdot \sum_{k=t_0+1}^t S_x^k, \end{aligned} \quad (5)$$

where $\Delta(\widehat{P}_x^t(\alpha_x))$ is the prediction obtained from the difference model (2) after substituting $\widehat{\boldsymbol{\theta}}_x = \widehat{\boldsymbol{\theta}}(\alpha_x)$ for components of $\boldsymbol{\theta}_x$ (and after omitting the error term); for the newly calculated parameters in model (5) the easily verifiable relations hold: $\widehat{A}_x = P_x^{t_0}$, $\widehat{B}_x = \widehat{b}_x + \frac{(t_0+1)}{2} \cdot \widehat{c}_x$, $\widehat{C}_x = \frac{1}{2} \cdot \widehat{c}_x$, $\widehat{D}_x = \widehat{d}_x$, $\widehat{E}_x = \widehat{e}_x$. The advantage of the calculation method based on the difference model is the possibility of setting the absolute term $\widehat{A}_x$ in the cumulative model (5) directly to the last known value in the threshold year t_0 , so we do not have to estimate this formal term, although formally important for extrapolation.

2.3 Comparison of model predictions with reality

The final data as of 12/31 of each year from the VZP yearbooks are shown in Graphs 1–5 with large fulfilled black circles until 2022. In 2002–2017, they do not visually differ from the quarterly data from the internal VZP databases. However, in 2022, massive increases were recorded in most age groups due to the influx of immigrants from Ukraine, as VZP was the majority health insurance company registering these refugees in the Czech Republic at that time. Of course, the prediction models from 2017 could not take this unexpected entry into account. However, since Ukrainians are not separated from the portfolio of other insured persons in the age structures of the annual VZP reports, their numbers in each five-year age group were qualifiedly estimated based on the balance of foreign migration of the Czech Republic by age and citizenship – see Tables H.12 and H.14 in (Demographic Yearbook of the Czech Republic – 2022, 2023). The numbers of VZP insured persons after deducting the estimated balance of Ukrainians from 2022 are shown as large circles without filling in Figures 1–5.

Already as part of solving the original problem of predicting the numbers of insured persons, a user application was created in MS Excel at 2017 (currently see Excel application of the prediction model, 2025), which allowed the selection of optional user specifications: For example, the

historical evolution of the number of VZP insured persons in each age group could be alternated by time series from the database of all health insured persons in the Czech Republic (a more conservative, but later universally useful variant), the future development of VZP re-registration balances could be selected in three variants of extrapolation: a) constant zero; b) more conservative logarithmic – later selected as optimal one; c) more progressive logarithmic. Furthermore, the graphs visualize quarterly predictions for the three deterministic models above and selected GLM models (currently only for ‘ABDE’ and ‘ABCDE’) under the mode of several different options of choosing discount factors α_x (displayed in the header of graphs), all with a projection horizon of $T=12/31/2032$. GLM models are supplemented in the legends with values of determination coefficients $\widehat{R}^2(\alpha_x)$.

Hereafter, for the couple of discounted regression models mentioned above, the graphs preferentially display extrapolation according to the value of the individually selected discount factors α_x (within the “0 = ALFA by choice” mode). After archiving the $\widehat{R}^2(\alpha_x)$ values following the experimental choice of $\alpha_x \in [0.65; 1.0]$ with a step of 0.01 in each age group the optimum alpha has been specified, how it was used for prediction in 2018. The visualization of these curves can be automatically called up within the mode „1 or 2 = /optALFA for ‘ABDE’ or ‘ABCDE’/“, which is also the current display in Graphs 1–5. However, since the chosen optimal selection method for α_x was used rather intuitively and heuristically in 2018, now an attempt was made to compare it with the results of α_x selections using the commonly recommended version of the in-sample method (Bělohradský – Šolc, 2018), which, instead of discounting within all historical data, minimizes only the unweighted sums of squares in the last few pre-threshold years. Thus, in Excel, values of $\alpha_x \geq 0.65$ were step by step archived and after that found such for which was in age group x minimized the following sum of squares:

$$SS_{2012-16} = \sum_{k=t_0-5}^{t_0-1} [P_x^k - \widehat{P}_x^k(\alpha_x)]^2, \quad (6)$$

i.e. for the last five years before the threshold period. This α_x option can be initialized in the Excel application by setting mode “3 or 4 /= minALFA for 5year’s old residua ‘ABDE’ or ‘ABCDE’/” for automatic recalling such minimum alpha’s (with the identification „-5Y’s ALFA“ in the header of created interactive charts).

Finally, we obtained the final variant of choosing discount factors α_x by minimizing through formula

$$SS_{2018-22} = \sum_{k=t_0+1}^{t_0+5} [P_x^k - \widehat{P}_x^k(\alpha_x)]^2, \quad (7)$$

which is an analogy (6), only the addition index is now running the final balances of insured persons as of 12/31/2018-2022. It is therefore the square error of the prediction with respect to the real numbers of VZP insured persons in groups (adjusted in 2022 by subtracting citizens of Ukrainian nationality). This option can be initialized in the Excel application under the mode “5 or 6 /= finALFA after 5year’s in ‘ABDE’ or ‘ABCDE’/” for automatic recalling such final alpha’s found after five years evolution. These final extrapolations are therefore the best which could be achieved by in-sample discounted-GLM method, if we were able to select the best α_x for the prediction in the following five years already in the threshold year $t_0 = 2018$ (but of course this can never be achieved exactly due to the stochastic nature of the modeled processes themselves). In headers of created interactive charts this option is displayed as “finALFA”.

Based on the comparison of the square roots of the sums of squares ad (7), which we obtained by substituting the values $\widehat{P}_x^t(\alpha_x)$ into (7) for α_x corresponding to the choices of *optimum*, *minimum* or *final alpha* in each age group x , we can now objectively assess the quality of the models' prediction after five years. In summary, the models can be evaluated by the square roots of the sums of squares ad (7) over all age groups $x = '0-4', \dots, '85+'$. (Such prediction for all deterministic and in all 18 age groups marginally optimized GLM models 'ABDE' and 'ABCDE' in 2018 is shown in Figure 2, for comparison.) The square roots of the sums of squares ad (7) can be interpreted as formal Euclidean distances of predictions from reality observed in the years 2018–22. In Figures 1–5, they are displayed within the legends for all plotted models after the designation ' $\sqrt{SS}$ '. 3.

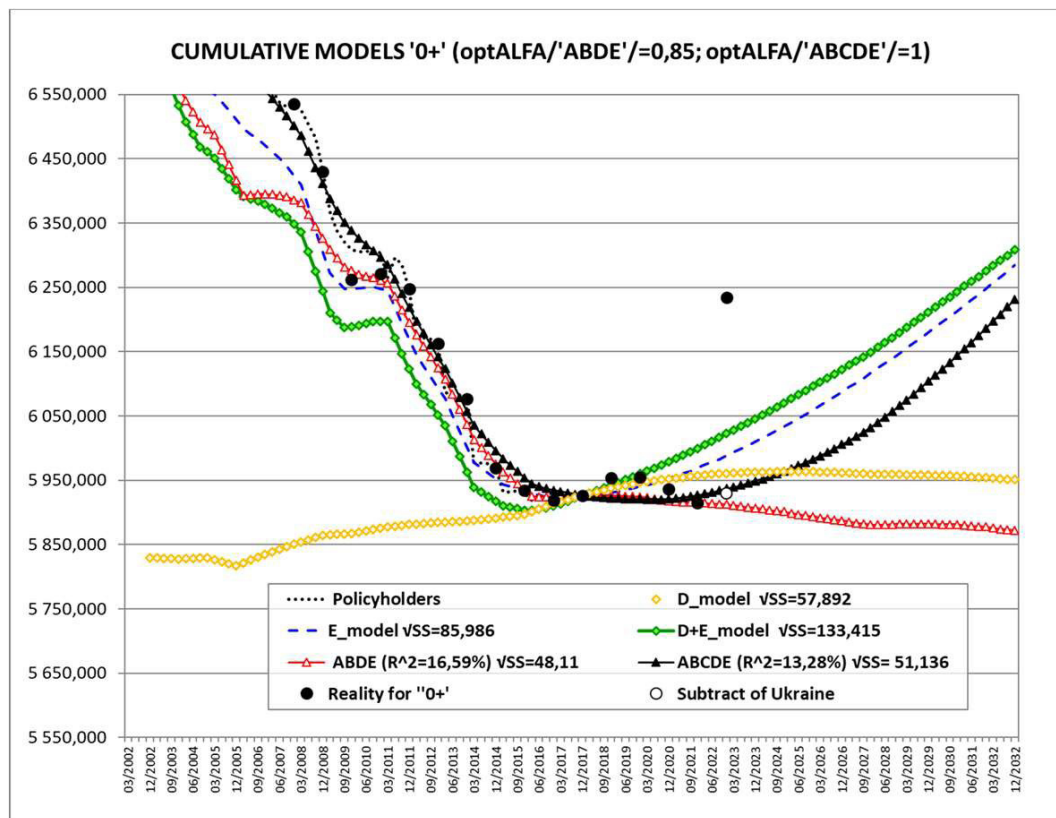


Fig. 1: Quarterly time series of the number of VZP CR insured people 2002-17 in group '0+' (estimated as the whole), their prospective and retrospective projections using three deterministic and two discounted regression models calibrated in 2017 in comparison with the real evolution in 2018-22 (see Legend inside graph)

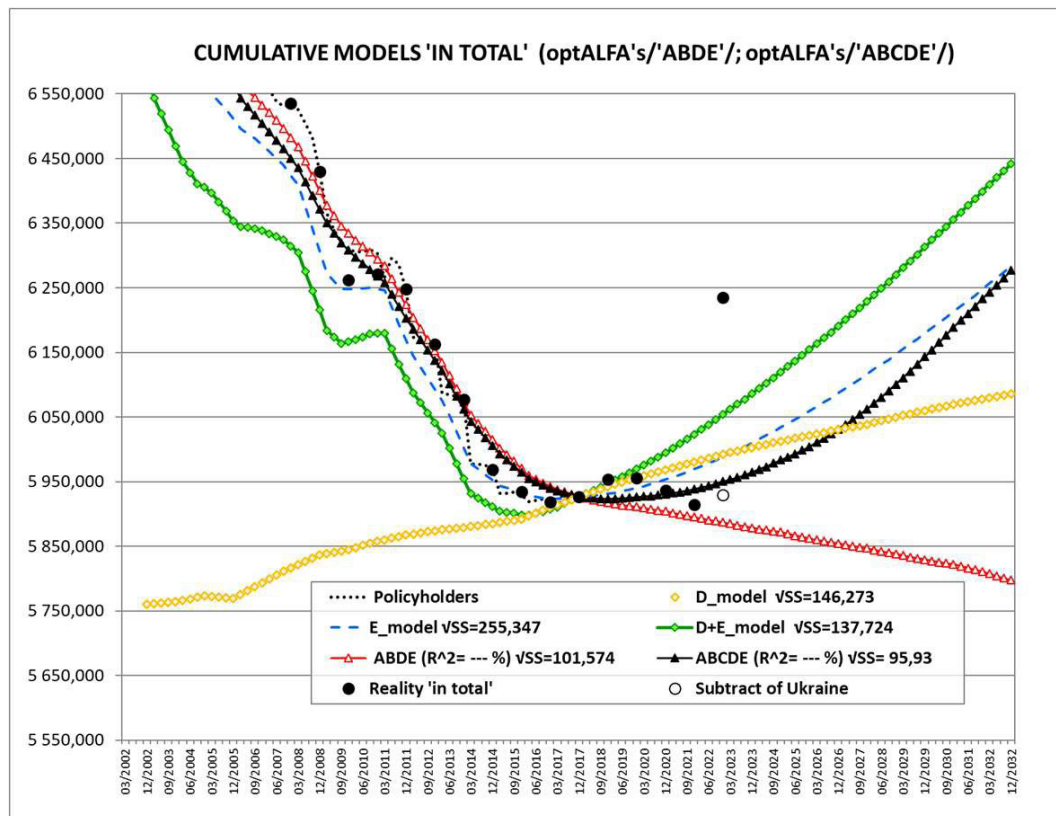


Fig. 2: Quarterly time series of the number of VZP CR insured people from 2002-2017 ,in total' (sum of age groups), their prospective and retrospective projections using three deterministic and two discounted re-gression models calibrated separately in each group '0-4', ... , '85+' in 2017 in comparison with the real evolution in 2018-22 (see Legend inside graph)

3 RESULTS

The quality of data fitting by extrapolation curves can be assessed by visually comparison the model curves with the actual values among 2018–2022 in Graphs 1–5 marked by circles (the threshold year $t_0 = 12/31/2017$ is located in the middle of the time axis of each graph). A more precise numerical evaluation of the quality of fitting for the variants of the models ,ABCDE' and ,ABDE' is found in Tables 1–3.

Table 1 takes the values of optimum alpha and their corresponding $R^2(\alpha x)$ coefficients decisive for the optimization of both generalized models from (Běláček, 2019). With the exception of extrapolations for the group of insured persons as the whole ('0+') and for the group of the youngest ('0-4'), where the coefficients of determination did not exceed 33.3% for any of the discounted linear models, the smoothing ability of the 'ABCDE' and 'ABDE' models is generally high in all other age groups (see columns 3 and 6 in Table 1). A more detailed analysis of the statistical significance of the model parameters showed (across age groups with the exception of '0+') that the demographic component is universally the most significant among the exogenous variables (see in Běláček, J., 2019 for more details). On the contrary, at the level of the 'ABCDE' models optimized in

Table 1, the least significant quadratic term by the time variable (coefficient Cx) appears to be. For most extrapolations, we could seemingly make do with the sub-model 'ABDE', whose linear trend component was statistically significantly different from zero ($p < 0.05$) in spectrum of age ranges '5-19' and '45+' and the re-registration component in the groups '10-14', '40-49' and '55-59'.

Tab. 1 Table of optimum discounts αx calibrated for weighted LSE estimators αx (using (2) at 2002-17 quarterly data in addition with criterions of maximization $R^2(\alpha x)$ as in part 2.2) for models 'ABDE' and 'ABCDE', their optimum $R^2(opt. \alpha x)$ values and the resulting $\sqrt{SS_{2018-22}}$ deviations from reality 2018-22 (in thousands of people)

model:	'ABDE'	'ABDE'	'ABDE'	'ABCDE'	'ABCDE'	'ABCDE'
group x	$opt. \alpha x$	$R^2(opt. \alpha x)$	$\sqrt{SS_{2018-22}}$	$opt. \alpha x$	$R^2(opt. \alpha x)$	$\sqrt{SS_{2018-22}}$
'0+'	0.85	16.59%	48.110	1.00	13.28%	51.136
'0-4'	1.00	27.44%	32.757	1.00	32.14%	10.084
'5-9'	0.69	52.68%	21.611	0.69	52.68%	21.931
'10-14'	0.99	82.47%	25.009	0.99	82.75%	14.098
'15-19'	0.86	76.50%	7.272	1.00	64.91%	12.379
'20-24'	1.00	39.75%	11.559	1.00	41.69%	27.264
'25-29'	1.00	47.21%	28.359	1.00	49.12%	58.509
'30-34'	1.00	66.04%	10.095	1.00	66.18%	5.586
'35-39'	0.93	72.07%	22.766	0.94	73.45%	9.961
'40-44'	0.99	49.01%	41.759	1.00	49.47%	30.001
'45-49'	1.00	83.03%	24.798	1.00	83.23%	27.327
'50-54'	0.96	71.66%	6.121	0.96	71.77%	1.908
'55-59'	1.00	79.01%	4.867	1.00	79.09%	18.840
'60-64'	1.00	79.93%	10.527	1.00	79.98%	9.127
'65-69'	0.75	85.02%	3.794	1.00	83.65%	5.736
'70-74'	1.00	93.05%	2.448	1.00	93.15%	17.111
'75-79'	0.94	85.73%	15.014	0.94	85.82%	14.237
'80-84'	0.86	36.66%	21.566	1.00	82.29%	17.384
'85+'	1.00	64.78%	57.438	1.00	67.10%	32.429
'in total'	---	---	101.574	---	---	95.930

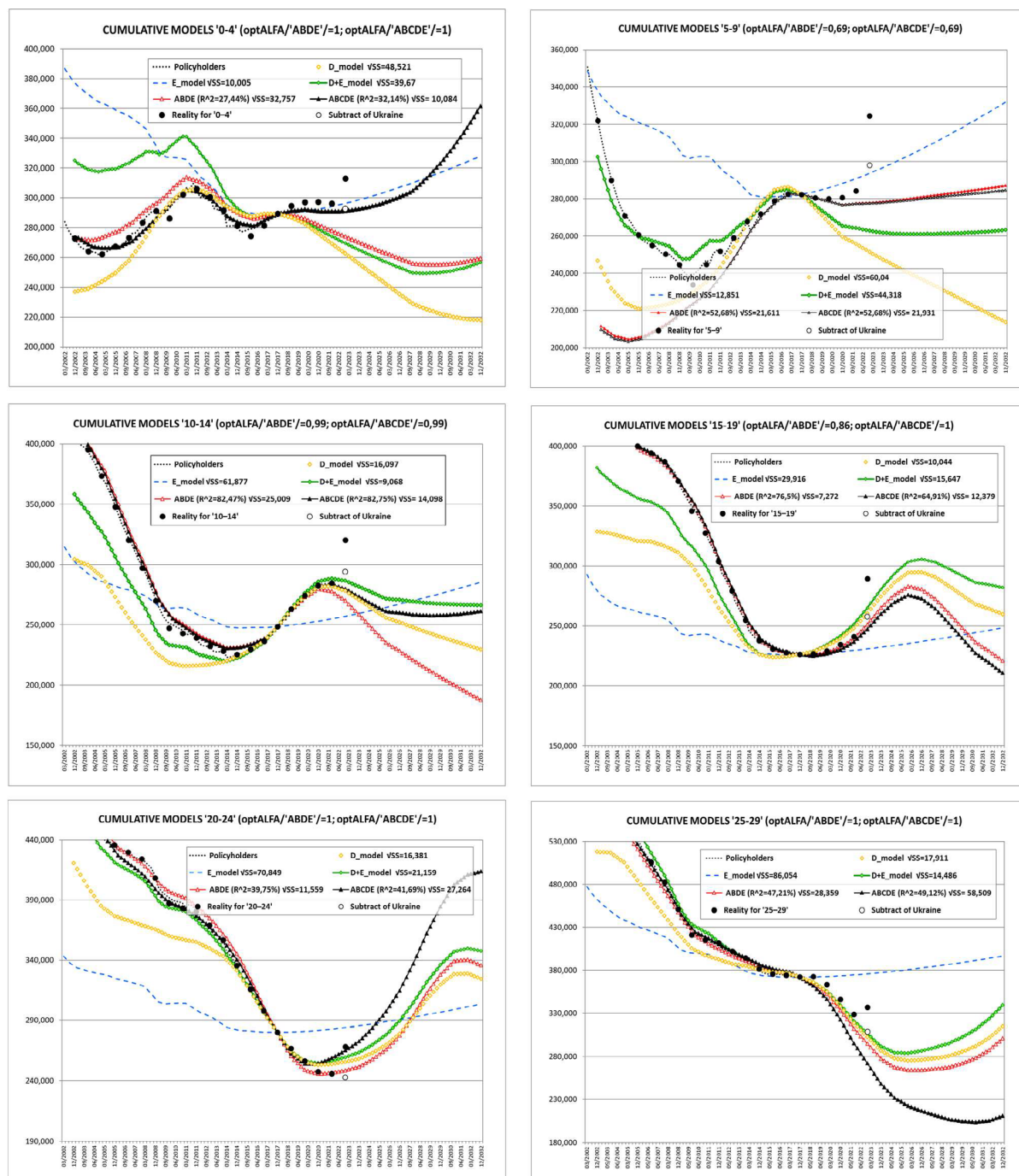


Fig. 3: a-f): Quarterly time series of the number of VZP CR insured people from 2002-17 in groups ,0-4', ..., ,25-29', their prospective and retrospective projections using three deterministic and two discounted re-gression models calibrated in 2017 in comparison with the real evolution in 2018-22 (see Legends inside graphs)

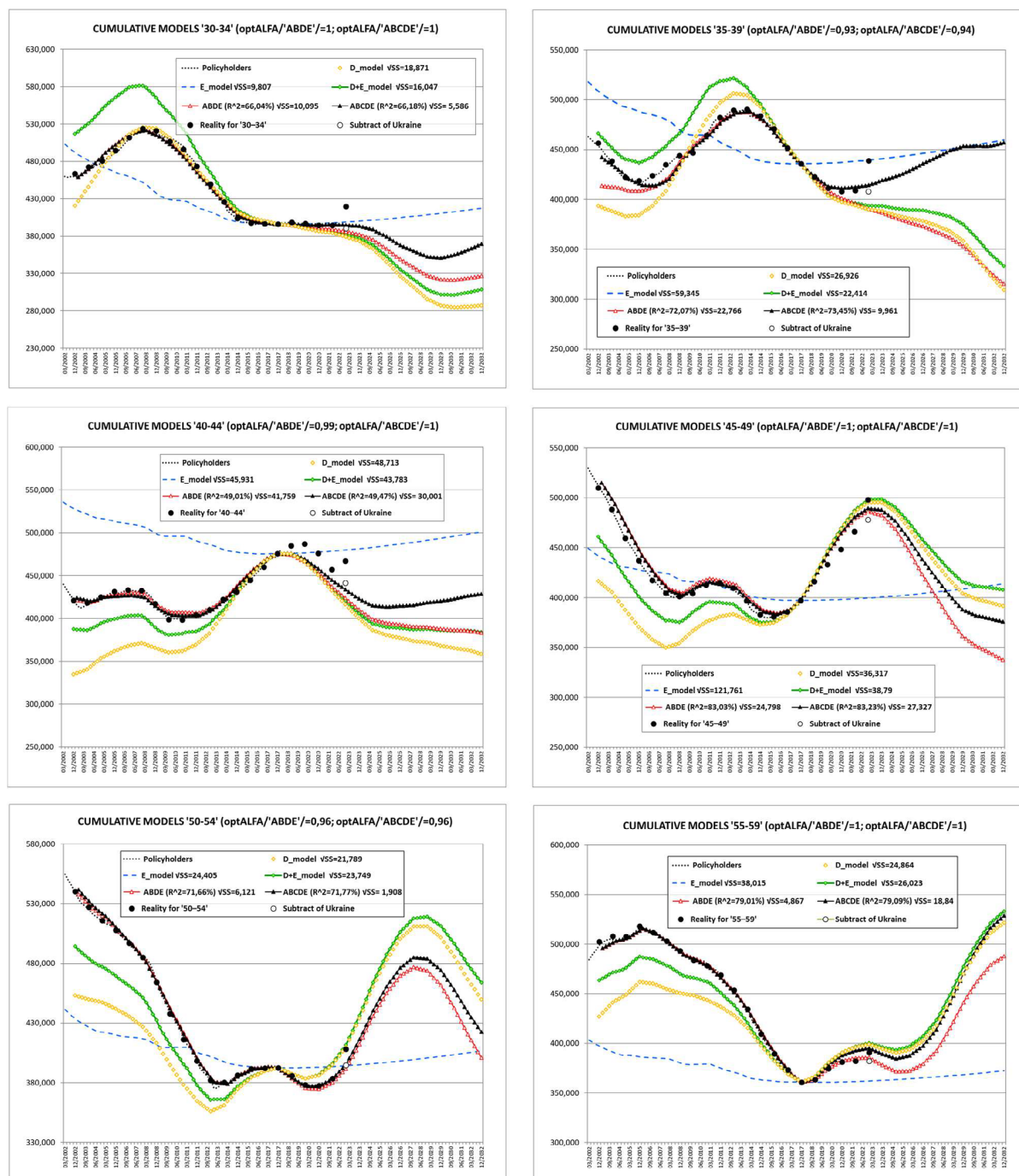


Fig. 4: a-f): Quarterly time series of the number of VZP CR insured people from 2002-17 in groups '30-34', ... , '55-59', their prospective and retrospective projections using three deterministic and two discounted regression models calibrated in 2017 in comparison with the real evolution in 2018-22 (see Legends in-side graphs)

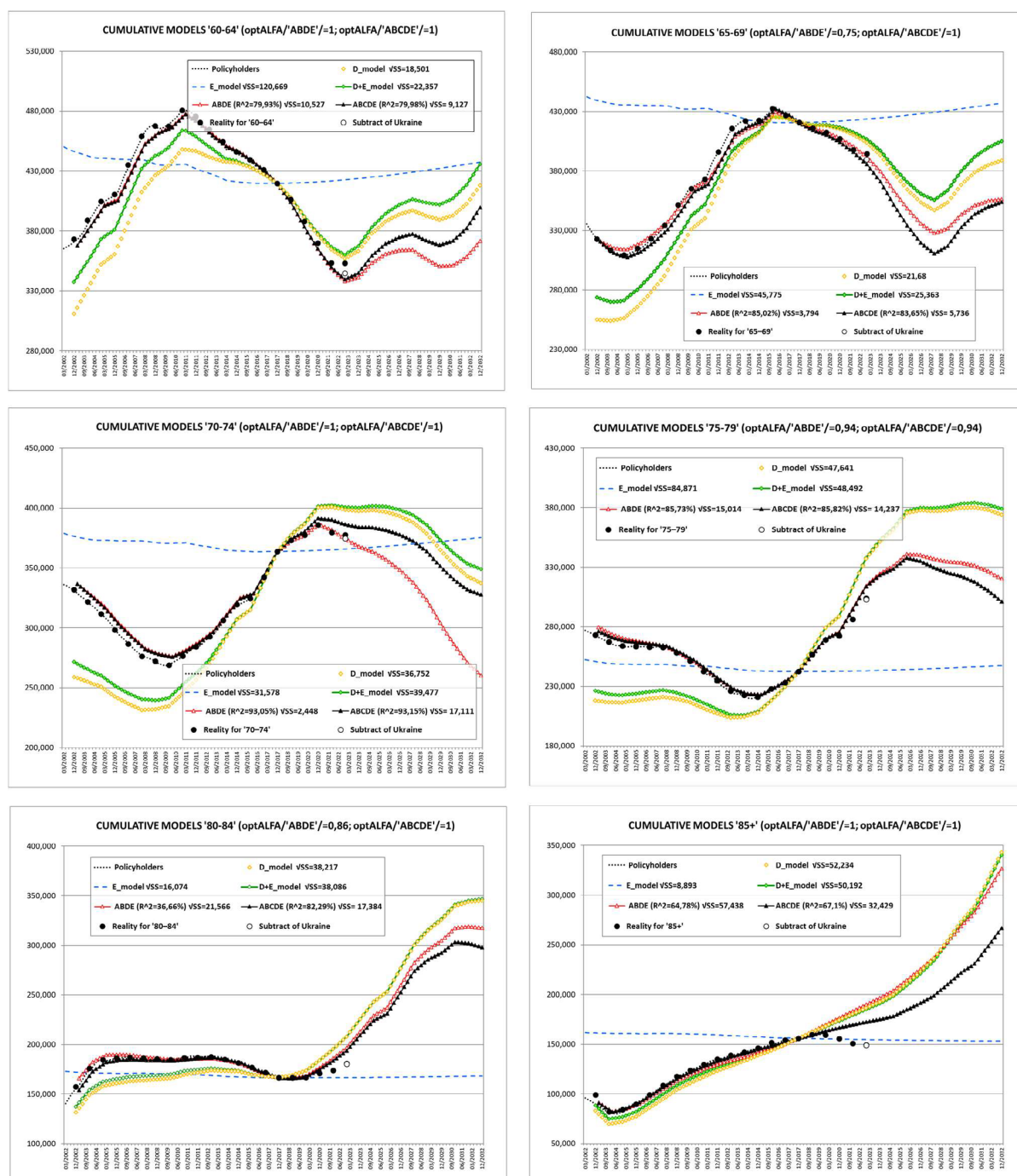


Fig. 5: a-f): Quarterly time series of the number of VZP CR insured people from 2002-17 in groups '60-65', ... , '85+', their prospective and retrospective projections from 2018 using three deterministic and two dis-counted regression models calibrated in 2017 in comparison with the real evolution in 2018-22 (see Leg-ends inside graphs)

However, we should be careful when interpreting these findings, because in some age groups, the explanatory role of formal polynomial terms is taken over by the currently steeper demographic or re-registration component inside some models. Overall, the higher level of statistical significance of the parameters in these models is proportional to the higher values of the coefficients of determination $R^2(\alpha x)$ shown in the Table 1.

The similarly structured Table 2 points to the significantly different results of the optimization of the 'ABDE' and 'ABCDE' models based on unweighted in-sample minimization in the last five years before the threshold period. It is not so much about how the minimum alpha values from Table 2 specifically differ from the previous optimum ones (in the minimizations ad (6) the choice $\alpha x > 1.00$ was also accepted), nor that the corresponding $R^2(\alpha x)$ values are generally lower than in Table 1 – this is a natural consequence of the fact that the models in Table 2 are parameterized

Tab. 2 Table of minimum discounts αx calibrated for unweighted LSE estimators αx (using (6) for 2012-16 annual data) for models 'ABDE' and 'ABCDE', their corresponding $R^2(\min. \alpha x)$ values and resulting $\sqrt{SS_{2018-22}}$ deviations from reality in 2018-22 (in thousands of people)

model:	'ABDE'	'ABDE'	'ABDE'	'ABCDE'	'ABCDE'	'ABCDE'
group x	$\min. \alpha x$	$R^2(\min. \alpha x)$	$\sqrt{SS_{2018-22}}$	$\min. \alpha x$	$R^2(\min. \alpha x)$	$\sqrt{SS_{2018-22}}$
'0+'	0.92	14.79%	82.662	0.95	15.12%	59.185
'0-4'	0.88	26.62%	14.618	1.05	43.77%	6.566
'5-9'	0.95	34.95%	72.639	0.90	29.36%	7.858
'10-14'	0.85	62.89%	19.176	0.85	75.50%	124.802
'15-19'	0.71	68.31%	3.414	0.89	76.44%	2.855
'20-24'	0.96	30.51%	7.433	0.98	36.40%	27.487
'25-29'	0.86	18.07%	40.537	0.83	24.81%	94.412
'30-34'	0.83	57.64%	11.822	0.89	60.01%	20.374
'35-39'	0.98	67.54%	21.672	0.92	72.95%	7.424
'40-44'	0.84	9.83%	86.747	0.89	27.96%	111.520
'45-49'	0.80	78.87%	30.436	1.03	85.06%	31.293
'50-54'	0.95	71.37%	5.574	0.96	71.77%	1.908
'55-59'	1.08	76.15%	5.091	0.99	77.73%	18.755
'60-64'	1.01	80.33%	11.323	0.82	17.23%	5.876
'65-69'	1.04	86.43%	23.133	1.04	86.60%	9.199
'70-74'	0.86	82.43%	7.087	1.04	92.28%	21.359
'75-79'	1.02	79.26%	19.329	1.07	75.69%	23.585
'80-84'	0.90	53.70%	26.010	1.00	81.83%	17.384
'85+'	0.61	0.34%	33.827	0.93	26.75%	19.216
'in total'	---	---	139.978	---	---	203.475

based on a different loss function than $R2(\cdot)$. What is now essential is to compare the corresponding values in the columns $\sqrt{SS_{2018-22}}$ for both models, which measure the formal distances of the model values from reality for each age group. Only for a few age groups did the in-sample optimization method using minimum alpha in the five prethreshold years perform significantly better results than the method originally used in 2018; in most cases, the second in-sample method is worse than the later observed reality. This is documented also by the cumulative deviations $\sqrt{SS_{2018-22}} = 139.978$ vs. 101.574 for the 'ABDE' model and $\sqrt{SS_{2018-22}} = 203.475$ vs. 95.930 for the 'ABCDE' model, at the expense of the second in-sample variant.

The decisive data for comparing the results of both discount optimization variants for the 'ABDE' and 'ABCDE' models is contained in Table 3 with the data compared for the final discounts αx , which were found ad poste as those minimizing the distances between real data and modeled

Tab. 3 Table of final discounts αx found due to step by step minimization of (7) for models 'ABDE' and 'ABCDE', their corresponding $R^2(fin. \alpha x)$ values and resulting $\sqrt{SS_{2018-22}}$ deviations from reality 2018-22 (in thousands of people)

model:	'ABDE'	'ABDE'	'ABDE'	'ABCDE'	'ABCDE'	'ABCDE'
group x	$fin. \alpha x$	$R^2(fin. \alpha x)$	$\sqrt{SS_{2018-22}}$	$fin. \alpha x$	$R^2(fin. \alpha x)$	$\sqrt{SS_{2018-22}}$
'0+'	0.82	16.42%	39.998	0.98	13.57%	49.291
'0-4'	0.76	26.22%	5.395	1.06	46.80%	6.511
'5-9'	0.54	35.52%	10.049	0.80	44.01%	5.456
'10-14'	0.80	56.40%	17.145	0.96	81.54%	7.196
'15-19'	0.66	61.32%	2.097	0.88	66.10%	2.646
'20-24'	0.45	9.73%	3.306	0.55	22.87%	8.736
'25-29'	0.92	33.40%	9.880	0.95	41.50%	12.913
'30-34'	0.97	62.62%	3.904	0.70	38.08%	3.140
'35-39'	0.71	41.05%	1.901	0.90	71.55%	6.380
'40-44'	0.69	14.74%	18.213	0.99	49.99%	20.448
'45-49'	1.01	83.78%	24.792	0.98	81.35%	26.800
'50-54'	0.87	58.66%	1.384	0.96	71.77%	1.908
'55-59'	0.64	55.33%	1.639	0.74	66.56%	7.906
'60-64'	0.90	45.33%	2.318	1.07	73.54%	3.048
'65-69'	0.66	82.68%	1.528	0.94	83.03%	1.779
'70-74'	1.09	84.87%	2.145	0.85	82.01%	2.114
'75-79'	0.95	85.60%	14.950	0.97	85.16%	13.878
'80-84'	1.00	81.83%	1.181	1.04	82.54%	5.173
'85+'	0.74	1.21%	15.627	0.60	2.38%	5.482
'in total'	---	---	44.590	---	---	43.514

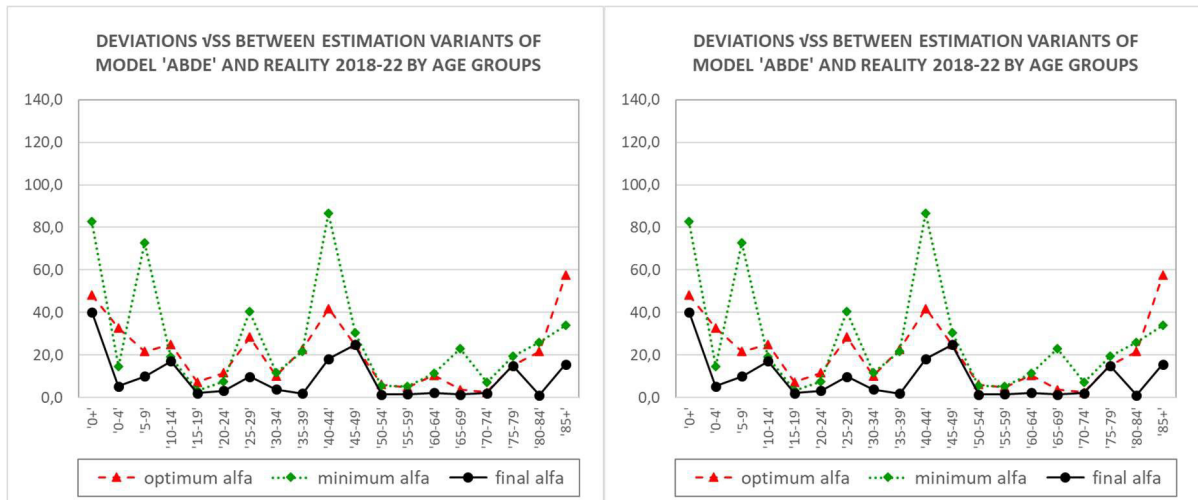


Fig. 6: a-b): Visual comparison of deviations between the real data 2018-22 against three calibration variants (see Legends in the charts) of model 'ABCD' or 'ABCDE', respectively, by age groups

values for the years 2018–22 (compare the 4th and 7th columns of Table 3). This means that there is no such αx in any age group x for which a better fit could be achieved in the given class of generalized-discounted regression models in 2018! In summary, the decisive data from Tables 1–3 are visualized in Figures 6a–b) for both models. The significant deviations from reality are noticeable in some age groups, especially for in-sample variant from Table 2, more clearly for 'ABCDE' model.

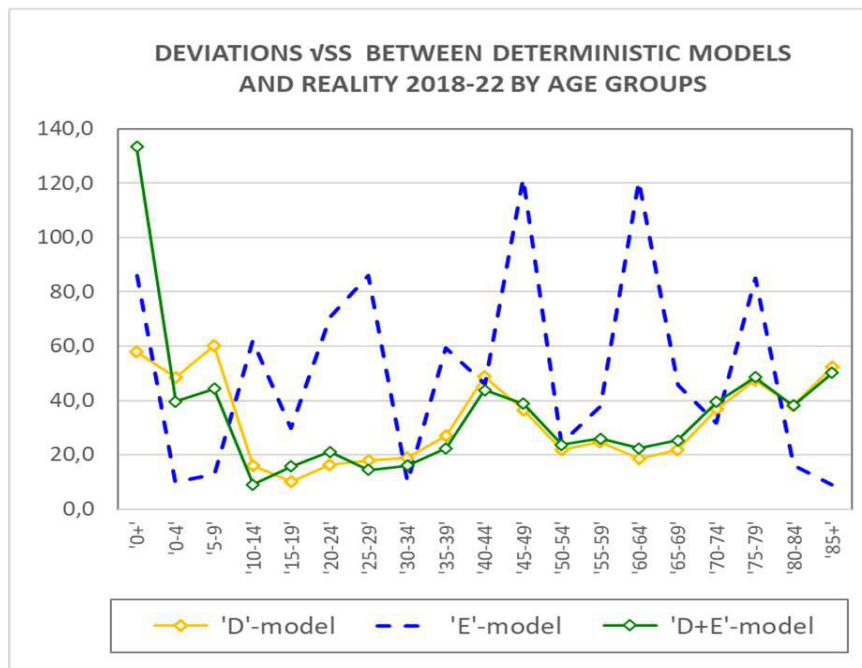


Fig. 7: Visual comparison of deviations between the real data 2018-22 against three deterministic models predictions (see Legend in the charts) by age groups)

For the variant from Table 1, the worst predictions were found in the groups '0+' (models with low smoothing and therefore predictive accuracy), '25-29' (we have found de facto inverse trend after 2017 here), '40-44' (underestimation of reality for both models) and '85+' (overestimation of reality for 'ABDE' model, where the linear trend component no longer had the ability to compensate for the losses of insured persons reflecting the COVID21 pandemic). In some age groups ('85+', '40-44' and '20-24'), we observed significant decreases in $R^2(\alpha x)$ relative to the 2018 model values with the transition to models with final discounts αx (see comparison of columns 3 and 6 in Tables 1 and 3). However, for most other age groups, the prediction based on optimum alpha's appears to be fairly successful (for some age groups, the predicted curves did not differ much from each other even with significant changes in alfa values).

Finally, we could be interested in comparing the prediction results with deterministic models from

Tab. 4 Table of resulting $\sqrt{SS_{2018-22}}$ deviations due to the deterministic model predictions toward the reality in 2018–2022 (in thousands of people)

model:	'D-model'	'D-model'	'D+E-model'
group x	$\sqrt{SS_{2018-22}}$	$\sqrt{SS_{2018-22}}$	$\sqrt{SS_{2018-22}}$
'0+'	57.89	85.99	133.42
'0-4'	48.52	10.01	39.67
'5-9'	60.04	12.85	44.32
'10-14'	16.10	61.88	9.07
'15-19'	10.04	29.92	15.65
'20-24'	16.38	70.85	21.16
'25-29'	17.91	86.05	14.49
'30-34'	18.87	9.81	16.05
'35-39'	26.93	59.35	22.41
'40-44'	48.71	45.93	43.78
'45-49'	36.32	121.76	38.79
'50-54'	21.79	24.41	23.75
'55-59'	24.86	38.02	26.02
'60-64'	18.50	120.67	22.36
'65-69'	21.68	45.78	25.36
'70-74'	36.75	31.58	39.48
'75-79'	47.64	84.87	48.49
'80-84'	38.22	16.07	38.09
'85+'	52.23	8.89	50.19
'in total'	146.273	255.347	137.724

(1). This purpose is served by Table 4, where the formal Euclidean distances of the actual numbers of VZP ČR insured persons as of 12/31/2018-2022 from the numbers predicted by the derived population projection ('D-model'), the middle variant of extrapolation of the VZP re-registration balance ('E-model') and the combination of both methods ('D+E-model') are available. Visualization of these distances in Figure 7 shows, that the applicability of the re-registration balance itself holds sporadically only in some age groups ('0-9', '30-34' and '85+'). The good predictive ability of both deterministic models with demo-projection suits in some younger age groups ('10-34'). Here, these deterministic models were competitive with predictions based on our best generalized regression models 'ABDE' and 'ABCDE', respectively. However, as the summary sums in the last row of Table 3 show, the deterministic models cannot be used universally in all age groups with better results. The indicators described above identify approximately those age groups where demographic or re-registration components were statistically most significant.

4 SUMMARY OF THE MAIN EMPIRICAL RESULTS

The main difficulty of the model developed in 2016 at VZP was the no transparent specification of the assignment from the beginning. If we were interested in predicting the future number of insured persons for the next few years, it would be sufficient to base ourselves on one-year time series from historical VZP yearbooks. However, relatively early after starting works on the model construction, a requirement emerged for predicting quarterly or at least semiannual balance sheets of insured persons, which may be important for the preparation of the so-called health insurance plans approved by VZP once a year (see Health Insurance Plans – VZP ČR, 2025). However, this requirement entailed the need to use data from VZP's internal databases and interventions to conceptualization of the entire prediction model, e.g. adjustment of the model with the aim of taking the seasonal component into account.

The specification of the projection horizon is another problem in itself. The pragmatic need for one-year plans of an insurance company can still be met relatively well by the method of simple indexing, where we compare the annual growth/decline indices in groups with a possible expert correction to the expected trends (common practice in VZP). But from the strategic point of view we should assume a horizon of at least 5 or even more years? The five-year period makes it possible to use the population forecast of the Czech Statistical Office for the Czech Republic available in five-year age groups. For situation, when VZP had approximately 55% share of the public health insurance market in the Czech Republic, this application has proven to be successful.

The natural inclusion of demographic and re-registration variables into models means a departure from methods of simple indexing and from a simple application of derived population projection, where predicted curves are simply prolonged by the results of the most current demo forecasts to the threshold year of prediction. In the applications used above, the population forecast and the extrapolated re-registration component became part of more complex regression models, where in particular the linear trend term in the time variable also takes into account the trends of the historical development of the insured population of VZP, which had been taken from consistent time series since 2002. However, the qualification of this trend to the future proved to be not easy to estimate inside the threshold years 2016–17, because at that time the total decrease in the number of insured persons in VZP began to visibly stagnate and it was not clear in which direction the current trends would further go on (see large fulfilled black circles between 2017–2022 on Figures 1–2 for an overall overview). In combination with carefully performed discounting within

five-year groups, it was mainly the population projection variable that made it possible to specify the realistic framework of the trends of the optimized extrapolation curves (see 'D-model' curves in Figures 3–5 for comparison with reality).

Of the two most important general regression models considered in this article, the 'ABDE' model, that one with only a linear term in the time variable, appears to be the more efficient. At comparable levels of the $R^2(\alpha x)$ coefficients shown in Table 1, it has the same or higher number of statistically significant parameters in most age groups than the 'ABCDE' model, that one with additional parabolic trend term! In terms of the best fit to historical data, as is seemed from the comparison of $R^2(\alpha x)$ and $\sqrt{SS_{2018-22}}$ values in Table 3, the entire 'ABCDE' model also demonstrated serious smoothing ability in some age groups. However, for the conventional in-sample method in the last five pre-threshold years, it turned out (see the last column of Table 2 in the groups '10–14' and '40–44', for example) that only the formal application of the full model (with a non-zero parabolic term by the time variable) can be misleading for the purposes of realistic prediction of the number of insured persons. However, this is a known undesirable property of prediction models with an uncorrected quadratic term by the time variable when applied to a longer time horizon.

The model for predicting the number of health insured people in five-year groups can be easily recalculated for both sexes. However, when applying the 2018 methodology to special groups of health insurance payers (employees, self-employed, other payers), the models used would have to be calibrated completely separately. Then, high-quality regression models could also serve as a basis for creating annual health plans of the insurance company, if they were calibrated and recalculated at least once a year based on consistently supplemented internal data. Then, we would record more or less continuous changes in the parameters of the fitting models, including their discounts αx , and in groups with lower achieved levels of $R^2(\alpha x)$, we could make more predictable corrections. And if we had systematic data on evolution of average incomes and average payments for the relevant groups of health insurance payers, we could also bring calculations of future total incomes and expenditures of the health insurance company under greater control in the medium horizon (say, for the next 3–5 years). Unexpected sudden changes in the number of insured persons (as was the case from 2022, for example, in connection with immigrants from Ukraine) can be easily corrected in these models at the level of input demographic or re-registration exogenous regressors.

5 CONCLUSIONS AND RECOMMENDATION

The prediction method of discounted linear regression models, when the calibration of discounts α is based on the maximum values of the coefficient of determination $R^2(\alpha)$, can be universally recommended based on the above experiences. The prerequisite for success is of course a sufficiently long history of the predicted time series (although when using discounts $\alpha \ll 1$ we could get by with an order of magnitude shorter). Furthermore, we must have a representative regression model with $R^2(\alpha)$, values not lower than (say) 35% and reasonably predictable time series of exogenous regression variables (that is, if we are not limited only to formal polynomial trends as in standard exponential smoothing).

If one of the significant explanatory components of the model is a demographic variable, population forecasts or projections can be used in justified cases. The auxiliary demo-prognosis can be effectively used even if it later turns out to be overestimated or underestimated compared to reality – the estimated level of the corresponding regression coefficient is optimized in the context

of a multiparametric regression model, which itself corrects trend inaccuracies in the demographic component. In any case, this approach is more universal than performing a formally simple derived population projection, which implicitly assumes that there is a constant trend in the data. However, the methodology used is unsuitable for time series located near the threshold year in the phase of a stagnant trend (as was the case in 2017 in aggregated age group ,0+'). A constant trend in the time series excludes the possibility of using a non-constant explanatory variable for smoothing the model, and therefore the values of discounted $R^2(\alpha)$, which should be used when optimizing more complex regression models, will themselves be close to zero and therefore will not lead to a more effective parameterization of the prediction model.

The parameter estimation methodology based on a linear difference model of type (2), which, thanks to the additional calculation ad (5), allows placing the predicted curve directly into the last known value of the predicted time series in the threshold year, can also be recommended due to the efficiency of fitting the extrapolation curve – model curves fitted directly into primary data do not have this property. Parameter estimates in the difference model according to formula (4) are essentially just a trivial modification of the standard unweighted least squares method, when we multiply the experimental design matrix and the column of the explained regression variable in the basic model by the same weighting factors $wt = \alpha x - t/2$ ($t = 1, \dots, t_0$) and perform the additional calculations described in paragraph 2.2. This simple algorithm works universally for any system of linear regression equations and is an alternation of formulas derived explicitly for exponential smoothing of polynomial regressions – given e.g. in (Cipra, 1986, pp. 57–72) or in (Exponential Smoothing, 2025). In contrast to those references, formula ad (4) allows to estimate parameters also for the choice of discount rates $\alpha > 1$, which may be also a justified choice in specific cases, as shown above.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, this is co-financed from Operational Programme Research, Development and Education.

REFERENCES

- BĚLOHRADSKÝ, A., ŠOLC, Z. 2018. Predikce příjmů veřejného zdravotního pojištění. In: Metodické compendium. Praha, Ministerstvo financí ČR. <https://www.mfcr.cz/cs/ministerstvo/kariera-a-vzdelavani/vzdelavani/odborne-studie-a-vyzkumy/2018/predikce-prijmu-verejneho-zdravotniho-po-33111>
- BĚLÁČEK, J. 2018. Predikce budoucích počtů pojištěnců VZP ČR – data, metodika a výsledky. In: Sborník příspěvků MEDSOFT 2018, pp. 7-19. Hotel Academic, Roztoky u Prahy, 3/20-21/2018. https://doi.org/10.35191/medsoft_2018_1_30_5_11
- BĚLÁČEK, J. 2019. Regresní model pro predikci budoucích počtů pojištěnců VZP ČR. In: Sborník příspěvků MEDSOFT 2019, pp. 4-8. Hotel Academic, Roztoky u Prahy, 3/19-20/2019. https://doi.org/10.35191/medsoft_2019_1_31_4_8
- BĚLÁČEK, J. 2023. Predikce budoucích počtů pojištěnců VZP ČR – ukázka modelů a porovnání výsledků. Paper presented on ČDS conference 2023, Hradec Králové, 5/26/2023. <https://www.czechdemography.cz/akce/konference/konference-cds-2023/prispevky-z-konference/>
- CIPRA, T. 1986. Analýza časových řad s aplikacemi v ekonomii. Praha, SNTL/Alfa.

- CZECH STATISTICAL OFFICE. 2013. *Population projection of the Czech Republic until 2100* [in Czech]. Czech statistical office. <https://www.czso.cz/csu/czso/projekce-obyvatelstva-ceske-republiky-do-roku-2100-n-fu4s64b8h4>
- CZECH STATISTICAL OFFICE. 2014. Population projection in the regions of the Czech Republic - up to the 2050 [in Czech]. Czech statistical office. <https://www.czso.cz/csu/czso/projekce-obyvatelstva-v-krajich-cr-do-roku-2050-ua08v25hx9>
- CZECH STATISTICAL OFFICE. 2023. Demographic Yearbook of the Czech Republic – 2022. Czech statistical office. <https://csu.gov.cz/produkty/demographic-yearbook-of-the-czech-republic-2022>
- CZECH STATISTICAL OFFICE. 2026. Age Distribution of the Population - 2012, ..., 2016. 2017. Czech statistical office. <https://csu.gov.cz/produkty/age-distribution-of-the-population-2016>
- VZP ČR. 2023. VZP ČR Yearbook for 2018-22 [in Czech]. www.vzp.cz/o-nas/dokumenty/rocenky
- VZP ČR. 2025. Health insurance plans. [in Czech] <https://www.vzp.cz/o-nas/dokumenty/zdravotne-pojistne-plany>
- WIKIPEDIA. 2025. Exponential smoothing. https://en.wikipedia.org/wiki/Exponential_smoothing

Electronic documentation:

Excel application of a predictive model. 2025. Available at <https://zenodo.org/uploads/16986175>

Contact information

Jaromír Běláček: e-mail: belacek@mendelu.cz