APPLYING STATE SPACE MODELS TO STOCHASTIC CLAIMS RESERVING

BY

RADEK HENDRYCH AND TOMAS CIPRA

ABSTRACT

The paper solves the loss reserving problem using Kalman recursions in linear statespace models. In particular, if one orders claims data from run-off triangles to time series with missing observations, then state space formulation can be applied for projections or interpolations of IBNR (*Incurred But Not Reported*) reserves. Namely, outputs of the corresponding Kalman recursion algorithms for missing or future observations can be taken as the IBNR projections. In particular, by means of such recursive procedures one can perform effectively simulations in order to estimate numerically the distribution of IBNR claims which may be very useful in terms of setting and/or monitoring of prudency level of loss reserves. Moreover, one can generalize this approach to the multivariate case of several dependent run-off triangles for correlated business lines and the outliers in claims data can be also treated effectively in this way. Results of a numerical study for several sets of claims data (univariate and multivariate ones) are presented.

KEYWORDS

Dependent run-off triangles, IBNR projections; missing observations, state space model; stochastic claims reserving.

JEL codes: C22, C32, G22

1. INTRODUCTION

Over last 40 years, numerous loss reserving methods in general insurance have been developed. Even though a substantial part of them are stochastic ones in terms of being based on stochastic models, only a limited number of them explicitly employ time series models. It is a surprising fact since claim figures might be represented as a sequence of discrete time data showing trend and

Astin Bulletin 51(1), 267-301. doi:10.1017/asb.2020.38

TABLE 1

RUN-OFF TRIANGLE OF OBSERVED INCREMENTAL CLAIMS AMOUNTS X_{ij} (THE MISSING OBSERVATIONS UNDER THE LAST OBSERVED DIAGONAL ARE IN BRACKETS).

Accident	Development year j							
year i	0	1	2	3	$\cdots$	$s-2$	$s-1$	s
0	$X_{0,0}$	$X_{0,1}$	$X_{0,2}$	$X_{0,3}$	$\cdots$	$X_{0,s-2}$	$X_{0,s-1}$	$X_{0,s}$
1	$X_{1,0}$	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$\cdots$	$X_{1,s-2}$	$X_{1,s-1}$	$(X_{1,s})$
2	$X_{2,0}$	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$\cdots$	$X_{2,s-2}$	$(X_{2,s-1})$	$(X_{2,s})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	$X_{s-1,0}$	$X_{s-1,1}$	$(X_{s-1,2})$	$(X_{s-1,3})$	$\cdots$	$(X_{s-1,s-2})$	$(X_{s-1,s-1})$	$(X_{s-1,s})$
s	$X_{s,0}$	$(X_{s,1})$	$(X_{s,2})$	$(X_{s,3})$	$\cdots$	$(X_{s,s-2})$	$(X_{s,s-1})$	$(X_{s,s})$

seasonal features. Moreover, the time series approach consisting in state space modeling enables to apply the Kalman recursions and to construct projections of IBNR (*Incurred But Not Reported*) loss reserves (including the corresponding mean squared errors) by means of filtering and smoothing algorithms within this methodology. In other words, one may look upon projections of IBNR reserves as if one interpolates missing values and predicts future values in time series. Moreover, it is possible to treat several claim figures of dependent business lines jointly in a multivariate time series of a fixed dimension taking the correlations among various triangles into account.

The usual organization of loss reserves data is based on so-called run-off triangles in the double-index format (see Table 1). The rows of such a triangle represent the accident (or origin) years and its columns the development years (its diagonals correspond to the calendar years). The value X_{ij} ($i = 0, 1, \ldots, s$, $j = 0, 1, \ldots, s-i$) denotes the incremental payment in the development year j for a loss event occurred in the accident year i (one assumes that all claims are settled after the development year s, see, e.g., Cipra (2010)). One observes payments till the calendar year corresponding to the positions of the last diagonal elements (i.e., $i + j = s$) and projects all the missing payments below the diagonal of the given run-off triangle. Then, the total sum of these projected (missing) values represents the total estimated loss reserve necessary to fulfill insurance obligations after the calendar year s. Note that some IBNR reserve estimators, for example, the chain ladder method (CL), use the cumulative form $C_{ij} = X_{i0} + \cdots + X_{ij}$ or the logarithmic form $Y_{ij} = \log X_{ij}$ assuming implicitly that $X_{ij} > 0$ for all i and j.

The state space modeling of claims development data requires that the elements of triangle are ordered to a time series. The orderings can be different. One can form a time series of subdiagonals corresponding to particular calendar years (such a time series is multivariatewith the dimension changing in time, see, e.g., de Jong and Zehnwirth (1983), Verrall (1989)). Alternatively, one can use the row-wise ordering and construct a univariate time series with

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 2

RUN-OFF TRIANGLE OF OBSERVED LOGARITHMIC INCREMENTAL CLAIMS AMOUNTS Y_{ij} ORDERED ROW-WISE TO A TIME SERIES $\{y_t\}$ (THE VALUES IN THE FIRST COLUMN ARE TAKEN AS THE INITIAL ONES, THE MISSING OBSERVATIONS UNDER THE LAST OBSERVED DIAGONAL ARE IN BRACKETS).

Accident	Development year j							
year i	0	1	2	3	$\cdots$	$s-2$	$s-1$	s
0	$Y_{0,0}$	y_1	y_2	y_3	$\cdots$	y_{s-2}	y_{s-1}	y_s
1	$Y_{1,0}$	y_{s+1}	y_{s+2}	y_{s+3}	$\cdots$	y_{2s-2}	y_{2s-1}	(y_{2s})
2	$Y_{2,0}$	y_{2s+1}	y_{2s+2}	y_{2s+3}	$\cdots$	y_{3s-2}	(y_{3s-1})	(y_{3s})
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	$Y_{s-1,0}$	$y_{(s-1).s+1}$	$(y_{(s-1).s+2})$	$(y_{(s-1).s+3})$	$\cdots$	$(y_{s.s-2})$	$(y_{s.s-1})$	$(y_{s.s})$
s	$Y_{s,0}$	$(y_{s.s+1})$	$(y_{s.s+2})$	$(y_{s.s+3})$	$\cdots$	$(y_{(s+1).s-2})$	$(y_{(s+1).s-1})$	$(y_{(s+1).s})$

missing observations, see, for example, Atherino *et al.* (2010). The row-wise ordering is also preferred in this paper (see, e.g., Table 2).

As far as the principle of the state space model (SSM) is concerned, one can apply the structural approach where the model is structured to the level, trend and periodic (mostly seasonal) components, see Atherino *et al.* (2010), Alpuim and Ribeiro (2003). The usual shape of the run-off observed from the incremental claims can be instead modeled as a Hoerl's curve, which increases rapidly to a peak and then dies off exponentially, see, for example, de Jong and Zehnwirth (1983), England and Verrall (2002), Pang and He (2012), Taylor *et al.* (2003). Another possibility comprises the log-normal models (denoted by Kremer (1982) as an ANOVA approach), where the logarithms of incremental values are supposed to be normal andtheir mean values are described by a parameter for each row i and a parameter for each column j, see, for example, England and Verrall (2002), Kloek (1998), Li (2006), Ntzoufras and Dellaportas (2002), Renshaw (1989), Verrall (1989, 1991, 1994), Wüthrich and Merz (2008), and Zehnwirth (1996). Alternatives for state space modeling loss reserves appear in the literature, for example, smoothing (nonparametric) approach, see de Jong (2005). This paper has been inspired by Atherino *et al.* (2010) in terms of using the row-wise ordering of the triangle elements to the time series $\{y_t\}$ with missing observations. However, we have found that this approach works well only for incremental data of a special character and not universally. Moreover, the substantial part of the loss reserving technique advocated by Atherino *et al.* (2010) consists in (*ad hoc*) identification of outliers in the investigated claims data which improve the outputs of the method to a considerable extent.

The aim of this paper is to investigate and select space models that adequately encompass most of the recursive features of loss reserving. There is a lot of possible alternatives combining various features of models and underlying data (e.g., one must fix on incremental or cumulative data in logarithmic

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

or non-logarithmic form facing the problem of impossible logarithmic transformation of negative incremental data). We have tested tens of various SSMs by means of various data sets (including the multivariate ones for dependent run-off triangles) which appear in the actuarial references. The models should enable effective recursive realizations of extensive simulations including a detailed residual analysis and estimating the loss reserve distribution even for multivariate run-off triangles.

The paper is organized as follows: Section 2 reviews the general form of linear Gaussian SSMs and the role of Kalman recursions when adopting these dynamic models. Section 3 presents several types of SSMs that are examined in this paper to receive suitable IBNR projections over all data sets from the Appendix under the criteria of simplicity and straightforward applicability. Due to fast recursive algorithms, they allow estimating numerically the distribution of IBNR reserves by means of simulations. Section 4 generalizes these models to the multivariate case of dependent run-off triangles for correlated business lines. Numerical results including technical details of the corresponding Kalman recursions and an extensive residual analysis are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Linear Gaussian SSMs and Kalman recursions

State space modeling is a very efficient technique to handle dynamic stochastic systems in a flexible way. In the context of claims reserving, it is sufficient to apply its linear Gaussian version in discrete time indexed by t:

$$y_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N(0, H_t), \tag{2.1}$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim \text{independent } N(0, Q_t), \tag{2.2}$$

where α_t is a vector $(m\times 1)$ of a latent state process identified by an observation vector $y_t(p\times 1)$ through the observation Equation (2.1). The development of α_t in time is described by means of the state Equation (2.2). Both equations are stochastic with mutually independent residuals $\varepsilon_t(m\times 1)$ and $\eta_t(k\times 1)$, which are also independent of $\alpha_1 \sim N(a_1, P_1)$. Some of matrices $Z_t(p\times m)$, $T_t(m\times m)$, $R_t(m\times k)$, $H_t(p\times p)$, and $Q_t(k\times k)$ are frequently time-invariant, but they may contain unknown parameters which need to be estimated. The dimensions of vectors α_t and y_t can be time-varying, which in such case should read as $(m_t\times 1)$ and $(p_t\times 1)$, respectively. The system of Equations (2.1) and (2.2) is usually called the dynamic linear model (DLM).

The SSM for claims data suggested by Atherino *et al.* (2017) has the form:

$$y_t = \mu_t + \gamma_t + h_t'u + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N\left(0, \sigma_\varepsilon^2\right), \tag{2.3}$$

$$\mu_{t+1} = \mu_t + \eta_t^\mu, \quad \eta_t^\mu \sim \text{independent } N\left(0, \sigma_\mu^2\right), \tag{2.4}$$

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \eta_t^\gamma, \quad \eta_t^\gamma \sim \text{independent } N\left(0, \sigma_\gamma^2\right) \tag{2.5}$$

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

with mutually independent white noises ε_t, η_t^{μ}, and η_t^{γ}. The model (2.3)–(2.5) is motivated by the behavior of claims process and is structured to the level component μ responding to the average claims values along each accident year, the periodic component γ capturing the development year effect, and the regression term $h_t'u$ responding to the intervention effect due to the presence of outliers (the denotation of *structural* time series models is just due to this structure).

The main objective of state space modeling is to obtain knowledge of the latent states α_t given the observations y_t. This is achieved using the Kalman recursive algorithms for filtering and smoothing, see, for example, Brockwell and Davis (2016), Durbin and Koopman (2012), Hamilton (1994), Harvey (1989), Johannssen (2016), and Shumway and Stoffer (2017). For instance, by means of the Kalman filtering algorithm, one obtains the one-step-ahead predictions and prediction errors:

$$a_{t+1|t} = \mathrm{E}(\alpha_{t+1}|y_t, \ldots, y_1), \quad e_{t+1|t} = y_{t+1} - Z_{t+1}a_{t+1|t}, \tag{2.6}$$

and the related covariance matrices:

$$P_{t+1|t} = \mathrm{Var}(\alpha_{t+1}|y_t, \ldots, y_1), \quad F_{t+1|t} = \mathrm{Var}(e_{t+1|t}) = Z_{t+1}P_{t+1|t}Z'_{t+1} + H_{t+1}. \tag{2.7}$$

The Kalman smoothing algorithm provides the smoothed values of the type:

$$a_{t|n} = \mathrm{E}(\alpha_t|y_n, \ldots, y_1), \quad P_{t|n} = \mathrm{Var}(\alpha_t|y_n, \ldots, y_1) \text{ for } t = n, \ldots, 1. \tag{2.8}$$

Moreover, the Kalman recursions allow interpolating the missing observations in the giventime series $\{y_t\}$ which is important just in our context of loss reserving. Various technicalities must be properly treated, for example, initial conditions of Kalman algorithms, estimation of unknown model parameters (usually, the classic (quasi-)maximum likelihood estimation (QMLE) or the expectation–maximization (EM) algorithm, see, e.g., Durbin and Koopman (2012) or Shumway and Stoffer (1982)) and others. Even though in general the unknown parameters of DLM can change in time, we suppose that such parameters are constant to avoid (numerical) problems with their estimation. These problems may be significant when the number of parameters is too high (moreover from the point of view of time series analysis, nearly one-half of observations are missing). In particular, the model (2.3)–(2.5) is formulated as homoscedastic with constant unknown (positive) variances σ_ε, σ_μ, and σ_γ. There are easily available software systems for the Kalman recursions, for example, the package KFAS (*Kalman filtering and smoothing*) in software system R (see, e.g., Helske (2017)) or STAMP (*Structural Time Series Analyser, Modeller and Predictor*, see Koopman *et al.* (2009)).

3. Univariate SSMs for claims data

In this Section, we present three types of linear Gaussian SSMs for the univariate run-off triangles. Despite their simplicity, they provide satisfactory results which are comparable withthe ones by more sophisticated and complex models. Under certain restrictions, they seem to work well for majority of routine actuarial data. Moreover, due to effective recursive algorithms, they enable to estimate comfortably the distribution of IBNR reserves by extensive simulations.

3.1. Log-normal SSM (*i*)

The first model is of the log-normal type applying the ANOVA principle to the logarithmic incremental data $Y_{ij} = \log X_{ij}$ (see Section 2). Log-normal models applied to claims reserving can be interpreted either as smoothing models or generalized linear models (GLMs), see Björkwall *et al.* (2011). The usual assumptions are that:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{independent } N(0, \sigma_\varepsilon^2), \tag{3.1}$$

that is, $X_{ij} \sim LN(\mu_{ij}, \sigma_\varepsilon^2)$ with

$$\mu_{ij} = c + \alpha_i + \beta_j, \tag{3.2}$$

where the parameters α_i and β_j model the row and column effects, respectively (one usually chooses $\alpha_0 = \beta_0 = 0$ for the sake of identification of parameters). More generally, the theory of GLMs can be applied in this context, see Merz and Wüthrich (2008) and others.

In the case of log-normal SSM, let us rewrite Table 1 into the form of Table 2 stressing the fact that the values in the first column ($Y_{i0}, i = 0, 1, \ldots, s$) are taken as the initial levels in the observation Equation (3.3). This adjustment plays a positive role when one initializes recursive algorithms over short data segments since one sets up the initial level correctly, particularly in the situation with missing observations (e.g., in the last row one observes only the first value and the remaining ones are missing). Moreover, the SSM assumes a stochastic behavior of the row effect (see, e.g., the state Equation (3.4)). As far as the problem of transformation of negative incremental values is concerned, we replace the negative values in the first row by the arithmetic mean of neighboring values in the same row. These interpolations link only to the negative values in the first row, while the others for $i > 0$ are canceled and taken as additional missing values for the Kalman recursions. The time series $\{y_t\}$ of length $(s+1) \cdot s$ contains in total at least $(s+1) \cdot s/2$ (i.e., 50%) missing observations; there can be more missing observations since in the models described in Sections 3.1 and 3.2 with logarithmic transformations the negative incremental values outside the first row are regarded as missing ones. The suggested interpolation for negative values in the first row might be seen as heuristic at first sight, and we have tried to avoid it by setting the particular entry in the first row as missing (in order to be interpolated together with other missing values in the given time series using the Kalman recursions). However, it could lead to numerical instabilities, or even degenerations in the Kalman recursions (e.g., when employing the diffuse initialization, the diffuse part might not end). Point out that this interpolation needs to be applied very rarely (in fact only in one run-off triangle denoted as 4_2, see Section 5).

The corresponding linear Gaussian SSM(with mutually independent residuals ε and η) can be written either in the double-index format for $i = 0, \ldots, s$ and $j = 1, \ldots, s$:

$$Y_{ij} = Y_{i0} + \beta_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{independent } N\left(0, \sigma_{\varepsilon}^2\right), \tag{3.3}$$

$$\beta_{ij} = \beta_{i-1,j} + \eta_{ij}, \quad \eta_{ij} \sim \text{independent } N\left(0, \sigma_{\beta}^2\right), \tag{3.4}$$

or in the time format using the row-ordering for $t = i \cdot s + j$:

$$y_t - Y_{i0} = \beta_t + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N\left(0, \sigma_{\varepsilon}^2\right), \tag{3.5}$$

$$\beta_{t+1} = \beta_{t-s+1} + \eta_t, \quad \eta_t \sim \text{independent } N\left(0, \sigma_{\beta}^2\right), \tag{3.6}$$

or as the DLM in (2.1)–(2.2) with the state vector $\vartheta_t = (\beta_t, \beta_{t-1}, \ldots, \beta_{t-s+1})'$ for $t = 1, 2, \ldots$:

$$y_t - Y_{i0} = (1,\ 0,\ \ldots,\ 0)\vartheta_t + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N\left(0, \sigma_{\varepsilon}^2\right), \tag{3.7}$$

$$\vartheta_{t+1} = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \vartheta_t + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \eta_t, \quad \eta_t \sim \text{independent } N(0, \sigma_{\beta}^2). \tag{3.8}$$

The vector (1, 0, …, 0) corresponds to the matrix Z_t in the observation Equation (2.1), the matrix of dimension ($s \times s$) in (3.8) corresponds to the matrix T_t in the state Equation (2.2), etc. In particular, the covariance matrix of the residual vector $R_t\eta_t$ in (3.8) with $R_t = (1, 0, \cdots, 0)'$ has the form:

$$\text{Var}(R_t\eta_t) = \text{Var}\begin{pmatrix} \eta_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sigma_{\beta}^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}. \tag{3.9}$$

There are only two unknown parameters σ_{ε} and σ_{β} in this DLM. As far as the initial state vector $\hat{\vartheta}_1 = (\hat{\beta}_1, \hat{\beta}_0, \ldots, \hat{\beta}_{-s+2})'$ is concerned, a diffuse initial state vector is employed (i.e., $\vartheta_1 \sim N(0, P_1)$ with $P_1 = \kappa \cdot I$, where I denotes the identity matrix of corresponding dimension and $\kappa >> 0$). The diffuse initialization was also considered but, in many cases, the diffuse part of Kalman recursion algorithm did not converge due to a high number of missing entries comparing to the number of non-missing inputs and estimated parameters (and when it converged the outcomes were very proximate to the ones gathered by the diffuse initial state vector). In contrast to the log-normal model (3.1)–(3.2) that assumes mutually independent rows of the corresponding run-off triangle, the state space equation of the type (3.4) enables to model the stochastic trend for the development components in particular columns. Moreover, we included a technical simplification reducing the number of unknown parameters (namely the row parameters α_i, see (3.2)) by the observed values Y_{i0}, since it avoids the problem of the parameter restriction for the model to be identified without impairing numerical outputs in routine actuarial data sets.

3.2. Log-normal SSM with Hoerl's curve modification (*ii*)

The second model amends the previous one described in Section 3.1 by Hoerl's curve (see, e.g., (18) or (21)). When modeling the trend over development years by this curve (its shape is similar to the run-off behavior observed in incremental claims), the corresponding smoothing effect can improve the IBNR prediction for volatile incremental data (on the other hand, it is not realistic that Hoerl's curve remains a good fit over all development years due to its quick decrease over more remote development years). The problem consisting in the logarithmic transformation of negative incremental values is solved in the same way as in the model discussed in Section 3.1.

Similarly, as in the model from Section 3.1 the corresponding linear Gaussian SSM (with mutually independent residuals ε and η) can be again written either in the double-index format for $i = 0, \ldots, s$ and $j = 1, \ldots, s$:

$$Y_{ij} = Y_{i0} + \beta_i \log(j+1) + \gamma_i \cdot j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{independent } N\big(0, \sigma_\varepsilon^2\big), \tag{3.10}$$

$$\beta_i = \beta_{i-1} + \eta_i^\beta, \quad \eta_i^\beta \sim \text{independent } N\big(0, \sigma_\beta^2\big), \tag{3.11}$$

$$\gamma_i = \gamma_{i-1} + \eta_i^\gamma, \quad \eta_i^\gamma \sim \text{independent } N\big(0, \sigma_\gamma^2\big) \tag{3.12}$$

or in the time format using the row-ordering for $t = i \cdot s + j$:

$$y_t - Y_{i0} = \beta_t \log(j+1) + \gamma_t \cdot j + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N\big(0, \sigma_\varepsilon^2\big), \tag{3.13}$$

$$\beta_{t+1} = \beta_{t-s+1} + \eta_t^\beta, \quad \eta_t^\beta \sim \text{independent } N\big(0, \sigma_\beta^2\big), \tag{3.14}$$

$$\gamma_{t+1} = \gamma_{t-s+1} + \eta_t^\gamma, \quad \eta_t^\gamma \sim \text{independent } N\big(0, \sigma_\gamma^2\big) \tag{3.15}$$

or as the DLM with the state vector $\vartheta_t = (\beta_t, \beta_{t-1}, \ldots, \beta_{t-s+1}, \gamma_t, \gamma_{t-1}, \ldots, \gamma_{t-s+1})'$ for $t = 1, 2, \ldots$:

$$y_t - Y_{i0} = (\log(j+1),\ 0,\ \ldots,\ 0, j,\ 0,\ \ldots,\ 0)\vartheta_t + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N(0, \sigma_\varepsilon^2), \tag{3.16}$$

$$\vartheta_{t+1} = \begin{pmatrix} \begin{matrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{matrix} & \\ & \begin{matrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{matrix} \end{pmatrix} \vartheta_t + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_t^{\beta} \\ \eta_t^{\gamma} \end{pmatrix},$$

$$\eta_t^{\beta}, \eta_t^{\gamma} \sim \text{independent } N(0, \sigma_\beta^2) \text{ and } N(0, \sigma_\gamma^2). \tag{3.17}$$

In particular, one has

$$Q = \text{Var}\left((\eta_t^{\beta}, \eta_t^{\gamma})'\right) = \begin{pmatrix} \sigma_\beta^2 & 0 \\ 0 & \sigma_\gamma^2 \end{pmatrix} \tag{3.18}$$

so that there are three unknown parameters σ_ε, σ_β, and σ_γ in this model. More generally, the diagonal matrix in (3.18) can be non-diagonal covering the covariance between the estimated parameters β and γ by the form:

$$\begin{pmatrix} \sigma_\beta^2 & \sigma_{\beta\gamma} \\ \sigma_{\beta\gamma} & \sigma_\gamma^2 \end{pmatrix}. \tag{3.19}$$

3.3. Chain ladder SSM (*iii*)

The third SSM is based on CL principle for cumulative data $C_{ij} = X_{i0} + \cdots + X_{ij}$ ($i = 0, 1, \ldots, s$, $j = 0, 1, \ldots, s - i$) so that it avoids naturally the problem of the logarithmic transformation of negative incremental values. One can characterize it as the CL approach with dynamic changes of the development factors by the random walk mechanism (for the stochastic model underlying the CL technique see, e.g., Costa *et al.* (2016), Mack (1993), or Renshaw and Verrall (1998)).

The corresponding linear Gaussian SSM (with mutually independent residuals ε and η) can be written either in the double-index format for $i = 0, \ldots, s$ and $j = 1, \ldots, s$:

$$C_{ij}/C_{i,j-1} = \beta_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{independent } N\left(0, \sigma_\varepsilon^2\right), \tag{3.20}$$

$$\beta_{ij} = \beta_{i-1,j} + \eta_{ij}, \quad \eta_{ij} \sim \text{independent } N\left(0, \sigma_\beta^2\right) \tag{3.21}$$

or in the time format using the row-ordering for $t = i \cdot s + j$:

$$C_{ij}/C_{i,j-1} = \beta_t + \varepsilon_t, \quad \varepsilon_t \sim \text{independent } N\left(0, \sigma_\varepsilon^2\right), \tag{3.22}$$

$$\beta_{t+1} = \beta_{t-s+1} + \eta_t, \quad \eta_t \sim \text{independent } N\left(0, \sigma_\beta^2\right). \tag{3.23}$$

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

The DLM format is analogous to the one in the previous models and is omitted here. Having estimated the ratios of cumulated claims (i.e., the left-hand sides of (3.20) and (3.22)), the particular(estimated) cumulated claims can be calculated by sequential multiplying the ratio by its (estimated) denominator (given C_{i0}). Another variant of this approach relates all cumulative values in each row to the initial one in the given row so that the ratios $C_{ij}/C_{i,j-1}$ in (3.20) or (3.22) are replaced by C_{ij}/C_{i0} (see also Alpuim and Ribeiro (2003), where one assumes autoregressive relations among development factors in particular columns). In any case, there are only two unknown parameters σ_ε and σ_β in this model.

4. Multivariate SSMs for dependent run-off triangles

The SSMs from Section 3 can be easily generalized for dependent run-off triangles when dealing with correlated lines of business. A single multivariate model exploiting such correlations may improve the IBNR projections generated in the loss reserving processin insurance companies, see, for example, Braun (2004), de Jong (2006, 2012), Merz and Wüthrich (2007, 2008), Merz *et al.* (2012), Peremans *et al.* (2018), Pröhl and Schmidt (2005), Shi *et al.* (2012), and Zhang (2010). Naturally, in multivariate models, one must estimate the additional parameters describing the correlation among particular triangles which might be in the context of state space methodology (especially, the Kalman recursions) easily performed. Let us demonstrate it only for the log-normal SSM (see the univariate model in Section 3.1); the multivariate generalization of the models in Sections 3.2 and 3.3 are analogous.

4.1. Multivariate log-normal SSM (*i*)

Let us suppose that we model N run-off triangles with the logarithmic incremental values $Y_{ij}(n) = \log X_{ij}(n)$ ($i = 0, 1, \ldots, s, j = 0, 1, \ldots, s - i, n = 1, \ldots, N$). The corresponding linear Gaussian SSM (with mutually independent residuals ε and η) written in the double-index format for $i = 0, \ldots, s, j = 1, \ldots, s$, and $n = 1, \ldots, N$ is a direct multivariate analogy of (3.3)–(3.4):

$$Y_{ij}(n) = Y_{i0}(n) + \beta_{ij}(n) + \varepsilon_{ij}(n), \quad \varepsilon_{ij}(n) \sim \text{independent } N(0, \sigma_\varepsilon(n, n)), \tag{4.1}$$

$$\beta_{ij}(n) = \beta_{i-1,j}(n) + \eta_{ij}(n), \quad \eta_{ij}(n) \sim \text{independent } N(0, \sigma_\beta(n, n)), \tag{4.2}$$

where moreover for $m, n = 1, \ldots, N$:

$$cov(\varepsilon_{ij}(m), \varepsilon_{ij}(n)) = \sigma_\varepsilon(m, n), \quad cov(\eta_{ij}(m), \eta_{ij}(n)) = \sigma_\beta(m, n). \tag{4.3}$$

Similarly the row-ordering in the time format for $t = i \cdot s + j$ is the analogy of (3.5)–(3.6):

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

$$y_t(n) - Y_{i0}(n) = \beta_t(n) + \varepsilon_t(n), \quad \varepsilon_t(n) \sim \text{independent } N(0, \sigma_\varepsilon(n,n)), \tag{4.4}$$

$$\beta_{t+1}(n) = \beta_{t-s+1}(n) + \eta_t(n), \quad \eta_t(n) \sim \text{independent } N(0, \sigma_\beta(n,n)). \tag{4.5}$$

Finally, the DLM in (3.7)–(3.9) with $\vartheta_t = (\beta_t(1), \ldots, \beta_{t-s+1}(1), \ldots, \beta_t(N), \ldots, \beta_{t-s+1}(N))'$ of dimension $N \cdot s \times 1$ for $t = 1, 2, \ldots$ can be rewritten for vectors $Y_{i0} = (Y_{i0}(1), \ldots, Y_{i0}(N))'$, $\varepsilon_t = (\varepsilon_t(1), \ldots, \varepsilon_t(N))'$, $\eta_t = (\eta_t(1), 0, \ldots, 0, \ldots, \eta_t(N), 0, \ldots, 0)'$ and matrices $\Sigma_\varepsilon = (\sigma_\varepsilon(m,n))_{m,n=1,\ldots,N}$, $\Sigma_\beta = (\sigma_\beta(m,n))_{m,n=1,\ldots,N}$ as:

$$y_t - Y_{i0} = \begin{pmatrix} 1 & 0 & \ldots & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix} \vartheta_t + \varepsilon_t, \tag{4.6}$$

$$\vartheta_{t+1} = \begin{pmatrix} \begin{matrix} 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{matrix} & & \\ & \ddots & \\ & & \begin{matrix} 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{matrix} \end{pmatrix} \vartheta_t$$

$$+ \begin{pmatrix} \begin{matrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \end{matrix} & & \\ & \ddots & \\ & & \begin{matrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \end{matrix} \end{pmatrix} \eta_t, \tag{4.7}$$

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

where the covariance matrices of the residual vector ε_t in (4.6) and η_t in (4.7) are

$$H_t = \mathrm{Var}(\varepsilon_t) = \Sigma_\varepsilon, \quad Q_t = \mathrm{Var}(\eta_t) = \Sigma_\beta. \tag{4.8}$$

The unknown parameters in this model are the ones in the (symmetric) matrices $\sum_\varepsilon$ and $\sum_\beta$.

5. Numerical study

5.1. Real data examples

Five data sets are used in the numerical study (maximally with three dependent run-off triangles, see Appendix). These data sets appeared in the actuarial literature on loss reserve modeling and some of them are used as benchmarks when comparing various approaches to loss reserving:

(1) **data set 1 (one run-off triangle):** motor bodily injury class of insurance business in one Australian state, see Atherino *et al.* (2010), Chukhrova and Johannssen (2017), Mack (1993), Pitselis *et al.* (2015), Taylor and Ashe (1983), and Verdonck *et al.* (2009);
(2) **data set 2 (one run-off triangle):** Belgian insurance industry, see Verdonck *et al.* (2009);
(3) **data set 3 (one run-off triangle):** motor insurance of Portuguese insurance company SPS, see Alpuim and Ribeiro (2003);
(4) **data set 4 (three run-off triangles):** auto insurance derived from US insurers (Schedule P of General Accident Insurance Company published by NAIC), see Zhang (2010);
(5) **data set 5 (two run-off triangles):** auto insurance for a major US insurer, see Avanzi *et al.* (2016), Shi *et al.* (2012).

For the sake of completeness, the data set 4_1 denotes the first run-off triangle in the data set 4 in the Appendix. In any case, the study is extensive so that we present only limited number of graphs or detailed numerical outputs. All data sets have the same $s = 9$ (i.e., the full rectangles have dimensions 10×10) with exception of the data set 3 with dimensions 13×10. We have investigated also other run-off triangles not reported here but with similar results, for example:

- AFG data (Automatic Facultative General Liability by Reinsurance Association of America, one run-off triangle), see Atherino *et al.* (2010), de Jong (2006), England and Verall (2002), RAA (1991);
- Belgian non-life insurance (two run-off triangles), see Avanzi *et al.* (2016), Verdonck and Van Wouwe (2011);
- auto liability and general liability business by Reinsurance Association of America (two run-off triangles), see Braun (2004), Merz and Wüthrich (2007, 2008), RAA (1991);
- a US insurer 1986–1995 (three run-off triangles), see de Jong (2012).

The main outputs of the numerical study are projections of IBNR reserves for particular (dependent) run-off triangles in data sets 1–5 including their prediction errors calculated by means of Kalman recursions in particular SSMs introduced in Sections 3 and 4. Moreover, there are two categories of these outputs: either (a) one calculates the moments of projected IBNR reserves applying the Kalman smoothing algorithm of the type (2.8) (see, e.g., Table 3 for the univariate data set 1 with projected values in black color and Figure 1 for the data the set 4of dimension 3 with projected values by dashed line using in both cases the SSM from Section 3.1) or (b) one employs the approach based on the simulation smoother(see Section 5.2.2) in combination with Kalman recursions and receives the histogram of projected loss reserves (see, e.g., Figure 2 for the data set 4 of dimension 3 again by the SSM from Section 3.1). It is necessary to stress the fact that, for example, in Figure 1 we display all observations including the initial values in the first column ($Y_{i0}, i = 0, 1, \cdots, s$) and not only the time series $\{y_t\}$ from Table 2.

5.2. Technical arrangements

The numerical calculations in this section were realized by means of the software package KFAS (see, e.g., Helske (2017)). Even though this R package is suitable for the state space modeling with the observations generally from the exponential family, we applied only the Gaussian distribution. The unknown parameters of the SSMs, that is, only the variance parameters in this numerical study, are estimated by the maximum likelihood approach applying the quasi-maximum likelihood estimate (QMLE) (the results by the alternative expectation–maximization (EM) algorithm are mostly comparable, see, e.g., Shumway and Stoffer (1982) and, moreover, one can use other software instruments than KFAS for Kalman recursions alternatively). The diffuse initial state vector was employed (i.e., $\vartheta_1 \sim N(0, P_1)$ with $P_1 = \kappa \cdot I$, where I denotes the identity matrix of corresponding dimension and $\kappa >> 0$). Refer also Section 3.1. Some technical aspects important from the practical point of view should be mentioned in more detail.

5.2.1. *Parameter estimation*

The R package KFAS solves various technicalities of linear state space modeling. Additionally, some novelties can improve the implementation efficiency. For instance, we have found that the random initializations of QMLE within the estimation of unknown parameters of the corresponding SSMs usually have a positive effect (the number 200 of such random initializations seems to be sufficient). In particular, when estimating unknown parameters of SSMs, one can repeatedly draw a vector of initial parameter values from, for example, multivariate uniform distribution and calculate the corresponding QMLE. Afterward, the estimation output with the maximal log-likelihood is selected. Since processing of EM estimation which is occasionally recommended for

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 3

PROJECTIONS IN DATA SET 1 (SEE TABLE A1 IN APPENDIX) USING THE LOG-NORMAL SSM (I) FROM SECTION 3.1 (THE PROJECTED VALUES BY KALMAN RECURSIONS ARE IN BLACK COLOR).

	Development year										
Acc. year	0	1	2	3	4	5	6	7	8	9	Row sums
0	357,848	822,608	834,578	820,891	454,979	303,418	273,646	185,981	313,296	67,948	—
1	352,118	809,436	821,215	807,747	447,694	298,560	269,264	183,003	308,280	77,926	**77,926**
2	590,507	1,357,436	1,377,190	1,354,605	750,788	500,688	451,560	306,899	579,542	130,684	**710,226**
3	310,608	714,015	724,406	712,526	394,916	263,363	237,522	178,894	304,855	68,740	**552,489**
4	443,160	1,018,721	1,033,547	1,016,595	563,447	375,754	373,702	255,293	434,999	98,076	**1,162,070**
5	396,132	910,615	923,869	908,714	503,655	369,499	334,195	228,291	388,931	87,672	**1,408,587**
6	440,832	1,013,370	1,028,121	1,011,255	615,818	411,460	372,167	254,229	433,053	97,581	**2,184,308**
7	359,480	826,362	838,391	905,425	502,570	335,841	303,819	207,570	353,569	79,625	**2,688,420**
8	376,686	865,915	964,209	949,593	527,191	352,389	318,906	217,985	371,512	83,654	**3,785,439**
9	344,014	867,733	881,386	868,221	482,169	322,447	292,025	199,859	341,395	77,201	**4,332,437**
										Projection of IBNR:	**16,901,902**
										Prediction error (CV):	**8.16%**

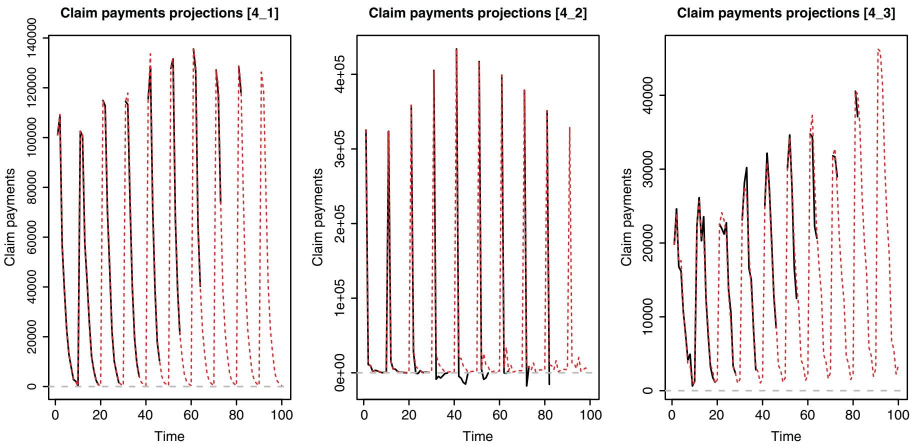

FIGURE 1: Projections in data set 4 (see Tables A4(a)–A4(c) in Appendix) using the log-normal SSM (i) from Section 4 (the projected values by Kalman recursions are plotted by dashed lines).

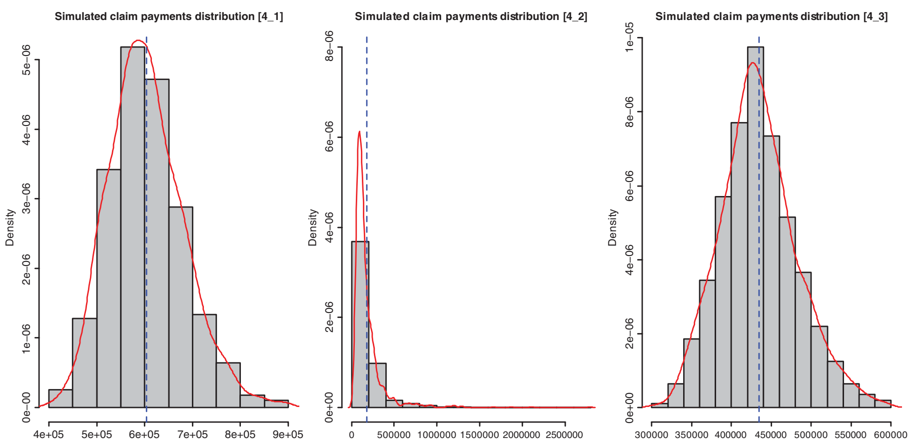

FIGURE 2: Histograms of projected IBNR reserves in data set 4 (see Tables A4(a)–A4(c) in Appendix) using the log-normal SSM (i) from Section 4 (outputs of simulation smoother using 1 000 simulations and the corresponding kernel smoothed densities).

the Kalman filter is relatively highly time-consuming in the context of loss reserving problem solved in this paper, we have used the described maximum likelihood estimators with the random initialization successfully for all data sets in the presented numerical study (see Section 5).

5.2.2. *Generating simulations in SSMs*

The simulation approach generally enables to obtain valuable information on the distribution of the projected IBNR reserves in the form of histograms or important quantiles (the quantiles are nowadays the main risk measures in finance and insurance and key figures in regulatory systems of the type of

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 4

QUANTILES OF PROJECTED IBNR RESERVES IN DATA SET 4 (SEE TABLES A4(A)–A4(C) IN APPENDIX) USING THE LOG-NORMAL SSM (*i*) FROM SECTION 4 (SELECTED EMPIRICAL QUANTILES OF 1 000 SIMULATIONS OBTAINED BY SIMULATION SMOOTHER).

Data	75% quantile	90% quantile	95% quantile	99% quantile
4_1	647,687	702,920	732,588	822,797
4_2	212,219	363,313	541,585	1,042,884
4_3	463,094	497,453	517,271	549,959

Solvency II). For this purpose, we apply the so-called *simulation smoother* introduced by Durbin and Koopman (2002) which has demonstrated consistent behavior. It is a procedure for drawing samples from the conditional distribution of state or residual vectors given the observations when assuming the Gaussian distribution of residuals. It enables generating a (typically high) number of realizations of states given the (fully estimated) SSM and all available observations (missing entries are allowed). These realizations correspond to the time series realizations which can be transformed into the run-off triangles realizations. Point out that the simulation smoother algorithm is fully supported, for example, by the KFAS package in R. Refer also to Durbin and Koopman (2012).

The simulation results can be presented by the histograms of projected IBNR reserves and the related kernel smoothed densities, for example, in Figure 2 (the data set 4 by the log-normal SSM from Section 3.1) or by the empirical quantiles selected usually in practice (see, e.g., Table 4 again for the data set 4 by the log-normal SSM from Section 3.1). One can conclude that the quantiles are more relevant for the application in practice, for example, for the data set y_4_2 (strongly skewed to the right), the mean projected IBNR reserve is approximately 55,800 using log-normal SSM from Section 3.1 (or only 4900 using the CL method with bootstrapping) while 75% quantile exceeds even 200,000. Additionally, estimated distribution of loss reserves might be very useful in terms of setting and/or monitoring their prudency levels.

5.2.3. *Inverse logarithmic transformation*

The log-normal models from Sections 3.1 and 3.2 suppose that the projections obtained for the logarithmic incremental values $Y_{ij} = \log X_{ij}$ with normal distribution are transferred back to the log-normal distribution corresponding to the original incremental values X_{ij}. Atherino *et al.* (2010) solve this problem analytically by the formulas for the first two moments of log-normal distribution based on the first two moments of normal distribution supposing that both distributions are related by means of the logarithmic transformation. Alternatively, the delta method might be applied in this context; however, it might be too approximative.

In case of performed simulation study, we supplemented the analytical approach by a natural one based on simulated run-off triangles. This

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 5

(A) CORRELATION MATRIX ESTIMATED FOR THREE TIME SERIES OF LOGARITHMIC INCREMENTAL VALUES ORDERED ROW-WISE IN PARTICULAR TRIANGLES OF DATA SET 4; (B) THE ESTIMATED MATRIX $\sum_\varepsilon$ IN LOG-NORMAL STATE SPACE MODEL (*i*) FROM SECTION 4 FOR DATA SET 4.

(a)	(b)
$\begin{pmatrix} 1 & 0.814 & 0.941 \\ 0.814 & 1 & 0.759 \\ 0.941 & 0.759 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -0.999 & -0.966 \\ -0.999 & 1 & 0.968 \\ -0.966 & 0.968 & 1 \end{pmatrix}$

approach: (*i*) takes exponentials on the simulated log values from a given run of the simulation smoother; (*ii*) creates the simulated (total) IBNR reserve by summing up the values obtained from (*i*) for a given run of the simulation smoother; and (*iii*) calculates sample averages and sample coefficients of variation of (total) IBNR reserve by means of corresponding Monte Carlo replicates.

5.2.4. *Multivariate run-off triangles*

The SSMs of the type (4.1)–(4.3) (or (4.6)–(4.8) rewritten to the equivalent DLM form) can treat effectively several run-off triangles of dependent business lines jointly taking the correlations among triangles into account. According to the references given in Section 4, such a simultaneous modeling may improve the loss reserving process considerably.

The correlation among particular triangles is modeled by the covariance matrices $\sum_\varepsilon$ and $\sum_\beta$ of the residuals in the observation equation and the parameters in the state equation, respectively (refer to (4.8)). Particularly, one can assume that the matrix $\sum_\beta$ is diagonal since in various real data examples the correlations among run-off triangles seem to be modeled sufficiently by the non-diagonal matrix $\sum_\varepsilon$.

Let us consider, for example, the data set 4 with three run-off triangles. The triangles are correlated (see the correlation matrix in Table 5(a) which is estimated for three time series of logarithmic incremental values ordered row-wise in particular triangles). The corresponding log-normal SSM from Section 4 reflects this fact by high correlations in the estimated matrix $\sum_\varepsilon$ (see Table 5(b)).

5.2.5. *Outliers*

There are sophisticated methods of robust statistics in the actuarial literature concerning the outliers in run-off triangles (see, e.g., Peremans *et al.* (2018), Pitselis *et al.* (2015), Verdonck and Van Wouwe (2011)). Atherino *et al.* (2010) apply an intervention approach that models the outliers by dummies added to the structural SSM; this approach assumes that one identifies the positions of particular outliers and includes subjective decisions (even though the identified outliers are then verified by statistical tests).

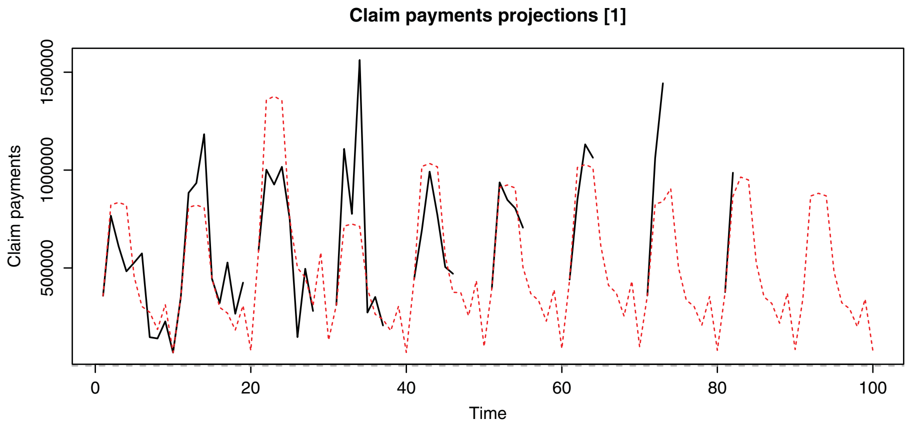

FIGURE 3: Projections in data set 1 (see Table A1 in Appendix) using the log-normal SSM (i) from Section 3.1 (the projected values by Kalman recursions are plotted by dashed lines).

In our case, we have decided to rely on the smoothing effect of Kalman recursions and avoid an explicit modeling of outliers in run-off triangles. Moreover, according to our opinion, it is sometimes difficult to distinguish whether "untypical" or extremal incremental observations in run-off triangles are really outliers or can be explained by evincible reasons. Let us consider, for example, Figure 3 for data set 1: Are the obvious extremes 1,108,250 and 1,562,400 from the fourth row ($i = 3$, $j = 1$, $j = 3$) outliers or realistic values? If we consider the column positions $j = 1$and $j = 3$ in particular rows, then the smoothing process overestimates the observed values in these positions at first (i = 2), then it underestimates them ($i = 3$) and for further rows it stabilizes on acceptable levels. It is also possible to robustify Kalman recursions to be insensitive to outliers, see, for example, Cipra and Romera (1997).

5.3. Numerical outputs

5.3.1. *Results*

Tables 6, 7, and 8 present the main results for models from Sections 3 and 4, respectively. For all data sets (2.1)–(2.5) from Section 5.1 and Appendix, these tables show the projected IBNR reserves (including their coefficients of variation CV in %) denoted as:

– IBNR (projections of missing values calculated by KFAS);
– IBNR.sim (simulated projections of missing values calculated by the simulation smoother in KFAS).

All technicalities are described in Sections 5.1 and 5.2. Particularly in models with logarithmic transformations, the negative incremental values in the first row are replaced by the arithmetic mean of neighboring values in the same row. The results for particular models are discussed below:

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 6

IBNR RESERVES IN THE LOG-NORMAL STATE SPACE MODEL (*i*) FROM SECTION 3.1 AND 4 (PROJECTIONS OF MISSING VALUES FOR IBNR, MEANS OF 1000 SIMULATIONS FOR IBNR OBTAINED BY THE SIMULATION SMOOTHER DENOTED AS IBNR.SIM, AND SELECTED EMPIRICAL QUANTILES OF THESE SIMULATIONS).

Data	IBNR	CV (%)	IBNR.sim	CV (%)	75% quantile	90% quantile	95% quantile	99% quantile
1	16,901,902	8	16,998,803	11	18,221,002	19,531,904	20,242,288	22,058,635
2	1,467,359,712	2	1,466,896,230	2	1,487,157,175	1,504,803,629	1,513,824,180	1,538,473,746
3	3,257,011	10	3,240,869	12	3,464,285	3,763,370	3,981,324	4,349,290
4_1	601,273	17	600,296	13	647,687	702,920	732,588	822,797
4_2	199,992	155	196,690	128	212,219	363,313	541,585	1,042,884
4_3	432,961	12	434,196	11	463,094	497,453	517,271	549,959
5_1	6,579,354	4	6,591,294	5	6,827,001	7,058,335	7,201,213	7,456,239
5_2	639,516	17	636,240	18	702,078	784,325	856,898	968,171

TABLE 7

IBNR RESERVES IN THE LOG-NORMAL STATE SPACE MODEL (*ii*) FROM SECTION 3.2 AND 4 WITH HOERL'S CURVE (PROJECTIONS OF MISSING VALUES FOR IBNR, MEANS OF 1000 SIMULATIONS FOR IBNR OBTAINED BY THE SIMULATION SMOOTHER DENOTED AS IBNR. SIM, AND SELECTED EMPIRICAL QUANTILES OF THESE SIMULATIONS).

Data	IBNR	CV (%)	IBNR.sim	CV (%)	75% quantile	90% quantile	95% quantile	99% quantile
1	16,905,832	8	16,867,064	7	17,800,582	18,594,940	19,023,707	19,469,387
2	1,467,437,048	2	1,466,482,616	1	1,481,912,027	1,496,022,483	1,499,761,698	1,504,429,057
3	5,470,128	80	4,395,409	26	5,004,901	6,108,053	6,843,724	7,719,928
4_1	593,622	9	591,933	6	620,895	644,586	654,898	665,213
4_2	145,545	87	122,515	35	148,570	185,931	209,123	233,106
4_3	463,271	11	456,508	11	493,010	530,274	544,051	559,889
5_1	6,966,577	4	6,962,218	4	7,149,196	7,361,211	7,422,736	7,503,312
5_2	711,156	15	703,183	11	760,072	818,941	844,772	866,704

TABLE 8

IBNR RESERVES IN THE CL STATE SPACE MODEL (*iii*) FROM SECTION 3.3 AND 4 (PROJECTIONS OF MISSING VALUES FOR IBNR, MEANS OF 1,000 SIMULATIONS FOR IBNR OBTAINED BY THE SIMULATION SMOOTHER DENOTED AS IBNR.SIM, AND SELECTED EMPIRICAL QUANTILES OF THESE SIMULATIONS).

Data set	IBNR	CV (%)	IBNR.sim	CV (%)	75% quantile	90% quantile	95% quantile	99% quantile
1	17,472,674	–	17,435,548	134	29,642,836	48,356,408	59,799,149	88,409,978
2	1,469,338,795	–	1,475,656,964	12	1,593,602,696	1,695,912,305	1,738,956,728	1,836,298,104
3	3,144,328	–	3,117,631	46	4,124,875	4,995,185	5,619,690	6,755,826
4_1	641,304	–	631,671	31	763,348	883,089	948,261	1,097,822
4_2	–4317	–	–7793	2675	122,245	265,560	349,565	506,685
4_3	458,717	–	438,030	40	545,559	654,595	737,169	881,581
5_1	6,778,109	–	6,805,270	44	8,837,636	10,600,184	11,521,724	13,740,705
5_2	846,291	–	476,443	210	990,184	1,831,308	2,384,050	3,480,765

(*i*) *log-normal SSM* (Sections 3.1 and 4): The model seems to be adequate for all data sets in Table 6 (both for the univariate and the multivariate ones). The residual analysis in Section 5.3.2 testifies to this conclusion. The accordance with the results by Mack's CL (MCL) and bootstrapping CL (BCL) in Section 5.3.3 is also relatively high. The only exception are the results for the data set 4_2 caused by the high number of negative incremental values (21 negative values including 2 in the first row) which the model treats as additional missing observations.

(*ii*) *log-normal SSM with Hoerl's curve* (Sections 3.2 and 4): Even though the results of the residual analysis and the accordance with MCL and BCL are acceptable, the simulation results for IBNR.sim of some data sets in Table 7 are not adequate and show very high CV (e.g., for the data set 3 if we again respect the argument that the high CV for the data set 4_2 is due to the high number of negative incremental values). An improvement can be achieved by trimming, for example, 10% of the highest and lowest simulated IBNR values.

(*iii*) *CL SSM* (Sections 3.3 and 4): Similarly, as the previous model (*ii*), this model shows acceptable results of the residual analysis and the accordance with MCL and BCL. However, the problem of the CL SSM consists in the calculation of standard deviation of the projected IBNR reserve (see Table 8): it is not feasible to calculate it analytically and its simulation calculation provides high values mainly due to the multiplication effect (the uncertainty of the projected values, e.g., in the last column cumulates multiplicatively the uncertainty of all previous columns). Moreover, modifications of the model (e.g., consisting in replacing ratios $C_{ij}/C_{i,j-1}$ by C_{ij}/C_{i0} mentioned in Section 3) neither brings a significant improvement. In comparison with the CL models by Mack (1993) and others, the mentioned problem consists mainly in the fact that the basic Mack's model assumes constant development factors for particular columns of run-off triangles, while in the CL SSM these factors are modeled stochastically (the analytical formulas for standard deviations of products of random variables are, therefore, very complex). On the other hand, one can compare both approaches empirically: for example, the out-of-sample studies show that numerical results are mostly comparable (see Section 5).

5.3.2. *Residual analysis*

The numerical results of the previous section are relevant only when a suitable residual analysis verifies that the corresponding SSMs fit the data adequately. The prediction residuals for this analysis should be primarily standardized by their standard deviations which are natural by-products of the Kalman scheme (see, e.g., Figure 4). The residual analysis can be realized by means of the classic econometrical instruments, namely Q–Q plots and Cullen and Frey graphs to verify the type of distribution (i.e., normality in our case, see Figures 5 and 6), various statistical tests to verify the independence of standardized residuals (see

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

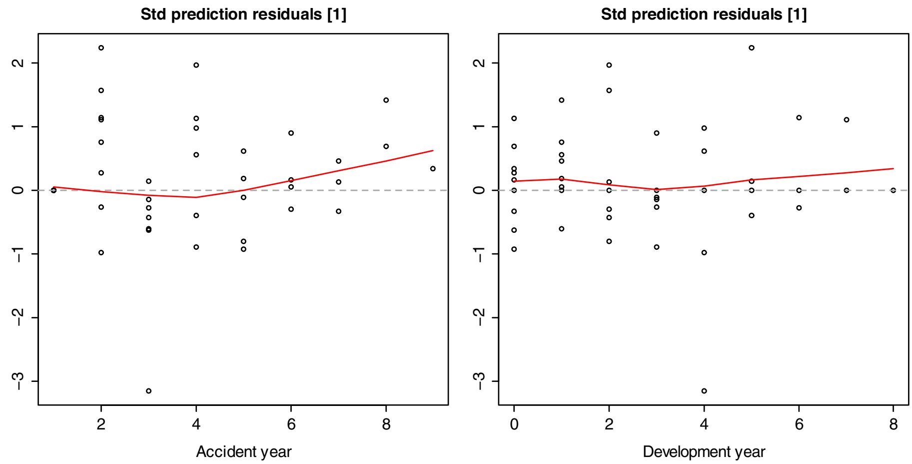

FIGURE 4: Standardized prediction residuals of projected claim payments in data set 1 (see Table A1 in Appendix) using the log-normal SSM (*i*) from Section 3.1.

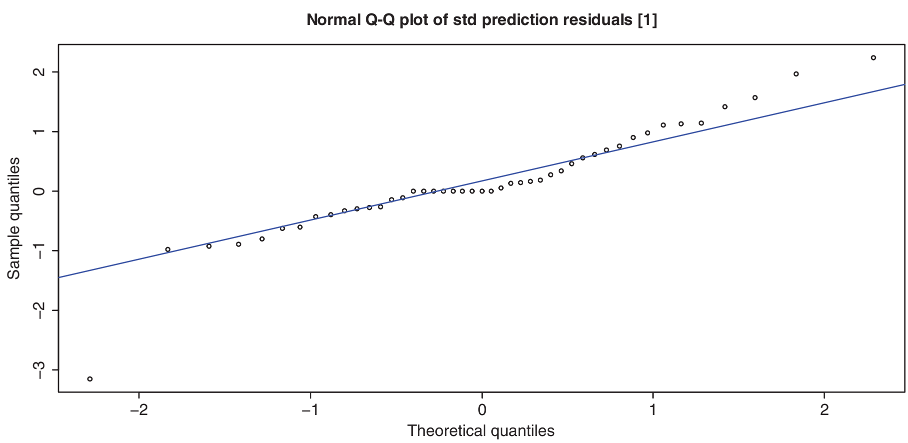

FIGURE 5: Q-Q plot of standardized prediction residuals of projected claim payments in data set 1 (see Table A1 in Appendix) using the log-normal SSM (*i*) from Section 3.1.

the correlogram in Figure 7 or the classic univariate or multivariate Ljung–Box test in Table 9) and the elimination of heteroscedasticity (see the modified univariate or multivariate Ljung–Box test based on the squared residuals in Table 9), and others. The results of this analysis are mostly acceptable for all models (*i*)–(*iii*) and all data sets (not only for the ones reported here).

5.3.3. *Comparing with other models*

One can compare the obtained numerical results with the ones calculated by available statistical software (including actuarial libraries) or reported in actuarial literature:

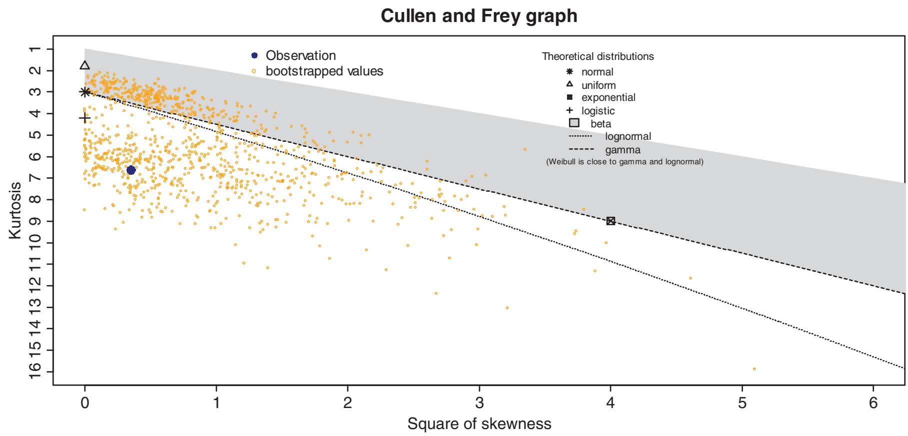

FIGURE 6: Cullen and Frey graph of standardized prediction residuals of projected claim payments in data set 1 (see Table A1 in Appendix) using the log-normal SSM (*i*) from Section 3.1.

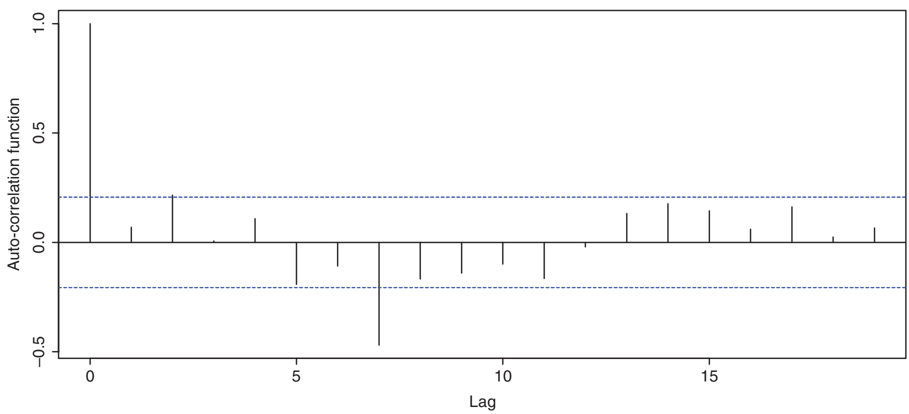

FIGURE 7: Correlogram of standardized prediction residuals of projected claim payments in data set 1 (see Table A1 in Appendix) using the log-normal SSM (*i*) from Section 3.1.

Table 10 shows the projected IBNR reserves (including CV in %) calculated by means of the software R (see Carrato *et al.* (2019)) and denoted as:

- IBNR.mcl (stochastic CL by Mack (1993));
- IBNR.bcl (bootstrapping in CL).

One can see that the accordance of SSM with CL approach is the highest for the model (*i*) (compare Table 6 with Table 10). As far as the models (*ii*) and (*iii*) are concerned, there appear disproportions when modeling some data sets (compare Tables 7 and 8 with Table 10) which are given by the character of incremental data (e.g., due to the high number of negative incremental values in the data set 4_2). Further disproportions may be caused by the fact that the

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE 9

LJUNG–BOX TEST AND MODIFIED LJUNG–BOX TEST BASED ON THE SQUARED STANDARDIZED PREDICTION RESIDUALS OF PROJECTED CLAIM PAYMENTS IN THE LOG-NORMAL STATE SPACE MODEL (*i*) FROM SECTION 3.1 AND 4.

Data	Univariate Ljung–Box test for residuals (p-values)	Univariate Ljung–Box test for squared residuals (p-values)	Multivariate Ljung–Box test for residuals (p-values)	Multivariate Ljung–Box test for squared residuals (p-values)
1	0.860	0.975	0.860	0.975
2	0.028	0.176	0.028	0.176
3	0.595	0.172	0.595	0.172
4_1	0.980	0.997	0.011	0.369
4_2	0.601	0.681		
4_3	0.908	0.689		
5_1	0.080	0.717	0.070	0.713
5_2	0.405	0.232		

TABLE 10

IBNR RESERVES CALCULATED BY MACK'S CL (IBNR.MCL) AND BOOTSTRAPPING CL (IBNR.BCL).

Data	IBNR.mcl	CV (%)	IBNR.bcl	CV (%)
1	17,998,814	13	18,272,982	16
2	1,463,319,571	3	1,465,385,235	2
3	3,119,812	8	3,121,484	14
4_1	624,247	5	625,612	5
4_2	–776	3633	4926	8925
4_3	428,781	7	429,662	8
5_1	6,439,764	5	6,455,058	5
5_2	487,401	19	492,682	20

CL approach does not exploit the correlation structure in the multivariate data sets 4 and 5.

The data set 1 is a typical benchmark when comparing various loss reserving methods (it was introduced originally by Taylor and Ashe (1983)). For this data set, Table 11 compares methods based on the classic CL principle (the upper part of this Table, see, e.g., Chukhrova and Johannssen (2017)) with the ones based on the SSM principle (the lower part of this Table). One can see that the results separately in each group are similar (with some exceptions, e.g., the higher IBNR reserve by Bornhuetter–Ferguson method, see (Chukhrova and Johannssen, 2017, p. 15, Table 5). In particular, the SSM results based on models (*i*)–(*iii*) are similar to the ones by Atherino *et al.* (2010) (mainly when applying the simulation methodology in these models and ignoring the

TABLE 11

IBNR RESERVES FOR THE DATA SET 1 (SEE TABLE A1 IN APPENDIX): COMPARISON OF CL AND SSM APPROACHES (SEE THE UPPER AND LOWER PART OF THIS TABLE, RESPECTIVELY).

Method	IBNR	CV (%)
Alpuim and Ribeiro (2003)	18,309,304	8.9%
Bornhuetter-Ferguson	20,548,942	5.9%
CL	18,680,856	13.1%
Chukhrova and Johannssen (2017)	18,307,113	7.5%
Atherino *et al.* (2010)	16,871,345	7.1%
SSM (*i*) from Section 3.1	16,901,902	8.2%
SSM (*i*) from Section 3.1: simulations	16,998,803	11.3%
SSM (*ii*) from Section 3.2	16,905,832	8.2%
SSM (*ii*) from Section 3.2: simulations	16,867,064	7.5%
SSM (*iii*) from Section 3.3	17,472,674	–
SSM (*iii*) from Section 3.3: simulations	17,435,548	134.3%

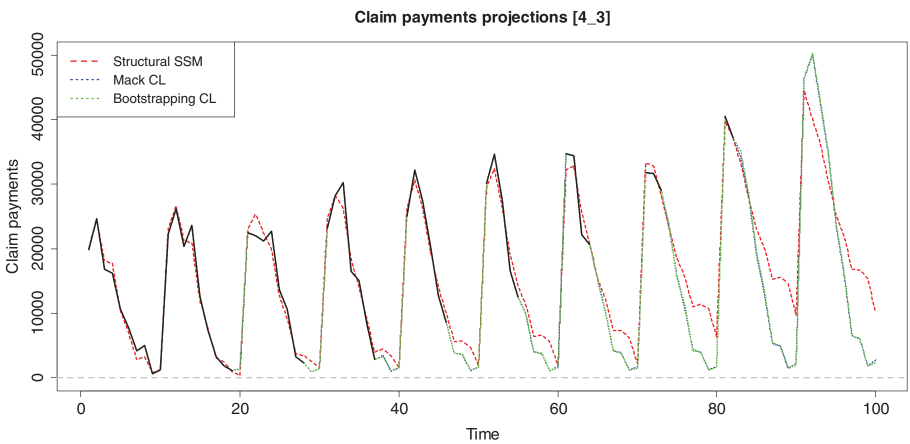

FIGURE 8: Projections in data set 4_3 (see Tables A4(c) in Appendix) using the structural SSM by Atherino *et al.* (2010).

higher IBNR in the model (*iii*) due to the extremely high standard deviation determined by the multiplicative effect).

If we compare explicitly the structural SSM by Atherino *et al.* (2010) and SSM (*i*) or (*ii*) based on the ANOVA principle, then it seems that the simulation process in latter ones replaces the need of the outlier identification in the structural SSM models recommended by Atherino *et al.* (2010). Moreover, it seems that the structural SSM approach is not suitable generally for each run-off triangle in practice. For example, Figure 8 shows the projection in the data set 4_3 using the structural SSM by Atherino *et al.* (2010). One can see that the structural model projects the missing data as in a given seasonal time

TABLE 12

BACKTESTING OF RUN-OFF TRIANGLE DIAGONALS: COMPARISON OF MAPE FOR THE LOG-NORMAL STATE SPACE MODEL (*i*) FROM SECTION 3.1 AND 4.

Data	SSM (*i*)	SSM (*i*): simulations	Mack's CL	Bootstrapping CL
1	24.23%	24.10%	32.90%	32.52%
2	10.62%	10.60%	10.13%	10.08%
3	56.87%	56.93%	40.69%	40.80%
4_1	47.19%	47.19%	51.09%	51.55%
4_2	884.47%	872.07%	340.98%	343.18%
4_3	30.29%	30.12%	26.19%	26.55%
5_1	13.75%	13.79%	15.83%	15.71%
5_2	190.03%	190.15%	195.59%	197.81%

TABLE 13

BACKTESTING OF RUN-OFF TRIANGLE DIAGONALS: COMPARISON OF MAPE FOR THE LOG-NORMAL STATE SPACE MODEL (*ii*) FROM SECTION 3.2 AND 4 WITH HOERL'S CURVE.

Data	SSM (*ii*)	SSM (*ii*): simulations	Mack's CL	Bootstrapping CL
1	31.92%	32.43%	32.90%	32.52%
2	10.62%	10.51%	10.13%	10.08%
3	61.53%	61.63%	40.69%	40.80%
4_1	45.60%	45.79%	51.09%	51.55%
4_2	1058.03%	1088.28%	340.98%	343.18%
4_3	20.06%	20.44%	26.19%	26.55%
5_1	13.08%	13.10%	15.83%	15.71%
5_2	190.02%	190.11%	195.59%	197.81%

series without respecting their original run-off character (see for comparison the projections by CL methods plotted also in Figure 8).

Additionally, the performance of the suggested models might be assessed also by means of back testing of run-off triangle diagonals (out-of-sample study): firstly, the original run-off triangles are transformed by eliminating their diagonals (i.e., the corresponding entries are set to NA) and deleting their last columns and rows. Secondly, the transformed run-off triangles are processed by employing selected modeling methods. Thirdly, the estimated outputs are compared with that previously eliminated from the original run-off triangles by some of standard evaluation criteria, for example, by the mean absolute prediction error (MAPE). Tables 12, 13, and 14 display the outcomes of this analysis. It is evident that the results are predominantly consistent with the MCL and the BCL methods. In other words, it is demonstrated that the modeling approaches proposed in Sections 3 and 4 are competitive to the classic ones.

TABLE 14

BACKTESTING OF RUN-OFF TRIANGLE DIAGONALS: COMPARISON OF MAPE FOR THE CL STATE SPACE MODEL (*iii*) FROM SECTION 3.3 AND 4.

Data	SSM (*iii*)	SSM (*iii*): simulations	Mack's CL	Bootstrapping CL
1	24.17%	32.93%	32.90%	32.52%
2	10.74%	10.58%	10.13%	10.08%
3	39.17%	37.44%	40.69%	40.80%
4_1	50.53%	55.09%	51.09%	51.55%
4_2	1040.82%	1112.84%	340.98%	343.18%
4_3	28.24%	29.67%	26.19%	26.55%
5_1	15.51%	17.19%	15.83%	15.71%
5_2	223.47%	248.20%	195.59%	197.81%

6. CONCLUSIONS

The paper contributes to the investigations on stochastic loss reserving based on the linear Gaussian SSMs which seem to be very efficient statistical instruments nowadays. In particular, one calculates the loss reserves by means of Kalman recursions as projections of missing observations in time series constructed by row-ordering run-off triangles (these time series are multivariate in the case of dependent run-off triangles). The empirical evidence realized by means of an extensive numerical study suggests that the SSMs which are used in the paper in the combination with the simulation smoother technique are adequate in routine actuarial situations. Moreover, they also deliver information on the distribution of the projected IBNR reserves in terms of histograms or quantiles which plays a key role when evaluating risk. In particular, one may empirically calculate various risk measures, for example, value at risk or expected shortfall, and set and/or monitor the prudency level in loss reserves.

Various potential topics for future research can be sketched briefly:

(1) One could try to suggest models for column-wise ordering of run-off triangles to time series (the time series of sub diagonals corresponding to particular calendar years were investigated, e.g., by de Jong and Zehnwirth (1983) or Verrall (1989)).

(2) Combination of IBNR projections provided by several alternatives of SSMs may improve the quality of combined projection (in general, combinations of various types of predictions are recommended in practical time series analysis).

ACKNOWLEDGEMENT

This paper is supported by the Czech Science Foundation 19-28231X. The authors thank to anonymous referees for helpful comments and suggestions improving the paper.

References

ALPUIM, T. and RIBEIRO, I. (2003) A state space model for run-off triangles. *Applied Stochastic Models in Business and Industry*, **19**, 105–120.

ATHERINO, R., PIZZINGA, A. and FERNANDES, C. (2010) A row-wise stacking of the runoff triangle: State space alternatives for IBNR reserve prediction. *ASTIN Bulletin*, **40**(2), 917–946.

AVANZI, B., LAVENDER, M., TAYLOR, G. and WONG, B. (2016) On the impact detection and treatment of outliers in robust loss reserving. In *Proceedings of the Actuaries Institute, General Insurance Seminar*, pp. 13–15, Melbourne.

BJÖRKWALL, S., HÖSSJER, O., OHLSSON, E. and VERRALL, R. (2011) A generalized linear model with smoothing effects for claims reserving. *Insurance: Mathematics and Economics*, **49**, 27–37.

BRAUN, C. (2004) The prediction error of the chain ladder method applied to correlated run-off triangles. *ASTIN Bulletin*, **34(2)**, 399–423.

BROCKWELL, P.J. and DAVIS, R.A. (2016) *Introduction to Time series and Forecasting*, 3^{rd} edition. Springer.

CARRATO, A., CONCINA, F., GESMANN, M., MURPHY, D., WUTHRICH, M. and ZHANG, W. (2019) *Claims reserving with R: ChainLadder-0.2.10 Package Vignette*. https://rdrr.io/cran/ChainLadder/.

CHUKHROVA, N. and JOHANNSSEN, A. (2017) State space models and the Kalman-filter in stochastic claims reserving: Forecasting, filtering and smoothing. *Risks* **5**, 30. doi:10.3390/risks5020030.

CIPRA, T. (2010) *Financial and Insurance Formulas*. Springer.

CIPRA, T. and ROMERA, R. (1997) Kalman filter with outliers and missing observations. *Test* **6**, 379–395.

COSTA, L., PIZZINGA, A. and ATHERINO, R. (2016) Modeling and predicting IBNR reserve: extended chain ladder and heteroscedastic regression analysis. *Journal of Applied Statistics*, **43(5)**, 847–870.

DE JONG, P. (2005) State space models in actuarial science. *Research Paper No. 2005/02*, Macquarie University, Sydney.

DE JONG, P. (2006) Forecasting runoff triangles. *North American Actuarial Journal*, **10**(2), 28–38.

DE JONG, P. (2012) Modeling dependence between loss triangles. *North American Actuarial Journal*, **16**(1), 74–86.

DE JONG, P. and ZEHNWIRTH, B. (1983) Claim reserving, state-space models and the Kalman filter. *Journal of the Institute of Actuaries*, **110**, 157–181.

DURBIN, J. and KOOPMAN, S.J. (2002) A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, **89**(3), 603–615.

DURBIN, J. and KOOPMAN, S.J. (2012) *Time Series Analysis by State Space Methods*, 2^{nd} edition. Oxford University Press.

ENGLAND, P.D. and VERRALL, R.J. (2002) Stochastic claims reserving in general insurance. *British Actuarial Journal*, **8**(3), 443–518.

HAMILTON, J.D. (1994) *Time Series Analysis*. Princeton University Press.

HARVEY, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

HELSKE, J. (2017) KFAS: Exponential family state space models in R. Journal of Statistical *Software*, **78**(10), 1–38. doi:10.18637/jss.v078.i10.

JOHANNSSEN, A. (2016) *Stochastische Schadenreservierung unter Verwendung vonZustandsraummodellen und des Kalman-Filters*. Hamburg: Verlag Dr. Kovac.

KLOEK, T. (1998) Loss development forecasting models: An econometrician's view. *Insurance: Mathematics and Economics*, **23**, 251–261.

KOOPMAN, S.J., HARVEY, A.C., DOORNIK, J.A. and SHEPHARD, N. (2009) *STAMP 8.2: Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants.

KREMER, E. (1982) IBNR claims and the two way model of ANOVA. *Scandinavian Actuarial Journal*, **1982**(1), 47–55.

LI, J. (2006) Comparison of stochastic reserving methods. *Australian Actuarial Journal*, **12**, 489–569.

MACK, T. (1993) Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, **23**(2), 213–225.

MERZ, M. and WÜTHRICH, M.V. (2007) Prediction error of the chain ladder reserving method applied to correlated run-off triangles. *Annals of Actuarial Science*, **2**(1), 25–50.

MERZ, M. and WÜTHRICH, M.V. (2008) Prediction error of the multivariate chain ladder reserving method. *North American Actuarial Journal*, **12**(2), 175–197.

MERZ, M., WÜTHRICH, M.V. and HASHORVA, E. (2012) Dependence modelling in multivariate claims run-off triangles. *Annals of Actuarial Science*, **7(1)**, 3–25.

NTZOUFRAS, I. and DELLAPORTAS, P. (2002) Bayesian modelling of outstanding liabilities incorpo-rating claim count uncertainty. *North American Actuarial Journal*, **6**(1), 13–28.

PANG, L. and HE, S. (2012) The application of state-space model in outstanding claims reserve. In *Conference Paper, International Conference on Information Management, Innovation Management and Industrial Engineering 2012*. Sanya.

PEREMANS, K., VAN AELST, S. and VERDONCK, T. (2018) A robust general multivariate chain ladder method. *Risks*, **6**, 108. doi:10.3390risks6040108.

PITSELIS, G., GRIGORIADOU, V. and BADOUNAS, I. (2015) Robust loss reserving in a log-linear model. *Insurance: Mathematics and Economics*, **64**, 14–27.

PRÖHL, C. and SCHMIDT, K.D. (2005) Multivariate chain-ladder. *Dresdner Schriften zur Versicherungsmathematik*. Technische Universität Dresden.

RAA (1991): *Historical Loss Development Study*. Reinsurance Association of America, Washington D.C.

RENSHAW, A.E. (1989) *Chain-ladder and interactive modelling (claims reserving and GLIM)*. *Journal of the Institute of Actuaries*, **116**, 559–587.

RENSHAW, A.E. and VERRALL, R.J. (1998) A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, **4**(4), 903–923.

SHI, P., BASU, S. and MEYERS G.G. (2012) A Bayesian log-normal model for multivariate loss reserving. *North American Actuarial Journal*, **16**(1), 29–51.

SHUMWAY, R.H. and STOFFER, D.S. (1982) An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, **3**(4), 253–264.

SHUMWAY, R.H. and STOFFER, D.S. (2017) *Time Series Analysis and Its Applications (With R Examples)*, 4th edition. New York: Springer.

STOFFER, D.S. and WALL, K.D. (1991) Bootstrapping state-space models: Gaussian maximum likelihood estimation of the Kalman filter. *Journal of the American Statistical Association*, **86**(416), 1024–1033.

TAYLOR, G.C. and ASHE, F.R. (1983) Second moments of estimates of outstanding claims. *Journal of Econometrics*, **23**, 37–61.

TAYLOR, G.C., MCGUIRE, G. and GREENFIELD, A. (2003) *Loss reserving: Past, present and future*. *Research Paper No. 109*. University of Melbourne 2003 (*an invited lecture to the 34th ASTIN Colloquium*, Berlin 2003).

VERDONCK, T. and VAN WOUWE, M. (2011) Detection and correction of outliers in the bivariate chain-ladder method. *Insurance: Mathematics and Economics*, **49**, 188–193.

VERDONCK, T., VAN WOUWE, M. and DHAENE, J. (2009) A robustification of the chain-ladder method. *North American Actuarial Journal*, **13**(2), 280–298.

VERRALL, R. (1989) A state space representation of the chain ladder linear model. *Journal of the Institute of Actuaries*, **116**, 589–610.

VERRALL, R. (1991) On the estimation of reserves from loglinear models. *Insurance: Mathematics and Economics*, **10**, 75–80.

VERRALL, R. (1994) A method for modelling varying run-off evolutions in claims reserving. *ASTINBulletin*, **24**(2), 325–332.

WRIGHT, T. (1990) A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries*, **117**, 677–731.

WÜTHRICH, M.V. and MERZ, M. (2008) *Stochastic Claims Reserving Methods in Insurance*. Chichester: Wiley.

ZEHNWIRTH, B. (1996) Kalman filters with applications to loss reserving. *Research Paper 27*, University of Melbourne, Department of Economics, Centre for Actuarial Studies.

ZHANG, Y. (2010) A general multivariate chain ladder model. *Insurance: Mathematics and Economics*, **46**, 588–599.

RADEK HENDRYCH (Corresponding author)
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics
Charles University
Prague, Czech Republic
E-Mail: hendrych@karlin.mff.cuni.cz

TOMAS CIPRA
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics
Charles University
Prague, Czech Republic
E-Mail: cipra@karlin.mff.cuni.cz

APPENDIX: DATA OF NUMERICAL STUDY

(1) *Data set 1 (one run-off triangle): Motor bodily injury class of insurance business in one Australian state*, see Atherino *et al.* (2010), Chukhrova and Johannssen (2017), Mack (1993), Pitselis *et al.* (2015), Taylor and Ashe (1983), and Verdonck *et al.* (2009):

TABLE A1

DATA SET 1: INCREMENTAL RUN-OFF TRIANGLE.

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	357,848	766,940	610,542	482,940	527,326	574,398	146,342	139,950	227,229	67,948
1	352,118	884,021	933,894	1,183,289	445,745	320,996	527,804	266,172	425,046	
2	590,507	1,001,799	926,219	1,016,654	750,816	146,923	495,992	280,405		
3	310,608	1,108,250	776,189	1,562,400	272,482	352,053	206,286			
4	443,160	693,190	991,983	769,488	504,851	470,639				
5	396,132	937,085	847,498	805,037	705,960					
6	440,832	847,631	1,131,398	1,063,269						
7	359,480	1,061,648	1,443,370							
8	376,686	986,608								
9	344,014									

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

(2) *Data set 2 (one run-off triangle): Belgian insurance industry*, see Verdonck *et al.* (2009):

TABLE A2

DATA SET 2: INCREMENTAL RUN-OFF TRIANGLE.

Acc. year	Development year									
	0	1	2	3	4	5	6	7	8	9
0	135,388,126	90,806,681	68,666,715	55,736,215	46,967,279	35,463,367	30,477,244	24,838,121	18,238,489	14,695,083
1	125,222,434	89,639,978	70,697,962	58,649,114	46,314,227	41,369,299	34,394,512	26,554,172	24,602,209	
2	136,001,521	91,672,958	78,246,269	62,305,193	49,115,673	36,631,598	30,210,729	29,882,359		
3	135,277,744	103,604,885	78,303,084	61,812,683	48,720,135	39,271,861	32,029,697			
4	143,540,778	109,316,613	79,092,473	65,603,900	51,226,270	44,408,236				
5	132,095,863	88,862,933	69,269,383	57,109,637	48,818,781					
6	127,299,710	92,979,311	61,379,607	50,317,305						
7	120,660,241	89,469,673	71,570,718							
8	134,132,283	87,016,365								
9	131,918,566									

(3) *Data set 3 (one run-off triangle): Motor insurance of Portuguese insurance company SPS*, see Alpuim and Ribeiro (2003):

TABLE A3

DATA SET 3: INCREMENTAL PAID RUN-OFF TRIANGLE.

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9,10,···
0	528,007	243,416	46,285	50,639	24,099	16,466	16,742	7216	1455	22,455
1	510,152	253,179	44,715	37,629	19,824	14,760	25,603	4439	4076	5342
2	636,545	294,445	83,638	48,801	32,538	26,295	13,018	11,955	15,495	20,727
3	713,804	409,426	94,553	56,563	99,388	61,665	33,304	58,247	22,758	55,339
4	1,013,575	517,621	92,439	81,543	89,632	47,411	40,182	55,110	36,424	
5	1,331,709	669,554	91,761	94,350	63,751	23,740	34,138	34,785		
6	1,711,542	752,618	226,533	135,677	70,288	53,371	57,800			
7	1,626,786	1,037,421	110,537	95,189	71,490	80,507				
8	1,564,913	495,069	94,301	189,018	82,651					
9	1,475,219	554,382	115,045	123,955						
10	1,432,485	535,867	142,918							
11	1,532,809	636,422								
12	1,747,414									

(4) *Data set 4 (three run-off triangles): Auto insurance derived from US insurers (Schedule P of General Accident Insurance Company published by NAIC)*, see Zhang (2010):

TABLE A4(A)

DATA SET 4_1: INCREMENTAL RUN-OFF TRIANGLE (PERSONAL AUTO INSURANCE 1).

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	101,125	108,796	56,697	38,489	22,743	12,819	7761	2763	2160	231
1	102,541	100,672	57,464	42,505	25,750	12,016	6385	2480	710	
2	114,932	112,772	70,416	47,422	22,218	10,239	5612	1613		
3	114,452	113,309	73,311	39,597	19,310	9269	4077			
4	115,597	128,014	71,604	39,275	17,886	10,362				
5	127,760	131,656	67,559	38,805	20,945					
6	135,616	126,678	64,792	40,271						
7	127,177	117,072	73,723							
8	128,631	118,172								
9	126,288									

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

TABLE A4(B)

DATA SET 4_2: INCREMENTAL RUN-OFF TRIANGLE (PERSONAL AUTO INSURANCE 2).

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	325,423	11,003	9635	1665	3269	2603	1199	228	−57	395
1	323,627	15,640	5240	4788	1743	545	467	181	−38	
2	358,410	27,920	−646	−985	2979	276	586	896		
3	405,319	−8678	−4808	−7014	−3905	−751	−457			
4	434,065	−4754	−7130	−12,859	−15,168	−1352				
5	417,178	5129	−8821	−6775	−208					
6	398,929	−142	−767	2520						
7	378,754	−17,657	8231							
8	351,081	−15,574								
9	329,236									

TABLE A4(C)

DATA SET 4_3: INCREMENTAL RUN-OFF TRIANGLE (COMMERCIAL AUTO INSURANCE).

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	19,827	24,622	16,756	16,193	10,681	7616	4158	4936	638	1263
1	22,331	26,149	20,309	23,567	12,602	7441	3239	1777	1156	
2	22,533	21,951	21,207	22,744	13,609	10,628	3301	2386		
3	23,128	28,200	30,214	16,521	15,086	8366	2832			
4	25,053	32,167	27,387	20,329	12,727	8517				
5	30,136	34,631	27,521	16,547	12,491					
6	34,764	34,361	22,229	20,633						
7	31,803	31,668	28,968							
8	40,559	37,108								
9	46,285									

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press

(5) *Data set 5 (two run-off triangles): auto insurance for a major US insurer*, see Avanzi *et al.* (2016), Shi *et al.* (2012):

TABLE A5(A)

DATA SET 5_1: INCREMENTAL PAID RUN-OFF TRIANGLE (PERSONAL AUTO LINE).

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	1,376,384	1,211,168	535,883	313,790	168,142	79,972	39,235	15,030	10,865	4086
1	1,576,278	1,437,150	652,445	342,694	188,799	76,856	35,042	17,089	12,507	
2	1,763,277	1,540,231	678,959	364,199	177,108	78,169	47,391	25,288		
3	1,779,698	1,498,531	661,401	321,434	162,578	84,581	53,449			
4	1,843,224	1,573,604	613,095	299,473	176,842	106,296				
5	1,962,385	1,520,298	581,932	347,434	238,375					
6	2,033,371	1,430,451	633,500	432,257						
7	2,072,061	1,458,541	727,098							
8	2,210,754	1,517,501								
9	2,206,886									

TABLE A5(B)

DATA SET 5_2: INCREMENTAL PAID RUN-OFF TRIANGLE (COMMERCIAL AUTO LINE).

Acc. year	Development year 0	1	2	3	4	5	6	7	8	9
0	33,810	45,318	46,549	35,206	23,360	12,502	6602	3373	2373	778
1	37,663	51,771	40,998	29,496	12,669	11,204	5785	4220	1910	
2	40,630	56,318	56,182	32,473	15,828	8409	7120	1125		
3	40,475	49,967	39,313	24,044	13,156	12,595	2908			
4	37,127	50,938	34,154	25,455	19,421	6728				
5	41,125	53,302	40,289	39,912	6650					
6	57,515	67,881	86,734	18,109						
7	61,553	132,208	20,923							
8	112,103	33,250								
9	37,554									

https://doi.org/10.1017/asb.2020.38 Published online by Cambridge University Press