Annales Univ. Sci. Budapest., Sect. Comp. 50 (2020) 269–282

MATCHED BOOTSTRAP PROCEDURE FOR INAR(1) PROCESSES

László Németh and András Zempléni

(Budapest, Hungary)

Dedicated to the memory of Professor János Galambos

Communicated by András Benczúr

(Received February 20, 2020; accepted April 18, 2020)

Abstract. Integer-valued autoregressive (INAR) processes is a relatively new field of time series analysis. Bootstrapping such data is not an evident problem, further assumptions are needed, or possibly the bootstrap samples will not form an INAR process. In this paper we present a new, non-parametric bootstrap method, based on the idea of block bootstrap. The procedure resamples blocks that perfectly match to the last element of the previous block. We present properties of this so called matched bootstrap approach and compare our method to other frequently used bootstrap procedures for INAR processes. We apply the methods to a natural disaster dataset.

1. Introduction

Time series analysis is a dynamically developing field of mathematics. Besides (mostly) continuous data of financial and some environmental time series,

Key words and phrases: INAR, bootstrap, block bootstrap, time series.
2010 Mathematics Subject Classification: 62F40, 62M10.
The Project was supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

integer-valued digital, environmental and social data also motivate new models in order to describe and predict real life processes better. One of the most common model is the autoregressive (AR) process, which is suitable for continuous data only. Its modification is the integer-valued autoregressive (INAR) process introduced by [1]. In fact, not every stationary, integer valued time series follow this type of structure, however INAR can be a good approximation for most of such real life data sets.

In the description of an INAR process we expect correlation among the nearby elements, similarly to the AR processes. The definition of an $(X_t, t \in \mathbb{Z})$ INAR process of order p by [2] and [10] is the following:

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t,$$

where $0 \leq \alpha_i < 1$ $(i = 1, 2, \ldots, p)$ are the autocorrelation coefficients, " $\circ$ " is the binomial thinning operator introduced by [18] meaning that $\alpha \circ X_t$ is a realization of $Bin(X_t, \alpha)$. The ε_t innovation process $(\varepsilon_t, t \in \mathbb{Z})$ is independent from X_s $(s < t)$. It contains independent, identically distributed, integer valued random variables. Usual choice for the innovations is the Poisson distribution. If $\sum_{i=1}^{p} \alpha_i < 1$ the given equation has a stationary solution.

In our paper we will analyse the first order INAR(1) processes:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t,$$

using the notations described above. We assume stationarity, so $0 \leq \alpha < 1$.

One can consider a stationary INAR process as a discrete Markov chain. The work of [13] summarizes the most important definitions and properties of Markov chains. Each value is a state and there is a p_{ij} transition probability between the given i and j values. The p_{ij} probabilities strongly depends on α and the innovation distribution, which imply some statements, as follows. Denote the support of the distribution of the innovations by D and the stationary distribution by Q and let t be a natural number.

- The transition probability $p_{ts} > 0$, where $s = u+d$, $0 \leq u \leq t$ are integers and $d \in D$.
- For every other state s we have $p_{ts} = 0$.
- The states constitute one class (i.e. $\forall\ s \in D$ can be reached from t).
- The chain is irreducible, aperiodic and positive recurrent. Each state t has a positive expected return time $1/Q(t)$ (by [13] Theorem 1.7.7.).

The main questions when analysing an INAR process are to estimate the expectation, variance and the α autoregression parameters. The mean and

variance can easily be estimated based on the observations. Similarly as in case of AR processes one can use the Yule–Walker equations to estimate α for INAR processes ([7]). However constructing confidence intervals and deriving the standard deviation of the given statistics are not straightforward.

A generally accepted way for constructing confidence intervals is using bootstrap simulations [8]. We present the frequently used methods in section 2. One of the most common method developed for time series is the circular block bootstrap by [15] which uses blocks of the original process. The block selection is automatized by [16] to maximize the effectiveness of estimating the expectation. Parametric bootstrap techniques are also available for time series, see e.g [11, 4]. A comparison study by [9] reveals that parametric bootstrap methods result in the most efficient estimates for INAR series. However the parametric assumptions might fail, therefore an effective non-parametric bootstrap can be desirable.

In section 3 we introduce a new, non-parametric, block resampling-based bootstrap technique, which is related to [5]'s method, called matched bootstrap, especially developed for INAR(1) processes. In order to its proper behaviour we need to shorten the original series. However, we prove that the loss of observations is negligible for long time series.

In section 4 we compare the parameter estimates using matched bootstrap, circular block bootstrap and parametric bootstrap for INAR processes, extending the work of [9]. We show that matched bootstrap procedure is effective on INAR series both with Poisson and uniform innovations, in contrast to other methods. As a real life application, we analyse the dataset of the most significant volcanic eruptions from the last 100 years [6] in section 5.

2. Bootstrap methods for INAR processes

2.1. The AR bootstrap

Since INAR processes are closely related to AR processes, it is logical to use parametric AR bootstrap for resampling an INAR process [4]. It might be suitable for some estimation problems, however it ruins the integer structure, so they are definitely not applicable to INAR-specific problems, like estimating the proportion of zeros in the process. Let $\mathbf{X} = X_1, X_2, \ldots, X_N$ be the observations of the INAR process. The steps of the procedure are the following:

1. Construct $X_i^* = X_i - \overline{\mathbf{X}}$ centred series, where $\overline{\mathbf{X}}$ is the sample mean.

2. Estimate the coefficient α by the Yule–Walker equation of $\mathbf{X}^*$, denote the estimate by $\hat{\alpha}$.

3. Calculate $\hat{\varepsilon_i} = X_i^* - X_{i-1}^* \cdot \hat{\alpha}$ for $i = 2, 3, \ldots, N$.

4. Centre the residuals as $\tilde{\varepsilon_i} = \hat{\varepsilon_i} - \frac{1}{N-1} \sum_{j=2}^{N} \hat{\varepsilon_j}$. Denote the empirical distribution of the $\tilde{\varepsilon}$'s by $\mathbb{F}$.

5. Generate bootstrap samples by $Y_i = Y_{i-1} \cdot \hat{\alpha} + \varepsilon_t^*$, $(i = 2, ..., N)$ where ε_t^* is a random element of $\mathbb{F}$. Y_1 can be equal to X_1, or a random element of $\mathbf{X}$.

The AR bootstrap has numerous desired properties, such as consistency and asymptotic normality when estimating the mean ([9]), however it is inconsistent for the variance ([19]). Since the residuals are not integer valued and may also be negative, this procedure is not able to give a possible realization of an INAR process.

The idea of AR bootstrap can not be modified for INAR processes, since the integer valued residuals can not be estimated directly. There are some residual based methods like the one introduced by [3] based on the work of [4], but they are mostly biased (proved by [9]). The key to constructing a residual based bootstrap method is to estimate the residual distribution in an unbiased way, which is usually a very difficult task.

2.2. Parametric INAR bootstrap

To overcome the problem of estimating the residual distribution one can use parametric INAR bootstrap as [9] suggest. In this approach the observed data is used only to estimate the parameters of the INAR process and the residuals (e.g. the parameter λ in case of Poisson distributed innovations), then bootstrap samples are generated using the estimated parameters. For this procedure an assumption for the family of the residuals distribution is needed - in our case this is the Poisson distribution. The algorithm is the following:

1. Estimate the auto-regressive parameter ($\hat{\alpha}$) of the time series (e.g. using the Yule–Walker estimator).

2. Since the marginal distribution of a Poisson(λ) based INAR(1) is also Poisson with parameter $\lambda/(1-\alpha)$, one can estimate λ by using the sample mean and $\hat{\alpha}$: $\hat{\lambda} = \overline{\mathbf{X}} \cdot (1 - \hat{\alpha})$.

3. Using a suitable starting value (e.g. $Y_1 = X_1$, or $Y_1 = \lfloor \overline{\mathbf{X}} \rfloor$) the simulated bootstrap process will be

$$Y_i = \hat{\alpha} \circ Y_{i-1} + \varepsilon_i,$$

where ε_i is a Poisson($\hat{\lambda}$) distributed random variable, independent from the past of the process.

The bootstrap sample will be integer-valued using the parametric INAR bootstrap. Under mild conditions the INAR bootstrap consistency is proved by [9]. This method is usually really effective, however one needs a correct assumption for the residuals' distribution, otherwise the estimates might be biased.

2.3. Circular block bootstrap

One of the most popular way for bootstrapping time series is resampling and attaching blocks of the original sample [15]. Theoretical results and asymptotic properties can be found in [12]. For choosing the optimal block size [16] proposed an adaptive algorithm based on the sample, which is corrected by [14]. This procedure is developed for estimating the mean of the process. A comparison study of [17] investigates different types of block bootstrap methods.

A potential error can occur if the starting observation of the chosen block is near to the end of the sample. In this case one may continue the given block at the beginning of the sample (circularity). Based on these guidelines the circular block bootstrap method is the following:

Let $\mathbf{X} = X_1, X_2, \ldots, X_n$ be the original time series, and ℓ be the selected block size (e.g. by [16]). Select $i_1, i_2, \ldots, i_{\lceil n/\ell \rceil}$ starting points uniformly from $1, 2, \ldots, n$. The bootstrap sample will then be:

$$\mathbf{Y} = X_{i_1}, X_{i_1+1}, \ldots, X_{i_1+\ell-1}, X_{i_2}, \ldots, X_{i_2+\ell-1}, \ldots, X_{i_{\lceil n/\ell \rceil}}, \ldots, X_{i_{\lceil n/\ell \rceil}+\ell-1},$$

circularly returning to X_1 if necessary. If ℓ is not a divisor of n, the last block will be shorter.

The circular block bootstrap method is non-parametric, therefore no assumptions for the time series structure other than stationarity is needed. Between blocks there is no correlation (conditionally on the sample) by the selection procedure. Thus despite the favourable asymptotic properties, for e.g. estimating the autocorrelation coefficients usually an extremely large sample is needed.

3. The matched bootstrap procedure

In this section we propose a modification of the non-parametric block bootstrap procedure in order to combine the advantages and overcome the disadvantages of the existing methods. The main idea of the new method is that instead of choosing the new block randomly (as in CBB), the starting value must be the same as the previous block ends. Matching the blocks by the identical values result in the bootstrap sample. This way each pair of adjacent observations in the generated bootstrap sample appears in the real sample, too. In spite of the number of possible blocks is being smaller (compared to CBB), the correlation structure can be more precise. The idea is similar as the one described in [5], but our method allows only perfect matching of the blocks.

It is possible that a block starts at the end of the series and can not be continued. For solving this problem we will also introduce circularity with some changes to save the correlation structure. Before sampling the blocks we shorten the original sample so that the new one starts and ends with the same value. By this reduction we might loose some observations in the beginning and at the end of the series. We will use only this shorter sequence in the bootstrap procedure. If a resampled block would reach the end of the reduced series we will continue it in the beginning as the values are the same. The reduced series is long enough to make the bootstrap estimates relevant by proposition 3.1.

Proposition 3.1. *Let* $\mathbf{X} = X_1, X_2, \ldots, X_N$ *be a realization of a stationary INAR process with autocorrelation* $0 \leq \alpha < 1$*.*

a) *One can choose* $1 \leq i < j \leq N$ *such that* $X_i = X_j$ *with probability tending to* 1 *as* $N \to \infty$*.*

b) *Choose* X_i *and* X_j *as described in a) such that they are the first and the last appearance of the same value in* $\mathbf{X}$*. The difference between* N *(the length of the original sequence) and the length of the sequence* $X_i, X_{i+1}, \ldots, X_j$ *is* $o(N)$ *with probability* 1*.*

Proof. a) Denote the stationary distribution of the Markov chain (generated by the $\mathbf{X}$ process) by Q (this is the same as the stationary distribution of $\mathbf{X}$). Let t be an arbitrary element of $supp(Q)$ and denote the first occurrence of t by X_{t_1} (the waiting time has finite expectation since the Markov chain is irreducible). As the Markov chain is stationary and irreducible the expected return time of t is $m_t = 1/Q(t)$. Since $Q(t) > 0$ for each $t \in supp(Q)$, $m_t < \infty$. Therefore as $N \to \infty$, $(N - t_1) \to \infty$ too, so $P(\exists t_2 | t_1 < t_2 < N, X_{t_2} = X_{t_1}) \to 1$.

b) Let t be an arbitrary element of $supp(Q)$. The expected return time of t is $m_t = 1/Q(t)$. Suppose t appears τ times in $\mathbf{X}$. The distance between the

first and the last appearance of t can be considered as sum of $\tau-1$ independent random variables from the distribution of the return time. Denote the value of the i-th return by R_i, the time of first appearance of t by W_1 and the time from the last t to the end of $\mathbf{X}$ by W_2. Then $N = W_1 + R_1 + \cdots + R_{\tau-1} + W_2$ holds. The ratio of the loss is $\frac{W_1+W_2}{N}$, where the numerator is a sum of two random variables, so it tends to 0 with probability 1 as $N \to \infty$. ■

In the proof we used arbitrary values, but in application it can be optimized by using a value that results in the longest sequence. In fact, for realistic parameters and sample sizes this type of reduction is practically always possible (the probability is close to 1). In table 1 we give the average proportion of the simulated samples where the reduction failed. Besides, in table 2 we present the average length of reduced sequences compared to the original one. One can see, that even for the most extreme case ($\alpha = 0.99$, $\lambda = 20$) a 100 size sample is usually enough to keep at least 75% of the original observations.

Sample type	$\alpha = 0.3$ $\lambda = 3$	$\alpha = 0.99$ $\lambda = 3$	$\alpha = 0.3$ $\lambda = 20$	$\alpha = 0.99$ $\lambda = 20$
n=10	0	1.06	4.54	16.73
n=20	0	0	0	0.57
n=50	0	0	0	0
n=100	0	0	0	0

Table 1: Percentage of simulated time series, where all of the observations were unique. The calculations are based on Poisson INAR processes, using given parameters and 100 000 simulations.

Sample type	$\alpha = 0.3$ $\lambda = 3$	$\alpha = 0.99$ $\lambda = 3$	$\alpha = 0.3$ $\lambda = 20$	$\alpha = 0.99$ $\lambda = 20$
n=10	0.787	0.572	0.594	0.396
n=20	0.891	0.652	0.788	0.535
n=50	0.957	0.737	0.915	0.677
n=100	0.978	0.797	0.957	0.763

Table 2: Average ratio of the length of reduced sequence and the original observations based on Poisson INAR processes, for the given parameters. The calculations are based on 100 000 simulations. The cases when no repeating observation occurred were considered as 0.

Note that an inappropriate starting point can undermine the effectiveness of the procedure, since the first steps are needed the process to reach the stationary distribution. However it is not a problem for realistic observations since

in real life the parametrizations are less extreme than the shown simulations, so the process reaches the stationary distribution fast. Now we continue by the description of the bootstrap procedure.

Let $\mathbf{X} = X_1, X_2, \ldots, X_N$ be an integer-valued stationary time series. Let $\mathbf{X_c} = X_i, X_{i+1}, \ldots, X_j$ be the maximal sequence, where $X_i = X_j$. If more than one maximal reduced series exists, chose one arbitrary. In order to produce an $\mathbf{Y} = Y_1, Y_2, \ldots, Y_M$ (usually $M = N$) bootstrap sample use the following steps:

1. Set $Y_1 = X_1$ (or randomly from $\mathbf{X}$) and block size b.

2. If the last simulated element of the bootstrap series is Y_k, choose one of $X_i, X_{i+1}, \ldots, X_j$ observations that equals Y_k randomly, denote it X_{k^*}.

3. Let $Y_k, Y_{k+1}, \ldots, Y_{k+b} = X_{k^*}, X_{k^*+1}, \ldots, X_{k^*+b}$. If $k^* + b \leq j$ and $k + b < M$, then go to 2. Otherwise apply one (or both) of the following corrections:

   a) End of sampling series:
   If $k^* + b = j + a$, where $a > 0$, return to the process at X_i, i.e. let $Y_k, Y_{k+1}, \ldots, Y_{k+b} = X_{k^*}, X_{k^*+1}, \ldots, X_j, X_{i+1}, \ldots, X_{i+a}$.

   b) End of bootstrap series:
   If $k + b \geq M$, set $b^* = M - k$ and use b^* as block size to complete the Y sequence.

4. Repeat steps 2-3 to get the Y bootstrap process.

Using matched bootstrap technique the bootstrap sample will maintain the correlation between each pairs of observations. The innovations come directly from $\mathbf{X}$, therefore no bias appears due to its estimation.

It is possible to use random block sizes during the procedure. It might be useful for small samples with large variance to avoid too much repetition. However, our simulations showed that in general fixed block length is similarly effective as random, coinciding with the results of [12] for regular block bootstraps, thus we investigate only the fixed block-size matched bootstrap method.

It is important to mention, that the matched bootstrap procedure is applicable only for INAR(1) processes. For higher order INAR, one may use larger overlapping sequences between neighbouring blocks. However, this modification dramatically reduces in the variability of possible blocks, therefore much longer observation series is needed to its applicability.

4. Comparison

We used 5 different bootstrap methods in our comparison study, namely the AR, parametric INAR (pINAR), CBB with Politis-White block size selection procedure and matched bootstrap using 10 and 1 size blocks. We considered these block sizes, as 10 seemed to be a reasonable size (not too small, not too large, considering that we investigated samples of size 100), while block size 1 is a special case, when the bootstrap sample is increased by a single element in each step 3 in the algorithm of the matched bootstrap procedure described in section 3.

We tested the effectiveness of methods on different integer valued time series. First, we generated stationary INAR processes with Poisson innovations using fixed parameters $\alpha = 0.4$ and $\lambda = 5$ (table 3). Second, the simulated samples had $\alpha = 0.4$ autocorrelation parameter as well, while the innovation process had the values from 0, 5 or 10 (table 4) with probability 1/3. The expected value of innovations is the same in both cases, however the structure is completely different.

Statistic	AR	pINAR	CBB	Matching 10	Matching 1
Autoregressive parameter					
CI coverage %	0.917	0.916	0.644	0.856	0.816
CI width	0.362	0.366	0.361	0.325	0.365
Standard error	0.0197	0.0199	0.0396	0.0188	0.0267
Mean of the process					
CI coverage %	0.931	0.943	0.881	0.869	0.858
CI width	1.645	1.666	1.449	1.528	1.484
Standard error	0.3669	0.3694	0.3327	0.3668	0.3839
Standard deviation					
CI coverage %	0.9	1	0.874	0.848	0.634
CI width	0.904	0.959	0.847	0.826	0.754
Standard error	0.1198	0.0686	0.114	0.1162	0.1689

Table 3: Simulated statistics for INAR process with $\alpha = 0.4$ and $\lambda = 5$ for 100 sized samples with Poisson distributed innovations. The autocorrelation parameter, mean and standard deviation of the bootstrap samples were calculated using different type of procedures. Percentage of confidence intervals covering the true value, average width of CI and average mean squared error for each statistic were estimated using 1000 simulated INAR processes. For bootstrap simulations we used 1000 bootstrap samples in each case.

Statistic	AR	pINAR	CBB	Matching 10	Matching 1
Autoregressive parameter					
CI coverage %	0.924	0.928	0.643	0.834	0.808
CI width	0.364	0.367	0.359	0.322	0.361
Standard error	0.0196	0.0197	0.0404	0.0185	0.0263
Mean of the process					
CI coverage %	0.925	0.746	0.885	0.880	0.845
CI width	2.694	1.663	2.365	2.466	2.543
Standard error	1.013	0.7076	0.919	1.017	1.153
Standard deviation					
CI coverage %	0.9	0	0.885	0.862	0.791
CI width	1.125	0.958	1.048	1.004	1.004
Standard error	0.1770	3.3504	0.1675	0.1674	0.2023

Table 4: Simulated statistics for INAR process with $\alpha = 0.4$ using uniform random innovations over $\{0, 5, 10\}$ for 100 length samples. The autocorrelation parameter, mean and standard deviation of the bootstrap samples were calculated using different type of procedures. Percentage of confidence intervals covering the true value, average width of CI and average mean squared error for each statistic were estimated using 1000 simulated INAR processes. For bootstrap simulations we used 1000 bootstrap samples in each case.

For each case, we estimated the parameter α, the mean and the standard deviation by the bootstrap samples. Using these simulated values we were able to construct confidence intervals for the given statistics. We present the percentage of the confidence intervals covering the true value, average width of the confidence intervals and the mean squared error from the theoretical parameter values in tables 3 and 4.

In general one can say, that a good bootstrap CI has large coverage rate, while the it is narrow and the standard error is small. These properties usually do not coexist, therefore the favourable procedures are different in the cases.

The AR bootstrap is usually quite stable, it has large percentage, narrow CI and small error, however the bootstrap samples are not integer valued. The parametric (Poisson-distribution based) INAR bootstrap is the best in case of Poisson distributed innovations. The α estimate is also acceptable for the uniform innovation distribution, but for the mean and especially for the standard deviation it does not have the needed coverage probability, when the innovation distribution is misspecified. The classical circular block bootstrap is strong for estimating the mean, but for the other parameters it underperforms.

Comparing the two matched bootstrap methods one can see, that 10 size blocks usually perform better than 1 size ones. Generally we can say, that

the performance of the matched bootstrap procedure is acceptable. The CI coverage ratio is larger than 0.8, the CI is relatively narrow and standard error for each statistic is one of the smallest. It is not sensitive to the distribution of innovations and the bootstrap series are integer valued.

Generally one can say, that the results of the AR and matched 10 methods can be considered in general for parameter and confidence interval estimation, while the other methods are able to perform well only in some of the cases.

5. Volcanic eruptions

We used the existing and new methods for analysing the number of significant volcanic eruptions of the last 100 years. The dataset was compiled by the National Centers for Environmental Information of the US Government, it can be downloaded from [6] and contains the annual number of the largest volcanic eruptions from 1900 to 2018 (for the exact definition we refer to [6]). We analysed the dataset using the AR, parametric INAR (pINAR), CBB with Politis–White block size selection procedure and matched bootstrap using 10 and 1 size blocks. The estimated autoregression parameter, mean and standard deviation are presented in table 5.

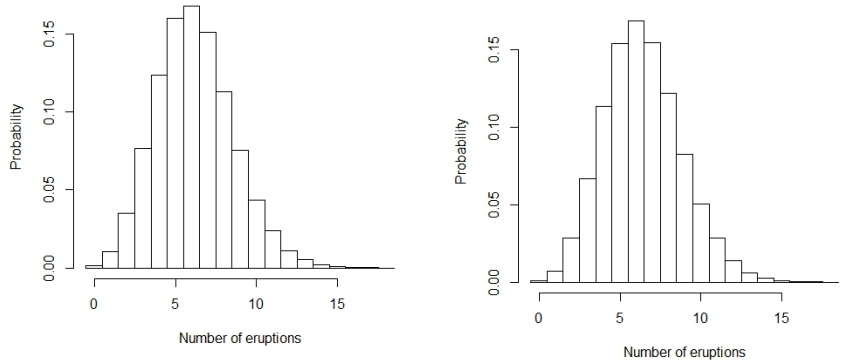
Method	α	mean	standard deviation
Original	0.245	4.18	2.439
AR	0.227 (0.05 - 0.41)	4.177 (3.64 - 4.73)	2.403 (2.03 - 2.8)
pINAR	0.227 (0.04 - 0.4)	4.174 (3.67 - 4.69)	2.019 (1.71 - 2.36)
CBB	0.151 (-0.03 - 0.33)	4.187 (3.67 - 4.77)	2.421 (1.98 - 2.87)
Matched 10	0.205 (0.02 - 0.36)	4.054 (3.37 - 4.79)	2.202 (1.82 - 2.56)
Matched 1	0.199 (-0.01 - 0.39)	4.044 (3.53 - 4.6)	2.21 (1.89 - 2.54)

Table 5: Estimated statistics for the time series of annual volcanic eruptions from 1900 to 2018. The estimates and 95% confidence intervals for the autocorrelation parameter, mean and standard deviation were calculated using different type of bootstrap techniques. The bootstrap estimations are based on 1000 simulations.

One can see, that the results of the AR method are close to the original estimated values. The pINAR underestimates the standard deviation, even the observed value is outside the confidence interval. The CBB underestimates the autocorrelation, so that the dependence is not even significant by this method. The confidence interval for α using matched 1 method also contains 0. In

contrast, matched 10 method result in acceptable estimates and confidence intervals for all of the statistics. It suggests smaller values than the estimates of the original sample, which might be explained by usually less eruptions, with some more intense periods. The confidence intervals are asymmetric in this case, as supposed to be for a process with non-normal distributed innovation structure.

Finally we present a prediction for year 2019 and compare it with the observed number of large eruptions. We estimated the average of the process by the mean of the sample. Using the estimated $\hat{\alpha}$ autoregression parameters from table 5 we calculated the $\hat{\lambda} = \overline{\mathbf{X}} \cdot (1 - \hat{\alpha})$ approximation for the Poisson distribution's parameter. This choice may be motivated by the fact that the innovations (new erupting volcanoes in the given year) are expected to fulfil the properties of the Poisson process. Then we simulated random values from the $Bin(X_{2018}, \hat{\alpha}) + Poi(\hat{\lambda})$ distribution. The histogram of marginal distribution for year 2019 can be seen on figure 1. The matched 10 and the AR bootstrap method both predicted 6 eruptions for year 2019 by using squared loss function. In fact, that year 7 large eruptions occurred, thus both method preformed well.



(a) Matched bootstrap with 10 sized blocks

(b) AR bootstrap

Figure 1: Monte Carlo simulation for the number of large eruptions in year 2019 based on the two best bootstrap technique. The histograms are based on 100 000 simulations.

6. Conclusion

We presented the matched bootstrap method for INAR(1) processes. This non-parametric block-based method selects new blocks fitting perfectly to the previous ones. The received time series has a structure similar to the original INAR observations, while the bootstrap estimates performs comparably to the previously known, best procedures as it was shown by our simulation study. As it is not sensitive for the innovation distribution, it can be used for a wide range of INAR(1) series without making assumptions.

A similar idea can be executed for classical block bootstrap. One can choose the upcoming block using a weighted distribution (kernel function) of the observations in order to rise the probability of choosing similar values as the end of the previous block. This procedure needs a deeper research to find the optimal weights and analyse the asymptotic and finite sample properties.

Unfortunately the generalization for higher order INAR processes is not evident. Trying to match more observations lowers the number of suitable blocks which makes the model less random. However, not claiming perfect matching, only similar values - based on a probabilistic decision as [5] - may extend the procedure for higher ordered processes as well.

Acknowledgement. We thank for the collection of the data for NOAA - National Centers of Environmental Informations.

References

[1] **Al-Osh, M. and A. Alzaid,** First-order integer-valued autoregressive (INAR(1)) processes, *J. Time Ser. Anal.*, **8** (1987), 261–275.

[2] **Alzaid, A. A. and M. Al-Osh,** An integer-valued pth-order autoregressive structure (INAR(p)) process, *Journal of Applied Probability*, **27** (1990), 314–324.

[3] **Bisaglia, L. and M. Gerolimetto,** Model-based INAR bootstrap for forecasting INAR(p) models, *Computational Statistics*, **34** (2019), 1815–1848.

[4] **Bühlmann, P.,** Sieve bootstrap for time series, *Bernoulli*, **3** (1997), 123–147.

[5] **Carlstein, E., Do, K., Hall, P., Hesterberg, T. and H. R. Künsch,** Matched-block bootstrap for dependent data, *Bernoulli*, **4** (1998), 305–328.

[6] **NOAA - National Centers of Environmental Informations**, Volcanic data and information, https://www.ngdc.noaa.gov/hazard/volcano.shtml, Accessed 4 February 2020., (2020)

[7] **Drost, F. C., van den Akker, R. and B. J. M. Werker,** Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **71** (2009), 467–485.

[8] **Efron, B.,** Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, **7** (1979), 1–26.

[9] **Jentsch, C. and C. Weiss,** Bootstrapping INAR models, *Bernoulli*, **25** (2019), 2359–2408.

[10] **Du J. and Li Y.,** The integer-valued autoregressive (INAR(p)) model, *Journal of Time Series Analysis*, **12** (1991), 129–142.

[11] **Kreiss, J.,** *Bootstrap Procedures for ar(∞)-process*, Springer, 1992.

[12] **Lahiri, S. N.,** Theoretical comparisons of block bootstrap methods, *The Annals of Statistics*, **27** (1999), 386–404.

[13] **Norris, J. R.,** *Markov Chains* (*Cambridge Series in Statistical and Probabilistic Mathematics*), Cambridge University Press, Cambridge, 1997.

[14] **Patton, A., Politis, D. N. and J. P. Romano,** Correction to "Automatic Block-Length Selection for the Dependent Bootstrap" by D. Politis and H. White, *Econometric Reviews*, **28** (2009), 372–375.

[15] **Politis, D. N. and J. P. Romano,** A circular block-resampling procedure for stationary data, in: R. Lepage and L. Billard (eds.) *Exploring the Limits of Bootstrap*, Wiley, New York (1992), 263–270.

[16] **Politis, D. N. and H. White,** Automatic block-length selection for the dependent bootstrap, *Econometric Reviews*, **23** (2004), 53–70.

[17] **Radovanov, B. and A. Marcikić,** A comparison of four different block bootstrap methods, *Croatian Operational Research Review*, **5** (2014), 189–202.

[18] **Steutel, F. W. and K. van Harn,** Discrete analogues of self-decomposability and stability, *Ann. Probab.*, **7** (1979), 893–899.

[19] **Weiss, C. H. and S. Schweer,** Bias corrections for moment estimators in Poisson INAR(1) and INARCH(1) processes, *Statist. Probab. Lett.*, **112** (2016), 124–130.

L. Németh and A. Zempléni
Department of Probability Theory and Statistics
Institute of Mathematics
Eötvös Loránd University
Budapest
Hungary
lnemeth@caesar.elte.hu
zempleni@caesar.elte.hu