Statistical Journal of the IAOS 39 (2023) 513–521
DOI 10.3233/SJI-220097
IOS Press

Automatizing model selection in an annual review of seasonal adjustment: A machine learning-inspired approach

Yingfu Xie
Statistics Sweden, Solna Strandvag 86, 17154 Solna
Tel.: +46 10 4794102; E-mail: Yingfu.Xie@scb.se

Abstract. In this paper, we bring to attention the problem of model selection with conflicting criteria in general and in annual reviews of seasonal adjustment in particular. Although partial concurrent seasonal adjustment and annual reviews are recommended by Eurostat, the problem of model selection in such reviews is seldom discussed in the literature, and our study is an attempt to fill this gap. In these reviews, revisions caused by model changes are very undesirable. The trade-off between different diagnostics, M- and Q-statistics, numbers of outliers, and revisions is hard to make to select the best model. In this study, a customary model selection procedure is described. Furthermore, we argue for using the manually chosen models as the "true" models, which makes it possible to employ a supervised machine learning-like approach to select weights for these diagnostics. It shows that this approach could work equally well as (if not better than) human statisticians, and thus facilitates an automatized procedure for model selection in such annual reviews. Although the approach has limitations as we describe, it is, to our best knowledge, the first study of its kind in the literature.

Keywords: Machine learning, statistical learning, partial concurrent, seasonal adjustment

1. Introduction

Seasonal adjustment is a fundamental step in the production of official statistics, which identifies and removes the seasonal fluctuations and calendar effects, such that the decision-makers and analysts can better understand the short and long-term movements in the time series [1]. The methodologies for seasonal adjustment have been developed constantly and the literature is huge, but the number of methods used in producing official statistics is quite limited. Much earlier works, such as the development of the US Census X-11 program by Shiskin et al. [2] and Cleveland and Tiao [3], can be found in the review by Pierce [4]. Later, Gómez and Maravall [5] proposed a parametric approach (called TRAMO-SEATS) based on the ARIMA (autoregressive integrated moving average) family, which has been successively adopted by the Census X-12-ARIMA and X-13a-Seats programs (cf. [6] and [7]). Although there are other seasonal adjustment methods using for example structural time series models ([8]), those methods cannot produce equally interpretable seasonally adjusted results and have never gained the same popularity in the official statistics as the X-12-ARIMA or the TRAMO-SEATS programs.

In the practice of seasonal adjustment in a national statistical office (NSO), the question is very common how often the practitioners may review the ARIMA models used in seasonal adjustment. The possible options are described in the ESS guidelines on Seasonal Adjustment [1]. Thereby, Eurostat recommends a partial concurrent revision policy, i.e., the models are to be re-identified once a year while the coefficients of the models may be updated instantly.

There is one important difference in such annual reviews compared with other occasions for model selection. In those reviews, the revisions caused by the change of models are highly undesired by both users of statistics and subject staff. For example, an intern guideline from the US Census Bureau described

that "... large revisions may damage the Census Bureau's credibility for producing high-quality data products" [9]. Consequently, we cannot rely only on statistical diagnostics at hand for model selection and a trade-off often has to be made between, for instance, a better fitness of a model and a smaller revision. As we will show in Section 2, it is not easy to make such trade-offs and it depends sometimes on experiences or even on personal preference.

The standard automatic model selection procedure implemented in X-12-ARIMA [10] or TRAMO-SEATS [11], relies heavily on the Akaike information criterion (AIC) [12] or its variants such as Hannan and Quinn criterion or the Bayesian Information Criterion (BIC) [10, Section 5.5]. However, the AIC and the related measures are less applicable in the annual reviews since they do not consider the revision problem. Besides, these measures are based on the maximum likelihoods of an underlying model and thus strongly affected by the number of outliers identified with that model. Incorporating more outliers normally improves the goodness of fitness of a model for the analyzed time series by increasing the likelihood and reducing the AIC or BIC. At the same time, a model is usually considered to be bad if it includes too many outliers, although there is no solid critical value for the acceptable number of outliers. Hence, solely using AIC or BIC to compare models with different numbers of outliers will be misleading in most cases and should be avoided ([10, p. 48]).

In this paper, a customary procedure for such an annual review applied in the seasonal adjustment procedures at Statistics Sweden is described, where the main statistical diagnostics, including revision errors, are listed and evaluated. The problem of model selection becomes obvious when there are no criteria for the trade-off between different diagnostics and/or revisions. To solve the problem, we propose to use the previous human-chosen models as the "true" ones. Following this proposal, we can rank these diagnostics and employ an approach like a supervised machine learning (ML) algorithm (cf. [13] and [14]) to select "optimal" weights among different diagnostics. To our best knowledge, this is the first study of this kind in the literature. Despite some limitations of this ML-inspired approach, we are going to show that this approach can work well, which makes it possible for us to build an automatic procedure for model selection, based on the obtained weights in the annual reviews of seasonal adjustment.

The remainder of this paper is organized as follows. Section 2 recalls the background of the partial concurrent revision policy of seasonal adjustment and introduces a customary procedure for model selection in annual reviews. The problem of how to select models and trade-offs between different diagnostics is discussed. In Section 3, the implementation and the result of an ML-inspired approach are presented. The paper ends with discussions and final remarks in Section 4.

2. Model selection in annual reviews

In the mainstream seasonal adjustment programs X-12-ARIMA [9] and TRAMO-SEATS [5], ARIMA models are used in the pretreatment step to handle missing values, forecasting, and backcasting to extend time series, deterministic exogenous variables such as calendar adjustment variables, and outliers. Ensuring the high quality of these ARIMA models is very important for the total quality of the seasonal adjustment. Therefore, it is a common question how often the practitioners in an NSO should review and eventually change the ARIMA models. The ESS Guidelines on Seasonal Adjustment [1, Section 4.2] lists four possible revision policies: the current adjustment, the concurrent adjustment, the partial concurrent adjustment, and the controlled current adjustment. Please see [1] for the details and discussions. The alternative recommended by [1] in an ordinary statistical production is the partial concurrent revision policy, that is to say, "the model, filters, outliers and calendar regressors are re-identified once a year and the respective parameters and factors re-estimated every time new or revised data become available" [1]. This revision policy is believed to be able to balance the accuracy of the models and the reduced revisions caused by model changes.

Although the annual review is a common and important step in the practice of seasonal adjustment for NSOs, the procedure and the methodology of model selections in such reviews are seldom discussed. There are works dedicated to specific diagnostics such as sliding spans [15] or the adequacy of seasonal adjustment [16], to the comparison of direct and indirect adjustment (cf. works in [17]), or some intern guidelines such as [9]. However, no research is available in the literature for model selections in annual reviews. Our study is an attempt to bring it to attention and attract more contributions in this area.

A customary review procedure usually begins with summarizing different statistical diagnostics (probably with different statistical programs). In this study, we use the Swedish Production Value Index (PVI) as an

Table 1
An example with summarized diagnostics and revision errors for Industry 20+21 (the statistics in the parentheses)

Industry	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21
Model ID	Current used	Model 1	Model 2	Model 3	Model 4	Model 5
Transform	ADD	MULT	MULT	MULT	MULT	MULT
Constant	0	0	0	0	0	0
P	0	0	0	1	2	1
D	1	1	1	1	1	1
Q	1	1	2	1	1	2
BP	0	0	0	0	0	0
BD	1	1	1	1	1	1
BQ	1	1	1	1	1	1
Season	Present	Present	Present	Present	Present	Present
Skewness	Not normal (1.88)	Normal (0.45)	Normal (0.5)	Normal (0.5)	Not normal (0.64)	Not normal (0.64)
Gearys_a	Not normal (0.65)	Normal (0.79)	Normal (0.77)	Normal (0.78)	Normal (0.77)	Normal (0.78)
Kurtosis	Not normal (9.47)	Normal (3.21)	Normal (3.64)	Normal (3.64)	Normal (4.03)	Normal (3.44)
ACF	Not Ok (0.0059)	Ok (0.477)	Ok (0.3743)	Ok (0.3411)	Ok (0.1563)	Ok (0.1002)
Q1	Not Ok (1.182)	Not Ok (1.255)	Not Ok (1.215)	Not Ok (1.209)	Not Ok (1.206)	Not Ok (1.181)
Q2	Not Ok (1.262)	Not Ok (1.296)	Not Ok (1.276)	Not Ok (1.27)	Not Ok (1.247)	Not Ok (1.236)
BIC	1080.89	1027.20	1036.49	1037.58	1032.18	1015.68
N_outliers	3	3	2	2	3	5
Outliers	TC01FEB2012	AO01FEB2012	AO01FEB2012	AO01FEB2012	TC01FEB2012	TC01FEB2012
	TC01MAY2012	LS01DEC2018	LS01DEC2018	LS01DEC2018	TC01MAY2012	TC01MAY2012
	AO01SEP2020	AO01SEP2019			LS01DEC2018	AO01SEP2018
						LS01DEC2018
						AO01SEP2020
rev1	0.00	0.06	0.01	0.01	0.01	0.05
rev2	0.00	2.38	1.84	1.86	1.95	1.96
rev3	0.00%	2.00%	1.58%	1.59%	1.68%	1.65%
RSF	Ok (0.52)	Ok (0.34)	Ok (0.49)	Ok (0.49)	Ok (0.53)	Ok (0.68)

example for illustration, but the analysis could be generalized to many other types of data. Table 1 is a summary of the NACE-industry 20+21 (Manufacture of chemical products and Manufacture of pharmaceutical products, NACE: Statistical Classification of Economic Activities in the European Community), from the currently used model and five other models (summaries for other industries are omitted in Table 1 for the sake of space). The first 9 rows define roughly an ARIMA model, and the other rows are for different diagnostics from the corresponding model. The diagnostics demonstrated here include those from the test of the existence of seasonality (row "Season"), the normality test ("Skewness", "Gearys-a", "Kurtosis"), the residual autocorrelation ("ACF"), summarized M-statistics ("Q1", "Q2"), BIC, the number of outliers, the outliers, the revision errors compared with the currently used models ("rev1", "rev2", "rev3"), and the test of the existence of residual seasonality ("RSF"). We refer to [10] the detailed descriptions of these diagnostics. For the revision errors, there are three different measures calculated. Denote $\{Y_t^0\}, t = 1, \ldots, N$ as the seasonally adjusted data of one series using the currently used model, and analogy $\{Y_t^i\}$ using the model i, the revision errors are defined respectively as $rev_1 = \frac{1}{N}\sum_{i=1}^{N}(Y_t^i - Y_t^0)$, $rev_2 = \frac{1}{N}\sum_{i=1}^{N}|Y_t^i - Y_t^0|$, and $rev_3 = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_t^i - Y_t^0}{Y_t^0}\right|$.

Some explanations and remarks are necessary for the content in Table 1, as listed below.

1) How to define a unique model is not trivial. In Table 1, the trend (Henderson) and seasonal filters are not included but are chosen automatically with default criteria [6] for different models. Otherwise, models with the same ARIMA parameters but different filters should reasonably be considered as different models. Similar reasoning applies to models with the same ARIMA parameters but different outliers; such models are treated as different models in this study.
2) Obviously, there are other possible statistical diagnostics. It is not our intention to list out all possible diagnostics in the study but to illustrate a basic working procedure.
3) It is neither our intention to give an optimal balance for different types of diagnostics in the summary. For example, there are 3 different normality tests and 3 revision error measures but only one BIC.

4) Six models are included in Table 1 for the Industry NACE 20+21, including the currently used model for the series and the 5 best models identified by the X-12-ARIMA program [6]. For other industries, an additional Airline model may be added if it is not already included. The best model in most cases is among these candidate models. However, it should be right straightforward to increase the number of candidate models if needed.

The second step in the procedure involves an inspection of the different models with different diagnostics and trying to find the best model. However, the main problem here is how to select the best model when different diagnostics conflict with each other. The normal automatic model selection procedure in X-12-ARIMA [6] and TRAMO/SEATS [5] relies heavily on the BIC measure. As we can see from Table 1, the BIC measure is not readily applicable in this case. For example, Model 5 has the smallest BIC measure for this series, however, there are 5 outliers identified compared with 2 or 3 outliers for other models. It is not strange that the BIC for Model 5 should be lower. Besides, there are no criteria for the trade-off between BIC and the revision errors; for instance, one can ask how much decrease of the BIC measure is acceptable for a 1% less revision of Rev2. The same issue applies also to the trade-off of other diagnostics.

The occurrence of outliers not only affects the comparison of BIC but also other diagnostics such as the normality test and the revision errors. Our experience shows that it is, above all, the occurrence or change of outliers that causes the most significant revisions. The interaction between BIC, outliers, and revisions further complicates the model selection problem. The default critical value for outlier identification in X-12-ARIMA [10] is developed from a simulation study [18] and depends on the length of the series, while TRAMO/SEATS [3] uses 3.5 as the default critical value for outlier identification, and no more than 5% of the number of observations as guiding principle for the number of outliers allowed in a series. Please see [19] for discussions of the choice of critical values. The question of what critical value should be used for outlier identification and how many outliers should be allowed in a series is in its own right an issue that deserves more research.

As we have seen, in many cases there are no readily available statistical criteria to guide the trade-offs between different models. The model selection in annual reviews is not only a science but also an art that depends on personal experience or preference. In Statistics Sweden, there is a conservative principle applied in model selections, saying that if there is no better model available do not change the current model, which is in the same spirit as [9], to reduce unnecessary revisions. In other cases, a practical convention is applied to guide the comparison of the importance of different diagnostics when there are conflicts. Among them is

$$\begin{aligned} &\textit{RSF} \geqslant \textit{Residual Normality} \geqslant \textit{Revisions} \geqslant \\ &M(Q)\textit{Statistics} \geqslant \textit{BIC} \geqslant \textit{Number of outliers} \geqslant \\ &\textit{Existence of Seasonality} \geqslant \textit{ACF} \end{aligned} \tag{1}$$

The guideline of Eq. (1) is based on our experience and preference. In the US Census Guideline [9], the emphasis is put to eliminate the RSF and to minimize revisions, as well as a sensible choice of outliers, which is basically in the same line as Eq. (1). The Guideline [9] doesn't cover other diagnostics.

3. Automatizing and a machine learning-inspired approach

Without clear guidance, there will be personal dependence problems in model selections and the reproducibility of the output cannot be guaranteed. Besides, it will lead to higher demand for time and human resources for the model selection in annual reviews. After all, there are thousands of seasonally adjusted time series from the Swedish official statistics system that require an annual review. To streamline the annual reviews and mitigate the aforementioned problems, one possible way would be to standardize and automatize the procedure of model selection. We explore the possibility below.

3.1. Ranking the diagnostics

One could rank the diagnostics first to facilitate the comparison of different models. For example, having ranked tied values with the mean of their ranks, the ranks for the diagnostics in Table 1 are shown in Table 2. The number of outliers and the outliers are excluded from Table 2, partly because it is unclear how to rank them, and partly as we mentioned before, the most effects of change of outliers should be covered by the revision errors. We rank the BIC measures differently only when the measures differ by not less than 2, a convention implemented in X-12-ARIMA [10]. A similar convention is applied to Rev1, Rev2, and Rev3, and they are ranked differently when their values differ by at

Table 2
The rank of the diagnostics in Table 1 (except the outliers and number of outliers)

Industry	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21	NACE_20_21
Model ID	Current used	Model 1	Model 2	Model 3	Model 4	Model 5
Season	3.5	3.5	3.5	3.5	3.5	3.5
Skew	5	2	2	2	5	5
Gearys_a	6	3	3	3	3	3
Kurtosis	6	3	3	3	3	3
ACF	6	3	3	3	3	3
Q1	3.5	3.5	3.5	3.5	3.5	3.5
Q2	3.5	3.5	3.5	3.5	3.5	3.5
BIC	6	2	4.5	4.5	3	1
rev1	1.5	6	1.5	4	3	5
rev2	1	6	3.5	3.5	3.5	3.5
rev3	3	6	3	3	3	3
RSF	3.5	3.5	3.5	3.5	3.5	3.5

least 5% (or at least 0.05 when compared with zero for Rev1 and Rev2, and at least 0.005 for Rev3). Please observe that ranking continuous variables this way could lead to a transitivity problem. Non-transitivity is assumed in our study. The discussion of transitivity is out of the scope of this paper and the readers with interest are referred to decision theory [20] and references therein.

3.2. True models?

In the next course of action, we could weigh different diagnostics and obtain a summarized rank for different models. However, an unavoidable obstacle is that we do not have the true models to evaluate the best-ranked models. Simulation studies could be a possible way to walk around the obstacle, but a simulation study would significantly alter the character of the problem in question. In our study, we are instead inspired by the rapid development of ML methods [13]. For example, to train a supervised ML algorithm, an important prerequisite is the labeled true targets. It is towards the true targets that the machine is to tune its models and parameters and learn from the data. Analogously, we propose here to use the human-chosen models as the "true targets", and then to find optimal weights to weight different diagnostics such that the chosen models from this approach would be as close to the manually chosen models as possible. We call it an ML-inspired approach. There are two arguments for our proposal. First, those models are the best possible models available at hand. More importantly, our main purpose is not to find the optimal models, which we, unfortunately, do not have the answer to, but rather to explore the possibility to standardize and automatize our working procedure for model selection in annual reviews.

Needless to say, the eventual best weights can only apply to this data set, as we have to re-train the weights for other data, which is the case for almost all ML algorithms.

3.3. Settings to train the weights

Our goal of this ML-inspired approach is thus to find the weights $\{w_i\}$ for different diagnostics in Table 2, such that as many of the best-ranked models are the same as the manually chosen models as possible. That is to say, the accuracy metric is defined as

$$Acc(w_i) = \frac{\#(\textit{Industries that the approach chose the same model as human})}{\textit{total number of industries}},$$

and our goal is to find the optimal weights $\{w_i^*\}, i = 1, \cdots, 12$, such that

$$\{w_i^*\} = \textit{argmax}_{w_i}(Acc).$$

Our data for PVI consists of 87 NACE industries or main industrial groups. We randomly split the data, about 75% of the total or 66 industries, as the train data, and the remaining as the test data set.

For the 12 diagnostics in Table 2, a loop is designed to assign values 1, 41 and 81 in turn to each diagnostic measure. The weight for diagnostic $i, i = 1, \cdots, 12$, is thus $w_i = \frac{K_i}{\sum_{j=1}^{12} K_j}$, where $K_i = \{1, 41, 81\}$. The range of weights for each diagnostic is from 0.11% $\left(\approx \frac{1}{11*81+1}\right)$ to 88% $\left(\approx \frac{81}{11*1+81}\right)$, which we believe is a reasonable range. The number of loops is $3^{12} = 531441$, which is moderate. We use the SAS® software for this study and an increase to $4^{12} = 16777216$ loops would cause some capacity problems. Please note that the number of loops is roughly equal to the number of combinations of different weights and should be enough for our purpose.

Table 3
One set of the best weights

BIC	rev1	rev2	rev3	Season	Skewness	Gearys	kurtosis	Q1	Q2	ACF	RSF
0.27%	0.27%	11.02%	11.02%	0.27%	0.27%	11.02%	11.02%	21.77%	0.27%	11.02%	21.77%

Table 4
Series with mismatched models using the weights from Table 3 and comments

NACE	Manually chosen		Chosen by ML-approach	
	Model	Comments	Model	Comments
10+12	Model1	BIC: 695.5. 0 outlier. Revision.	Current	ACF: not Ok. BIC: 686.9. 1 outlier.
16.2	Model1	BIC: 835. 0 outliers. Revision.	Current	BIC: 830.7. 1 outlier.
17.11	Model1	BIC: 878.9. 0 outlier. Revision.	Current	BIC: 883.6. 0 outlier.
19	Model1	No season. BIC: 979.7. 2 outliers. Revision.	Current	No season. BIC: 991.4. 3 outliers.
20+21	Model1	BIC: 1027.2. 3 outliers. Higher Rev.	Model2	BIC: 1036.5. 2 outliers.
21	Model1	BIC: 1136.5. 3 outliers. Higher Rev.	Model2,3,4	BIC around 1145. 2 outliers.
27	Model1	BIC: 851.8. 2 outliers. ACF not Ok.	Model4	BIC: 860.6. 2 outliers. Similar Rev.
29+30	Model1	BIC: 865.7. 3 outliers. Revision.	Current	BIC: 975.2. 2 outliers.
30	Model2	BIC: 963.4. 9 outliers. Better Rev1.	Model5	BIC: 927.5. 11 outliers. Better Rev2.
32	Model1	BIC: 837. 2 outliers. Revision.	Current	BIC: 840.6. 2 different outliers.
36–39	Current	BIC: 821.5. 0 outlier. ACF not Ok.	Model2	BIC: 815.2. 1 outlier. ACF Ok. Rev.
51	Current	BIC: 781.7. 3 outliers.	Model1	BIC: 781.7. 3 outliers (1 different type). Revision.
58	Model1	BIC: 977.6. 4 outliers. Revision.	Current	BIC: 991.8. 3 outliers.
85	Model1	BIC: 698.4. 2 outliers. Revision.	Current	BIC: 714.8. 0 outlier.
90–93	Airline	BIC: 822.5. 1 outlier. Revision.	Current	BIC: 806.2. 3 outliers. Skewness and Q1 not Ok.
94–96	Model2	BIC: 761.7. 4 outliers. Revision.	Current	BIC: 765.3. 4 outliers (1 different). Gearys-a Test not Ok.
90–96	Model1	BIC: 709. 4 outliers. Revision.	Current	BIC: 720.4. 4 outliers (1 different).

Note also that we have not tried to balance the uneven numbers of different types of diagnostics (e.g., 3 normality tests vs 1 BIC) in this approach. Once again, it is because we are not attempting to find the optimal model, but to search for the optimal weights $\{w_i^*\}$ for the diagnostics.

3.4. The result and discussions

3.4.1. The result with discussions

For the train data set, the highest accuracy score is 84.8%. I.e., for 56 series over 66 in the train data, the best weights yield a model that is the same as the manually chosen. Applying those best weights to the test data set, the average accuracy score is 73%, which indicates overfitting to some degree. The average score for the whole data set is about 82.8%. We also considered a benchmark scenario where all 12 diagnostics have equal weight. In this scenario, the accuracy score is about 69% for the whole data set.

The best accuracy score (about 83%) is to some extent lower than our expectation. This score seems not to be much higher than the benchmark scenario (69%). However, it is worth noting that all diagnostics have been ranked in the beginning. If there is no conflict among different diagnostics for a time series, it would be easy to pick up the best model by simply choosing the model with the best-ranked diagnostics. It explains the relatively high score of the benchmark scenario. It is undoubtedly of greater interest for us to study cases with conflicting diagnostics.

To better understand this ML-inspired approach, using the chosen weights shown in Table 3, we list all series (industries) with mismatched models in Table 4.

In Table 4, in addition to the chosen models for each series, we include those diagnostics that are different for the competing models, hoping that they might explain why the outcome from the ML-inspired algorithm differs from the manually chosen one. Other diagnostics are omitted for the sake of space. From Table 4, some observations can be made.

I) The ML-inspired approach works pretty well. In almost all of the 17 series in Table 4, there are no clearly outperforming models, even if we inspect them closer afterward. In those cases, difficult trade-offs have to be made between BIC and revision errors, both of which are mainly caused by different (numbers of) outliers. Only in one case for NACE 51, we may prefer the manually chosen model (given that the change of outlier type in the model chosen by the approach is not justified; see III below). In all other cases, we see no reason why we cannot use the models chosen by the ML-inspired approach.
In addition, we carried out a Kolmogorov-Smirnov test [21] on the empirical distribution

functions (EDFs) of forecasting errors from the two competing models. Our main study was based on the Swedish PVI data until September 2021. Following [22], one-step ahead forecasting errors between October 2021 and December 2022 were calculated and the absolute values are used as the samples for the test. In none of the 17 industries in Table 4, the null hypothesis that the two EDFs are from the same sample distribution can be rejected at a significance level of 0.05. As for the one-sided Kolmogorov-Smirnov test, while 10 human-chosen models have higher absolute mean forecast errors among the 17 industries, only in one case (Industry 90–96) the difference is significant at p-value 0.038 (the human-chosen model has smaller forecasting error). The Kolmogorov-Smirnov test confirms the complexity of the model selection problem and the fact that the ML-inspired approach yields a satisfactory result.

II) Interestingly, it seems that the approach complies with the conservative principle and the convention (Eq. (1)) better than human statisticians. The result from this approach suggests keeping the currently used models for 78 series, while in the manual review, currently used models are kept for 68 series. It implies that the algorithm seems to act more consistently than human beings.

III) We emphasize once again that it is difficult to handle outliers in such reviews. Except for one series (NACE 30), the numbers of outliers are less than 5% of the number of observations and might be regarded as reasonable for cases in Table 4. In 6 series in Table 4, the manually chosen models yield fewer outliers, while 5 series yield more. Still, the slight changes in the number of outliers have led to the model selection problem. In this study, all the outliers are significant for the default critical values in the X-12-ARIMA program [10]. Other aspects of the plausibility of the outliers are nevertheless not evaluated. It would be interesting to incorporate outlier treatment into this ML-inspired approach. We leave it for future work.

3.4.2. Drawbacks of the approach

One problem discovered with this approach is that there are no unique solutions, i.e., there are several sets of weights that yielded the same highest score. The problem is probably not so surprising with consideration of the relatively small sample size (87 industries and about 530 models). Besides, our accuracy metric $Acc(w_i)$ is actually discrete with the finite sample, which causes non-unique solutions. We expect that it will not be an issue if we for example could increase the sample size to more than 1000 time series and more than 6000 models.

Another problem with the result is that some diagnostics, such as the test of the existence of seasonality in the original series ("Season") or the test of the existence of residual seasonality ("RSF"), are insensitive. This may seem surprising at first glance since the test of residual seasonality is a highly prioritized measure (see Eq. (1)). Recall nevertheless that we are comparing different models given the data. All competing models may have similar performance regarding some diagnostics depending on the properties of the input time series. For example, a re-investigation shows that there are only 2 models of the total of about 530 models that have failed the test of the existence of residual seasonality. For the example (industry NACE 20+21) shown in Table 2, many diagnostics such as Season, Q1, Q2, and RSF do have the same rank among all the 6 competing models. The insensitivity of some diagnostics compounds the aforementioned problem with non-unique solutions.

It is noteworthy that non-unique solutions will not necessarily lead to problems. As we have noticed, with the other sets of weights (not reported) that achieved the highest score of $Acc(w_i)$, the mismatched series are almost the same as those in Table 4 in our case.

4. Conclusions and final remarks

In this paper, we took up the problem of model selection in annual reviews, which is an important step to guarantee the quality of seasonal adjustment with a partial concurrent revision policy recommended by Eurostat [1]. We introduced a customary procedure for this task. It was illustrated that without statistical criteria for a trade-off among different statistical diagnostics, not least between the BIC measure and revision errors, the model selection could be difficult, personally dependent, and time-demanding. This issue is common for general model selection problems when there is no single dominant criterion but instead conflicting diagnostics. It would be interesting to see more work in this area, which is absent to our best knowledge.

In order to mitigate the problems, we attempted to standardize and automatize the customary working pro-

cedure. For this purpose, we proposed to use the human-chosen models as the "true" ones and to search for the "optimal" weights such that the best-ranked models, after weighting the diagnostics with those weights, would be as close to the manually chosen models as possible. This proposal overcame the issue that there were no true models for the evaluation and was inspired by supervised ML algorithms.

Our study showed that this ML-inspired approach worked pretty well. For the cases in which the approach chose different models than human statisticians, the approach-chosen models were equally well with respect to the diagnostics and the Kolmogorov-Smirnov test of forecasting errors. Interestingly, the approach seemed to comply with the practical, implicit praxis (Eq. (1)) more consistently than human statisticians. Following the result, we do not see an obstacle to using the weights chosen by this ML-inspired approach and automatizing the procedure for model selection.

As one would expect, this ML-inspired approach has its limitations. In general, ML algorithms cannot give satisfactory results for small data sets with no clear pattern, not least for time series data. Similar to many other ML algorithms, our results obtained for this particular data set cannot be readily generalized; for other input data, the training process must be redone and the resulting weights will be generally different.

For this particular study, there is plenty of room for improvement. A drawback of our result is that the solution is non-unique. Some measures can be taken to mitigate the problem with non-unique solutions. So far only a limited number of candidate models, 6 to 7 models for every series are evaluated, and including more ARIMA models in such studies should be straightforward if it will not cause a capacity problem. Similarly, the sample size could be increased if possible. Otherwise, one could employ some resampling methods to increase the sample size. More statistical diagnostics can be taken in, and some insensitive diagnostics could be excluded from the evaluation beforehand, after an inspection. In addition, the design of weights and the loop in this study are at the elementary level and can be refined as well.

Despite the limitations and the drawbacks, our study is, to our best knowledge, the first one in the literature to propose using the manually chosen models as the "true" ones, in order to be able to apply ML-like approaches. Furthermore, we have shown an example of how machine-learning approaches could improve statistical learning and facilitate official statistical productions. We hope our study will inspire more coming work in this direction.

Acknowledgments

We are very grateful to the two anonymous referees for their insightful comments, which greatly improved the readability of this paper. We thank Dr. Elezović at Statistics Sweden for the discussions and valuable comments.

References

[1] Eurostat. ESS Guideline on Seasonal Adjustment. 2015. Available at: https://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf [accessed 24 January 2023].

[2] Shiskin J, Young AH, Musgrave JC. The X-11 Variant of the Census Method-II Seasonal Adjustment Program. Technical Paper No. 15. 1967. U.S. Bureau of the Census.

[3] Cleveland WP, Tiao GC. Decomposition of seasonal time series: A model for the census X-11 program. Journal of the American Statistical Association. 1976; 71: 581–587.

[4] Pierce DA. A survey of recent developments in seasonal adjustment. The American Statistician. 1980; 34(3): 125–134.

[5] Gómez V, Maravall A. Automatic Modeling Methods for Univariate Time Series. In A Course in Time Series, R.S. Tsay, D. Pena, and G.C. Tiao (eds). New York: John Wiley, 2000. pp. 171–201.

[6] Findley DF, Monsell B, Bell WR, Otto MC, Chen B-C. New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program. Journal of Business & Economic Statistics. 1998; 16(2): 127–152. doi: 10.1080/07350015.1998.10524743.

[7] Findley DF. Some recent developments and directions in seasonal adjustment. Journal of Official Statistics. 2005; 21(2): 343–365.

[8] Aston JAD, Koopman SJ. An Implementation of Component Models for Seasonal Adjustment Using the SsfPack Software Module of Ox. Proceedings of the 13th Federal Forecasters Conference. 2003. pp. 64–71.

[9] McDonald-Johnson KM, Monsell B, Fescina R, Feldpausch R, Hood CCH, Wroblewski M. Seasonal Adjustment Diagnostics: Census Bureau Guideline, version 1.1. The US Census Bureau. 2010. Available at: http://www.census.gov/content/dam/Census/library/working-papers/2010/adrm/G18-0-v1-1-Seasonal-Adjustment.pdf [Accessed 3 February 2023].

[10] U.S. Census Bureau. X-13ARIMA-SEATS Reference manual. (Version 1.1); 2017. Available at: https://www2.census.gov/software/x-13arima-seats/x-13-data/documentation/docx13as.pdf [accessed 24 January 2023].

[11] Maravall A. An application of the TRAMO-SEATS automatic procedure; Direct versus indirect adjustment. Computational Statistics & Data Analysis. 2005; 50(9): 2167–2190. doi: 10.1016/j.csda.2005.07.006.

[12] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19(6): 716–723. doi: 10.1109/TAC.1974.1100705.

[13] Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE. 2018; 13(3): e0194889. doi: 10.1371/journal.pone.0194889.

[14] Bruce Ratner. Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Third Edition. Chapman & Hall; 2017.

[15] Findley DF, Monsell B, Shulman HB, Pugh MG. Sliding-spans diagnostics for seasonal and related adjustments. Journal of the American Statistical Association. 1990; 85(410): 345–355. doi: 10.1080/01621459.1990.10476207.
[16] McElroy T, Roy A. A review of seasonal adjustment. International Statistical Review. 2022; 90(2): 259–284. doi: 10.1111/insr.12482.
[17] Manna M, Peronaci R. (eds) Seasonal Adjustment. European Central Bank: Frankfurt am Main. 2003.
[18] Ljung GM. On outlier detection in time series. Journal of the Royal Statistical Society. Series B. 1993; 55(2): 559–567.
[19] McDonald-Johnson KM, Hood CC. Outlier Selection for RegARIMA Modeling. Proceedings of the American Statistical Association, Section of Business and Economic Statistics. [CD-ROM Paper No. 00438]. Alexandria: American Statistical Association. 2001.
[20] Hansson SO. Decision Theory: A brief introduction. Textbook for Uppsala University. 1994. Available at https://www.researchgate.net/publication/210642121_Decision_Theory_A_Brief_Introduction [accessed 24 January 2023].
[21] Gibbons JD, Chakraborti S. Nonparametric Statistical Inference. 5th ed. New York: Chapman & Hall. 2010.
[22] Hassani H, Silva ES. A kolmogorov-smirnov based test for comparing the predictive accuracy of two sets of forecasts. Econometrics. 2015; 3: 590–609.