RESEARCH REPORT SERIES
(Statistics #2017-02)

Maximum Entropy Extreme-Value Seasonal Adjustment

Tucker McElroy
Richard Penny [1]

[1]Statistics New Zealand

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: January 11, 2017

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Maximum Entropy Extreme-Value Seasonal Adjustment

Tucker McElroy*and Richard Penny†

Abstract

Some economic series in small economies exhibit meager (i.e., non-positive) values, as well as seasonal extremes. For example, agricultural variables in countries with a distinct growing season may exhibit both of these features. Multiplicative seasonal adjustment typically utilizes a logarithmic transformation, but the meager values make this impossible, while the extremes engender huge distortions that render seasonal adjustments unacceptable. To account for these features we propose a new method of extreme-value adjustment based on the maximum entropy principle, which results in replacement of the meager values and extremes by optimal projections that utilize information from the available time series dynamics. This facilitates multiplicative seasonal adjustment. The method is illustrated on New Zealand agricultural series.

Keywords. Anomalies, Extremes, Seasonality, Seasonal Adjustment, Signal Extraction.

Disclaimer This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau or Statistics New Zealand.

1 Introduction

Economic time series may exhibit negligible or zero values during periods of inactivity; such series can arise in industries that are weather-contingent, such that in certain seasons little or no activity occurs. This behavior can also arise in smaller economies, where activity is reduced during certain times of the year. Moreover, non-positive values, referred to as *meager* values henceforth, can arise due to accounting constraints. Seasonal extremes are a second type of possible feature. Both meager values and extremes create a challenge for multiplicative seasonal adjustment methods such as the log-additive procedures, which rely upon log transformations, or the multiplicative procedures, which rely upon division. Extremes exert a tremendous impact on seasonal adjustments. The

*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, tucker.s.mcelroy@census.gov

†Statistical Methods Division, Statistics New Zealand, 401 Madras Street, Christchurch, New Zealand 8053, richard.penny@stats.govt.nz

topic of this paper is extreme-value adjustment and seasonal adjustment of time series with extremes and/or meager values, and we focus on some examples published by Statistics New Zealand (StatsNZ).

StatsNZ, like all National Statistics Offices (NSO), produces a large number of time series as part of its work. Many are published, but there are also time series that are used for internal analytical purposes, or are aggregated to produce published outputs. For series with seasonal variation, StatsNZ produces seasonally adjusted time series using the X-12-ARIMA (X12) methodology (Findley, Monsell, Bell, Otto, and Chen (1998)) of the X-13ARIMA-SEATS software (U.S. Census Bureau, 2011). For StatsNZ, the X12 methodology typically works well in production: the method is robust (with respect to different data structures), computationally stable, and capable of accommodating the volatility of New Zealand time series.

However, StatsNZ has a small number of anomalous series that X12 does not model well. New Zealand relies heavily on agricultural production in its economy, and some of the series exhibit meager values. Indeed, some quarterly agricultural series have between one and three quarters where the vast majority of the annual output occurs, with the other quarters having negligible activity. The Apple exports series (middle panel of Figure 1) has negative values in 1988, 1990, 1991 and 1992, while the Avocado exports series (left panel of Figure 1) has two quarters with value zero (i.e., no activity). Moreover, the smaller quarterly values tend to occur in the same Winter quarters as these zero values. In both these cases, the multiplicative seasonal adjustment is not possible, and additive adjustment can result in the undesirable feature of negative seasonally adjusted values. (While the input series can have small negative values due to accounting reasons, negative seasonally adjusted values have no economic validity, especially for those quarters with significant input values (e.g., second quarter 1998 in middle panel of Figure 1).) Even in cases where no zeros or negative values are present, small or large extremes can have an undue influence on the dynamics of the multiplicatively seasonally adjusted outputs, as in the case of the Berry exports series (right panel of Figure 1).

Although for some extremal time series the additive X12 method is feasible, typically the multiplicative X12 method is preferable for StatsNZ series; this is because the substantive trend and seasonal effects manifest in a multiplicative fashion. Moreover, as a matter of interpretation of seasonal and nonseasonal components, many economists believe that seasonal adjustment should be multiplicative. Even in cases where multiplicative X12 can be applied, the impact of extremes can lead to unacceptable seasonal adjustments. Resolving the impact of such effects is the subject of this paper. We briefly review some previously explored solutions that are defective; this is contrasted with our proposed methodology.

For series with meager values one possible approach is to additively shift the data upwards away from negatives and zeroes, by adding a small constant, and then proceeding with the log transformation. However, this will not yield well-defined components (discussed in Section 3.3

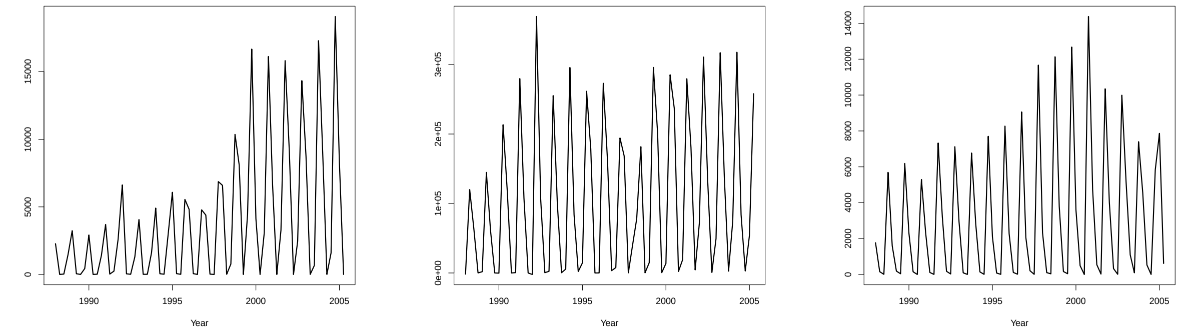

Figure 1: Left panel: Avocado Exports series. Middle panel: Apple Exports series. Right panel: Berry Exports series.

below), and moreover the results become contingent on the constant's value (one must remove the constant at the end of the process). Alternatively, one could replace the meager values with very small positive values, as this allows us to multiplicatively adjust within the X12 framework. The replacement strategy amounts to an *ad hoc* method for extreme value adjustment, and work at StatsNZ has found that the resulting adjustments are even more unstable than those arising from an additive adjustment. Ultimately, it does not resolve the fact that meager values exert an exorbitant impact on the results of the seasonal adjustment. (In the language of Peña (1990), these are influential observations.)

Another solution explored at StatsNZ is to ignore the meager values and small extremes by collapsing the time axis, essentially redefining the sampling frequency from four quarters per year to a say, triadic period. This procedure might seem viable, if one believes that the meager values do not hold meaningful information for the purposes of understanding trend dynamics. Also, because the series in question are often not directly published, and moreover are providing a negligible contribution to total actual quarterly economic activity, it may seem sensible to ignore these types of features. However, in practice the number of meager values can vary by year so that the sampling frequency becomes variable; this is also true of the Apple exports series (middle panel of Figure 1), where the first quarter is initially less than 1% of the annual value but increases to over 10% of the annual for the last 3 years. Thus, in this approach any attempt to use standard time series models becomes impossible. Moreover, straight omission of meager values generates accounting difficulties for higher aggregates of component series, if one desires that seasonally adjusted totals be equal to the sum of the seasonally adjusted components, as many users of NSO seasonally adjusted outputs expect.

However, one might consider replacing extreme values with predictions based upon a fitted time series model – this is related to the collapsing strategy, but respects the calendrical structure, i.e., that there are four quarters per year. Our replacement procedure is formulated according to the maximum entropy principle, thereby rendering the time series more amenable to signal extraction procedures (such as seasonal adjustment) that are predicated upon Gaussian data. Essentially, the original time series is transformed such that its entropy increases; this is the same philosophy as using a Box-Cox transform. Alternatively, one might consider a non-Gaussian model with a non-Gaussian signal extraction procedure, but the modeling becomes more complex – either way, one is accounting for influential observations; see the review in McElroy (2016). Whereas it is common in the outlier literature (see Fox (1972) and Chang, Tiao, and Chen (1988)) to perform the replacement via the device of an additive outlier (AO) regressor (Ljung, 1993), we instead use stochastic projection formulas because the missing value is a random variable, not a fixed quantity.

Our approach replaces the extremes by a model-based imputation, or replacement, which takes into account the dynamics in each individual series. Specifically, extreme values are omitted and treated as missing, being first identified via significant increases to the sample's Gaussian entropy arising from their imputation. Although there is information content in these extremes, this is completely ignored by the imputation method, which is based on treating the omissions as *missing completely at random*; this is done to increase the Gaussian entropy, as is explained in Section 2.

The modified data can be log transformed, and by computing the likelihood based upon having missing values, the model can be fitted. Seasonal adjustment can then proceed by regular X12 applied to the transformed series, where imputations are supplied for missing values and extremes, as well as forecasts and backcasts, utilizing conditional expectation formulas arising from the fitted model. Uncertainty can also be quantified for these estimates, through the imputation mean squared error formulas. The key requirement is that there exist at least d contiguous values in the time series, where d is the order of the differencing polynomial for the nonstationary time series model.

We describe the maximum entropy principle for adjusting anomalous values in Section 2, making connections to prior literature based upon regression methods. The computational aspects of likelihood evaluation, extreme-value adjustment, and seasonal adjustment are discussed in Section 3, and Section 4 provides an evaluation of the methodology through simulations. In Section 5 we apply this method on the quarterly agricultural series of price exports for Avocado, Apple, and Berry commodities at StatsNZ, covering 1988 through the second quarter of 2005, in each case producing a multiplicative decomposition into seasonal and non-seasonal components. Section 6 summarizes our findings.

2 Maximum Entropy, Missing Values, and Extremes

Classical projection formulas for forecasting, imputation, and signal extraction are based upon linear estimators, and are only optimal for a Gaussian process; see McElroy (2008). The validity of these linear filtering methods is impugned when the data is non-Gaussian, and hence it is desirable to either develop non-linear filtering techniques suitable for non-Gaussian time series, or to first modify the data vector such that its cumulants more closely resemble those of a Gaussian random vector. The maximum entropy principle (Jaynes (1957a, b)) can be utilized to address the second approach.

In general, the maximum entropy principle states that the analyst should seek a probability distribution of maximal entropy subject to relevant empirical constraints. For example, the random vector of maximal entropy given a particular mean vector and covariance matrix is the multivariate Gaussian. Moreover, entropy maximization – subject to the constraint that the data vector is drawn from a difference stationary Gaussian model – is equivalent to maximum likelihood estimation; replacing extreme values by the Gaussian projection further maximizes the entropy, as we show below.

The maximum entropy principle is typically used by specifying certain constraints on a distribution, which may take the form of moments set equal to known quantities; having obtained the maximizing distribution, some empirical estimate of the constraint is then substituted (Golan et al., 1996). We apply this principle to derive the log Gaussian likelihood as criterion function. Let $X = [X_1, \ldots, X_T]'$ be a sample from $\{X_t\}$, and let $\mu = \mathbb{E}X$ and $\Sigma = \mathrm{Cov}(X)$. Conditional on these quantities, the maximum entropy distribution of X is $\mathcal{N}(\mu, \Sigma)$ (Park and Bera, 2009). Let $\mathcal{M}$ be a Gaussian model for $\{X_t\}$, which specifies some mean vector and covariance matrix. In this paper we consider the class of difference-stationary time series models discussed in McElroy and Monsell (2015), which presumes the existence of a degree d unit-root polynomial δ such that $W_t = \delta(B)X_t$ and $\{W_t\}$ is stationary. Bell (1984) discusses the representation of such processes in terms of d initial values of $\{X_t\}$, which can be any contiguous d values, say $X_* = [X_{t^*+1}, \ldots, X_{t^*+d}]'$ for some t^*. Supposing that $0 \leq t^* \leq T-d$, and that $\{W_t\}$ is uncorrelated with X_*, it follows from Theorem 1 of McElroy and Monsell (2015) that the quadratic form $X' \, \Sigma^{-1} \, X$ can be factorized into two terms involving only X_* and $\{W_t\}$.

Let W consist of W_t for $t \in \{d+1, \ldots, T\}$, and suppose the model $\mathcal{M}$ specifies a mean vector η_β for W, where β is a parameter. Because $\{W_t\}$ is stationary and Gaussian under the model, it has a spectral density of the form $\sigma^2 \, f_\theta$, where σ^2 is the innovation variance and θ is a parameter. The covariance matrix of W is then described via $\sigma^2 \, \Gamma_\theta$, where Γ_θ is a Toeplitz matrix of dimension $T-d$ (corresponding to f_θ). It then follows that

$$[X-\mu]' \, \Sigma^{-1} \, [X-\mu] = [X_* - \mu_*]' \, \Sigma_*^{-1} \, [X_* - \mu_*] + \sigma^{-2} \, [W - \eta_\beta]' \, \Gamma_\theta^{-1} \, [W - \eta_\beta], \qquad (1)$$

where $\mu_* = \mathbb{E}X_*$ and $\Sigma_* = \mathrm{Cov}[X_*]$. The factorization results of McElroy and Monsell (2015) also show that

$$\log\det\Sigma = \log\det\Sigma_* + (T-d)\,\log\sigma^2 + \log\det\Gamma_\theta. \qquad (2)$$

Combining the results of Park and Bera (2009) with (2), and dispensing with all factors that do not depend on the parameter vector $[\beta, \theta, \sigma^2]$, we obtain the Gaussian entropy

$$H(\theta, \sigma^2) = .5\,(T-d)\,\left(\log\sigma^2 + \log\det\Gamma_\theta\right). \qquad (3)$$

Notice this does not depend on β. A related quantity, which does depend on β, is -2 times the log Gaussian likelihood, which can be expressed – utilizing (1) and (2) – as

$$\mathcal{L}(\beta, \theta, \sigma^2) = \log \det \Gamma_\theta + (T-d) \, \log \sigma^2 + \sigma^{-2} \, [W - \eta_\beta]' \, \Gamma_\theta^{-1} \, [W - \eta_\beta]. \tag{4}$$

This quantity is called the Gaussian divergence. Replacing the quadratic form

$$Q = (T-d)^{-1} \, [W - \eta_\beta]' \, \Gamma_\theta^{-1} \, [W - \eta_\beta] \tag{5}$$

by its expectation $q(\theta) = (T-d)^{-1} \, \mathrm{tr}\{\Gamma_W \, \Gamma_\theta^{-1}\}$ yields an asymptotic form of the Gaussian divergence:

$$\mathcal{W}(\theta, \sigma^2) = \log \det \Gamma_\theta + (T-d) \, \log \sigma^2 + (T-d) \, q(\theta) \, \sigma^{-2}.$$

This function is the Kullback–Leibler divergence between two Gaussian random vectors with the same mean, and respective covariance matrices Γ_W and $\sigma^2 \, \Gamma_\theta$. We suppose the model $\mathcal{M}$ is correctly specified, so that there is some true unknown $[\widetilde{\beta}, \widetilde{\theta}, \widetilde{\sigma}^2]$, which we are seeking to determine. Then $\Gamma_W = \widetilde{\sigma}^2 \, \Gamma_{\widetilde{\theta}}$, and we can write

$$q(\theta) = \widetilde{\sigma}^2 \, (T-d)^{-1} \, \mathrm{tr}\{\Gamma_{\widetilde{\theta}} \, \Gamma_\theta^{-1}\},$$

which equals $\widetilde{\sigma}^2$ if $\theta = \widetilde{\theta}$. The following result connects entropy (3) to the Gaussian divergence, allowing us to claim that minimizing the latter accomplishes entropy maximization.

Theorem 1 *Given the Gaussian model $\mathcal{M}$ for $\{X_t\}$, the maximum entropy distribution has parameter vector $[\widetilde{\theta}, \widetilde{\sigma}^2]$ obtained as the minimizer of the Kullback–Leibler divergence $\mathcal{W}(\theta, \sigma^2)$, and the parameter estimate $[\widehat{\beta}, \widehat{\theta}, \widehat{\sigma}^2]$ is obtained as the minimizer of the Gaussian divergence $\mathcal{L}$.*

Now consider the case that some observations are extremal, and suppose $\mu = 0$ to simplify the discussion. Let $X_\circ$ denote moderate observations, and X_∂ be the block of extremes – the notation designates the interior ($\circ$) and boundary (∂) of the data vector's set of values. There exists some permutation matrix Π such that

$$X = \Pi \left[\begin{array}{c} X_\circ \\ X_\partial \end{array} \right]. \tag{6}$$

Employing the notation $\Sigma_\circ = \mathrm{Cov}[X_\circ]$ and $\Sigma_\partial = \mathrm{Cov}[X_\partial]$, elementary results for random vectors (Mardia, Kent, and Bibby (1979)) yield

$$X' \, \Sigma^{-1} \, X = X_\circ' \, \Sigma_\circ^{-1} \, X_\circ + [X_\partial - \widehat{X_\partial}]' \, \Sigma_{\partial|\circ}^{-1} \, [X_\partial - \widehat{X_\partial}], \tag{7}$$

where $\widehat{X_\partial} = \mathbb{E}[X_\partial | X_\circ]$ and $\Sigma_{\partial|\circ} = \mathrm{Cov}[X_\partial - \widehat{X_\partial}]$. One consequence of this expression is that $X_\circ' \, \Sigma_\circ^{-1} \, X_\circ$ can be computed by calculating $\widetilde{X}' \, \Sigma^{-1} \, \widetilde{X}$, where

$$\widetilde{X} = \Pi \left[\begin{array}{c} X_\circ \\ \widehat{X_\partial} \end{array} \right].$$

That is, the likelihood corresponding to the regular values alone is identical to that arising from replacing the extremes X_∂ by their estimate $\widehat{X_\partial}$. If the extremes were omitted, the missing value likelihood could be evaluated the same way. This observation can be advantageous when certain algorithms are available for computing $X' \, \Sigma^{-1} \, X$; we only need to supply imputations $\widehat{X_\partial}$ for the missing values, and then can proceed. This is the reasoning behind the skipping method described in Gómez, Maravall, and Peña (1999), as is further discussed in Section 3.

Ljung (1993) advocated handling missing values by inserting arbitrarily large values followed by extreme-value adjustment based on additive outlier regression techniques. We reverse this philosophy: treat extremes X_∂ as missing values (i.e., excise them) and replace them via $\widehat{X_\partial}$. If a set of r values X_∂ are extreme, and are replaced by $\widehat{X_\partial}$, then the quadratic form in (7) is reduced by the non-negative quantity

$$S_\partial = [X_\partial - \widehat{X_\partial}]' \, \Sigma^{-1}_{\partial|\circ} \, [X_\partial - \widehat{X_\partial}], \tag{8}$$

and the Gaussian divergence gets computed as if X_∂ were missing. (If a series has actual missing values as well, these can be collected in the X_∂ vector as well.) The quantity S_∂ is the difference between the Gaussian divergence with and without extreme-adjustment. In Section 3, we discuss how the Gaussian divergence can be computed for difference stationary series with arbitrary patterns of missing values. Because $S_\partial \geq 0$, *extreme-value adjustment increases entropy.* However, we should only be concerned with statistically significant increases to entropy – otherwise, we might replace all the entries by the sample mean. To that end, consider the null hypothesis

$$H_0 : \mathcal{M} \text{ is correct.}$$

If S_∂ is computed using the true parameter vector $[\widetilde{\theta}, \widetilde{\sigma}^2]$, then under H_0 it has a χ^2_r distribution (Mardia, Kent, and Bibby (1979)), where r is the length of X_∂. However, in practice we must utilize parameter estimates, computed by maximizing the Gaussian divergence based on $X_\circ$; under H_0, these are maximum likelihood estimators and hence converge in probability as $T \to \infty$ to the true parameters. Therefore, applying Slutsky's theorem, the statistic $\widehat{S}_\partial$ obtained by plugging in the parameter estimates satisfies

$$\widehat{S}_\partial \xrightarrow{P} S_\partial \sim \chi^2_r.$$

Comparing $\widehat{S}_\partial$ against the χ^2_r quantiles is asymptotically correctly sized, and we refer to this procedure as our *extreme-value test.* The power of this test is generated by scenarios where $X_\partial - \widehat{X_\partial}$ is large in magnitude.

3 Algorithms and Computations

We describe how the Gaussian divergence can be evaluated when missing values are present, how to test for extremes and adjust for them, and finally how seasonal adjustment can be completed. Much of this material is based on McElroy and Monsell (2015).

3.1 Gaussian Divergence

Let X denote the entire potential sample of size T. We know that potential extremes can be replaced by their conditional expectations, in which case the calculation of the Gaussian divergence proceeds as if these were also omitted. Therefore, we let all r omitted and deleted values be denoted X_∂, and the remaining m values of X are denoted $X_\circ$.

We suppose that transformations have already been applied. Any meager values in the original scale can be viewed as potential extremes, and hence omitted; the remaining values are positive, and hence the logarithm is well-defined. Utilizing the notation of Section 2, we suppose the contiguous initial values X_* are observed, which amounts to assuming the existence of some contiguous set of d observations – otherwise, the Gaussian likelihood need not factorize (McElroy and Monsell, 2015). Let P denote a permutation that interchanges the indices $\{t^*+1, \ldots, t^*+d\}$ with $\{1, \cdots, d\}$, leaving all other indices unaltered. Furthermore, define the $(T-d) \times T$ dimensional differencing matrix Δ via

$$\left[\begin{array}{cccccc} \delta_d & \cdots & \delta_1 & \delta_0 & 0 & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \cdots & 0 & \delta_d & \cdots & \delta_1 & \delta_0 \end{array} \right],$$

where $\delta(B) = \sum_{j=0}^d \delta_j B^j$. Then the matrix

$$\Lambda = \left[\begin{array}{c} [1_d \; 0] \, P \\ \Delta \end{array} \right],$$

where 1_d denotes a d-dimensional identity matrix, maps X to the initial values X_* concatenated to the differenced values W. It is known that Λ is invertible, and hence by (6)

$$X_\circ = [1_m \; 0] \, \Pi^{-1} \, X = [1_m \; 0] \, \Pi^{-1} \, \Lambda^{-1} \left[\begin{array}{c} X_* \\ W \end{array} \right].$$

However, because X_* is observed, it is featured among the entries of $X_\circ$, and hence there exists a permutation Q such that

$$Q \, [1_m \; 0] \, \Pi^{-1} \, \Lambda^{-1} = \left[\begin{array}{cc} 1_d & 0 \\ \underline{A} & \underline{B} \end{array} \right]$$

for some matrices $\underline{A}$ and $\underline{B}$. Letting

$$R = \left[\begin{array}{cc} 1_d & 0 \\ -\underline{A} & 1_{m-d} \end{array} \right] Q,$$

we obtain (using † to denote inverse transpose)

$$R \, X_\circ = \left[\begin{array}{c} X_* \\ \underline{B} \, W \end{array} \right]$$

$$\Sigma_\circ = R^{-1} \left[\begin{array}{cc} \Sigma_* & 0 \\ 0 & \sigma^2 \, \underline{B} \, \Gamma_\theta \, \underline{B}' \end{array} \right] R^\dagger.$$

Evidently, R is an invertible transformation of the observed variables that reduces them to initial values and a linear combination of differenced variables, and hence facilitates factorization of the likelihood. Taking the lower rows of R enables us to discard the initial values:

$$\begin{aligned} D &= [0 \; 1_{m-d}] \, R = [-\underline{A} \; 1_{m-d}] \, Q \\ D \, X_\circ &= \underline{B} \, W. \end{aligned}$$

Finally, we can generalize the result (1) to the case of missing values:

$$[X - \mu]' \, \Sigma^{-1} \, [X - \mu] = [X_* - \mu_*]' \, \Sigma_*^{-1} \, [X_* - \mu_*] + \sigma^{-2} \, [W - \eta_\beta]' \, \underline{B}' \left[\underline{B} \, \Gamma_\theta \, \underline{B}' \right]^{-1} \underline{B} \, [W - \eta_\beta], \quad (9)$$

where we can compute $\underline{B} \, W$ via $D \, X_\circ$. The mean vector η_β is typically written as a linear combination β of known regressors, and hence $\underline{B} \, \eta_\beta$ is easily calculated. We also have an analogue of (2). Substituting into (4), and omitting the terms involving the initial values, yields

$$\mathcal{L}(\beta, \theta, \sigma^2) = \log \det \left(\underline{B} \, \Gamma_\theta \, \underline{B}' \right) + (T - d) \, \log \sigma^2 + \sigma^{-2} \, [W - \eta_\beta]' \, \underline{B}' \left[\underline{B} \, \Gamma_\theta \, \underline{B}' \right]^{-1} \underline{B} \, [W - \eta_\beta]. \quad (10)$$

This Gaussian divergence can be minimized using standard algorithms, and model fit can be checked by computation and examination of the residuals. We have encoded this procedure (for specified SARIMA models) in R.

3.2 Extreme Value Adjustment

For extreme-value adjustment, one could construct the Gaussian divergence (10) for various choices of extremes X_∂, and compare the entropy via $\widehat{S}_\partial$. Note that it is not necessary to compute $\widehat{S}_\partial$ via the formula (8); we only have to take the difference of the two Gaussian divergences. Clearly, there are 2^T possible subsets of extremes to potentially consider, so a practical algorithm is needed to quickly identify and process the most salient anomalies.

Once the extremes are identified through statistical tests, they are replaced by imputations arising as best estimates of values *missing completely at random*, i.e., in this context, Gaussian conditional expectations. Although the bare fact that these extremes are blotted out tells us something – in particular, that they decrease the sample's entropy – this information is not utilized in the missing value imputation procedure, because that would not accomplish the purpose of attenuating the extremes. Moreover, taking into account the fact of identification in the imputation procedure would require a more nuanced modeling of the data process, which is at odds with the objective of our procedure.

We pursue a forward addition strategy, gradually adding extremes as needed so as to significantly increase entropy. First, fit the model with no extremes to obtain the parameter estimates. Then compute $\widehat{S}_\partial$ based on excising each single data point in turn, based on the same parameter estimates. (Alternatively, the model can be fitted to each such excision; while this is preferable from the standpoint of accuracy, it is very expensive computationally, with T maximizations to be computed.) Then rank the observations by the p-values (ignoring, for sake of efficacy, the effects of sorting and multiple-testing) and consider batches of the most extreme observations.

Next, consider a sequence of χ^2_1 tests, which are formulated as comparing the Gaussian divergence with the $(k+1)$ most extreme observations excised against the k most extreme observations excised. Beginning with $k=0$, this compares deleting the most extreme observation against retaining it. If the test rejects, we excise this observation, and now take this as a our new base model; then compare deleting the two most extreme observations against this base model. When for some k we fail to reject, we have k excisions defining our extreme-value adjusted time series, and we refit the model to get final parameter estimates. At this point, residuals can be assessed for goodness-of-fit using standard tools (but with significance ignoring the prior extreme-value adjustment).

This method can be compared with other available procedures. Our algorithm for sifting batches of extremes is based upon the procedure of X12, which takes the Ljung (1993) AO approach. After the initial pass through the data has obtained a regression coefficient corresponding to an AO at each data point in turn, the t-statistics are sorted and compared to *ad hoc* critical values. Ljung (1993) advocates using an AO regressor to account for, and ultimately replace, an extreme value. If X_∂ is treated as a deterministic quantity, then $\widehat{X}_\partial$ is estimated by Generalized Least Squares (GLS); formula (7) still holds true, and S_∂ is also χ^2_r. However, the estimate $\widehat{X_\partial}$ can be quite different in a GLS formulation, as opposed to our general approach that allows for both fixed effects (through the mean vector) and stochastic effects.

Gómez, Maravall, and Peña (1999) make the point that the parameter estimates for GLS in Ljung (1993) rely upon the extremes. If the model fitting criterion is based upon $X_\circ' \, \Sigma_\circ^{-1} \, X_\circ$, rather than $X' \, \Sigma^{-1} \, X$, then the potential extremes have been eliminated from the calculation, and hence won't impact the parameter estimates. This point is also discussed in Brubacher and Wilson (1976). Gómez, Maravall, and Peña (1999) discuss the efficient computation of $X_\circ' \, \Sigma_\circ^{-1} \, X_\circ$, which is achieved by running a suitable algorithm – such as Durbin-Levinson or the Kalman filter – upon $X' \, \Sigma^{-1} \, X$ but with X_∂ replaced by $\widehat{X_\partial}$. In the recursive Durbin-Levinson algorithm, the update then becomes numerically zero, indicating that the particular step in the loop of the algorithm can be skipped – hence the term *skipping method*.

This skipping method yields an efficient algorithm for computing the Gaussian divergence, at least in the case of stationary data. Whereas the Durbin-Levinson is intended for stationary data, Gómez, Maravall, and Peña (1999) show how the Kalman filter can be utilized with a skipping method to compute the Gaussian divergence for nonstationary processes. However, because of their diffuse initialization of the Kalman filter, they take the initial values to be at the beginning of the series – if any of the first d values are missing, these will be treated as deterministic unknowns, as the authors discuss on p.348. Latter missing values, for $t > d$, are treated as stochastic.

Missing values can also be treated in a state space approach through the alternate initialization of Ansley and Kohn (1985), as discussed further in Bell and Hillmer (1991). This treatment does not require the initial values to be at the start of the series, and all missing values can be treated properly as a mixture of deterministic and stochastic effects. In essence, missing values that occur on the boundary of the sample are estimated by forecasting and backcasting, and are never treated as deterministic – so why should missing values in the interior of the sample be treated as deterministic? This is the philosophy behind the treatment of McElroy and Monsell (2015), which is adopted here. Our approach is equivalent to that of Bell and Hillmer (1991), but does not require state space algorithms, being applicable to long memory processes.

Once the extremes have been identified, we need their imputations, which can be expressed in terms of the notation of the previous subsection. In general the Gaussian conditional expectation is

$$\widehat{X_\partial} = \mathbb{E} X_\partial + \mathrm{Cov}[X_\partial, X_\circ] \, \mathrm{Cov}[X_\circ]^{-1} \, [X_\circ - \mathbb{E} X_\circ].$$

These missing values could be in the interior of the sample, or outside the boundary, corresponding to forecasts and backcasts.

Proposition 1 *Given the Gaussian model $\mathcal{M}$ with true parameters $[\beta, \theta, \sigma^2]$, the missing/omitted vector of r observations X_∂ is estimated via*

$$\widehat{X_\partial} = [0 \, 1_r] \, \Pi^{-1} \, \Lambda^{-1} \left[\begin{array}{l} X_* \\ \eta_\beta + \Gamma_\theta \, C(\theta) \, (W - \eta_\beta) \end{array} \right],$$

with $\underline{B} \, W = D \, X_\circ$ and

$$C(\theta) = \underline{B}' \left[\underline{B} \, \Gamma_\theta \, \underline{B}' \right]^{-1} \underline{B}.$$

The error covariance matrix is

$$\Sigma_{\partial|\circ} = [0 \, 1_r] \, \Pi^{-1} \, \Lambda^{-1} \left[\begin{array}{cc} 0 & 0 \\ 0 & \sigma^2 \, (\Gamma_\theta - \Gamma_\theta \, C(\theta) \, \Gamma_\theta) \end{array} \right] \Lambda^\dagger \, \Pi^\dagger \, [0 \, 1_r]'.$$

These formulas are used to replace meager values, adjust extremes, and provide forecasts and backcasts as needed for filtering. As discussed in McElroy and Monsell (2015), filtered estimates can be obtained by applying a linear filter $\Psi(B) = \sum_{j \in \mathbb{Z}} \psi_j B^j$ to such an adjusted/extended series; in the case of an infinite length filter, such as arises from model-based signal extraction, we determine the maximal coefficient index such that all coefficients are below some threshold, such as machine precision. For instance, if applying the filter with coefficients $\psi_j = .9^{|j|}$, then $|j| > 175$ implies $\psi_j < 10^{-8}$. Then T is determined accordingly, generating as many forecasts as needed. The signal extraction uncertainty can also be computed utilizing $\Sigma_{\partial|\circ}$.

3.3 Seasonal Adjustment

We write the untransformed data in lower case letters, reserving capital letters for the transformed data process $\{X_t\}$ with seasonal $\{S_t\}$ and non-seasonal $\{A_t\}$. We seek a decomposition of $\{x_t\}$ into a product of seasonal $\{s_t\}$ and non-seasonal $\{a_t\}$ components:

$$x_t = s_t \cdot a_t. \tag{11}$$

Before applying a logarithm, we must handle the meager values. A naïve approach involves simply adding a small constant $c > 0$ to x_t, such that $x_t + c > 0$ for all t, before log transforming. This method has serious drawbacks. If we obtain S_t and A_t from an additive decomposition of $\log(x_t + c)$, then reversing this relationship entails

$$x_t = \exp\{S_t\} \cdot \exp\{A_t\} - c,$$

which will not be equal to $\exp\{S_t\} \, (\exp\{A_t\} - c)$, and no multiplicative decomposition of the form (11) is possible. For instance, setting $s_t = \exp\{S_t\}$ and $a_t = x_t / s_t$ yields

$$a_t = \exp\{A_t\} - c \cdot \exp\{-S_t\},$$

which indicates that if the seasonal adjustment is adequate in the log domain, and c is non-negligible, then there will be some remaining seasonal oscillations in a_t. This is unsatisfactory. (This discussion also leaves aside the issue of the selection of c, which is not unique and will have an impact on the signal extraction estimates in the log domain.)

Our approach avoids these obstacles. The observed data is modified by excising meager values and extremes, taking a log transform of the other values. By minimizing the Gaussian divergence and searching over different batches of extremes, as described in the previous subsection, we ultimately produce an extreme-value adjusted $\{\widetilde{X}_t\}$. If working with the X12 methodology, seasonal adjustments are produced by the application of a fixed X-11 filter $\Psi(B)$, as described in Ladiray and Quenneville (2001). Extending the sample by forecasts and backcasts, we compute S_t via

$$S_t = \Psi(B) \, \widetilde{X}_t.$$

Subtracting, we obtain $A_t = \widetilde{X}_t - \widehat{S}_t$. Exponentiating, we obtain

$$\widetilde{x}_t = \exp\{S_t\} \cdot \exp\{A_t\} \tag{12}$$

for each time t. Note that $\widetilde{x}_t$ equals x_t for most times t, but differs if t corresponds to a meager value, identified extreme, or forecast/backcast. Although it seems natural to set $s_t = \exp\{S_t\}$ and $a_t = \exp\{A_t\}$, this has the drawback that (11) would be violated for any t such that $\widetilde{x}_t \neq x_t$. However, if we assign the discrepancy to the seasonal or non-seasonal, we can ensure that (11) holds for all t. For the agricultural series studied in this paper, meager values are associated with certain seasons of low productivity (due to weather patterns in New Zealand), and hence are naturally associated with the seasonal. Therefore, instead of $s_t = \exp S_t$, we should set $a_t = \exp A_t$ and make the definition

$$s_t = x_t / a_t. \tag{13}$$

Note this division is well-defined, because the non-seasonal is non-zero. Now for non-excised values, we have $s_t = \widetilde{x}_t / a_t = \exp S_t$, but in general $s_t = \exp S_t \cdot (x_t / \widehat{x}_t)$. So if the value is meager, and equals zero, then $s_t = 0$ as well; or if the meager value is negative, then $x_t / \widehat{x}_t$ must be negative, and $s_t < 0$ as well. So s_t will have the same meager behavior as the original x_t, but looks like a regular seasonal at non-meager time points.

4 Simulation Study

In order to evaluate the proposed methodology, we conducted a short simulation study. The first task was to generate time series data resembling observed meager series, with a facility for generating a greater incidence of anomalies through tuning parameters. The second task is to develop criteria by which the extreme value seasonal adjustment methodology would be deemed successful.

To generate simulations, we adopt one of the SARIMA models later utilized in our empirical studies, and pass heavy-tailed innovations through the ARMA polynomials. Our model process $\{X_t\}$ is the quarterly SARIMA(3,1,0)(0,1,1), with $\{W_t\}$ given by

$$(1 - .037B - .046B^2 - .046B^3)W_t = (1 - .055B^4)\epsilon_t,$$

with innovation scale $\exp\{0.247\}$. The innovation scale is equal to the standard deviation when the innovations are Gaussian, but we also generated student t innovations with 10, 5, and 2 degrees of freedom. Then we define a left-censored process $x_t = 1_{\{X_t > \tau\}} \exp\{X_t\}$, where τ is a threshold chosen to be -2.

Each simulation was re-centered and re-scaled before exponentiation, to ensure the data was not dominated by one or two values. The left-censoring ensures that any suitably small values (less than e^{-2}) are set to zero. Such synthetic data retains the features of NZ agricultural data, while also having a non-negligible incidence of zero values. Note that the degree of freedom is related to the innovations' marginal tail index, which governs the magnitude of extremes rather than their ubiquity.

Generating such series of length $T = 40, 60, 80$, we applied our methodology with $\alpha = .01, .05, .10$. We expect the method to correctly identify anomalies, although these are actually defined in terms of the Gaussian likelihood in this paper, so that each simulation can have a different number of extremes. Nevertheless, we expect more extremes with a lower degree of freedom in the t innovations, and by increasing α we should be able to increase our detection. Table 1 reports the proportion of extremes detected (given by the number of imputations in each simulation, divided by sample size and the total number $1,000$ of replications) in each case, which is roughly increasing in α and stable in T. The proportion does not increase with lower degrees of freedom, because this parameter only impacts the magnitude of extremes.

Process	Proportion Extremes			SA Inadequacy		
dof 2	.01	.05	.10	.01	.05	.10
$T = 40$	.0343	.0712	.1325	0	0	0
$T = 60$	.0261	.0693	.1281	0	2	0
$T = 80$	.0231	.0683	.1261	4	0	5
dof 5	.01	.05	.10	.01	.05	.10
$T = 40$	.0273	.0659	.1340	0	0	0
$T = 60$	.0201	.0679	.1345	0	0	0
$T = 80$	.0167	.0690	.1345	0	0	0
dof 10	.01	.05	.10	.01	.05	.10
$T = 40$	.0258	.0635	.1353	0	0	0
$T = 60$	.0183	.0651	.1798	0	0	0
$T = 80$	.0143	.0674	.1345	0	0	0
dof ∞	.01	.05	.10	.01	.05	.10
$T = 40$	.0317	.0590	.1277	0	0	0
$T = 60$	.0176	.0877	.1824	0	0	0
$T = 80$	.0134	.0665	.1366	0	0	0

Table 1: Detection of extremes and seasonal adjustment quality measures for four simulated processes. Three sample sizes ($T = 40, 60, 80$) were considered, four degrees of freedom (dof) for data processes, and extreme-value adjustment thresholds determined by $\alpha = .01, .05, .10$. The proportion of extremes counts the number of identified extremes divided by sample size, averaged over all $1,000$ simulations. SA Inadequacy gives the number of series (out of $1,000$) for which the lag four sample autocorrelation was significantly different from zero.

While the anomaly count is mainly a check on the overall soundness of the method, actual performance should be assessed via adequacy of seasonal adjustment. Many measures of adequacy appear in the literature (see McElroy (2012) for an overview), but we focus on the lag four sample autocorrelation of the differenced adjusted time series, and determine the statistical significance by using the Bartlett formula assuming a fitted MA(3) process is truth. This means our significance levels are higher than what a white noise hypothesis would generate. Table 1 indicates a very low incidence of inadequacy, according to this metric, the only problems occurring when the simulations are extremely heavy-tailed (two degrees of freedom corresponds to infinite variance). Note that setting degrees of freedom equal to ∞ corresponds to Gaussian innovations.

Given that the synthetic data appears to replicate the observed series' main features, one could propose modeling the NZ data via the left-censoring mechanism described above. However, even if the tail index and left-censoring threshold τ could be accurately estimated, seasonal adjustment would then proceed via nonlinear algorithms, which typically requires greater computational resources as well as additional human judgment. Supposing that one imputed a value for $\exp\{X_t\}$ when $X_t \leq \tau$, the resulting adjusted series would still have extremes – only the censoring of meager values has been accounted for. Non-Gaussian signal extraction methods would be required for further analysis. In a sense, our simple Gaussian model is a conscious mis-specification – which avoids complex modeling and attendant nonlinear signal extraction – that we hope is an adequate approximation to the true data process.

5 New Zealand Agricultural Data

We now apply the paper's method to three agricultural time series from New Zealand, each of which seems to demand multiplicative seasonal adjustment, but also has some extremely low values. These series are constant price exports for commodities: Avocado, Apple, and Berry, for the period of 1988 through the second quarter of 2005.

5.1 Avocado

We first consider the New Zealand Avocado series. A quarterly SARIMA(3,1,0)(0,1,1) model was identified and fitted (to the logged, trimmed data). The differenced series $\{W_t\}$ followed the model

$$(1 - .017B - .046B^2 - .009B^3)W_t = (1 - .141B^4)\epsilon_t,$$

with innovation variance $\exp\{0.219\}$. The lag 24 Ljung Box statistic has p-value .377, and the sign test had p-value .093. The model could be simplified, as some of the coefficients are quite small. As there were two zero values, these were automatically omitted, and the extreme-value procedure with $\alpha = .05$ was applied to the logged data. In addition, anomalies were detected: a large extreme at 1989.Q2, and two meager values on 1996.Q2 and 1997.Q3. Having imputed replacements for these anomalies, the resulting extreme-value adjusted series (left panel of Figure 2) was seasonally adjusted via the X12 method, using a 3 x 5 seasonal and a Henderson 9 trend. In exponential scale, the final results are given in the right panel of Figure 2.

The resulting non-seasonal passes through the center of the meager series (the missing value estimates for the log data are not displayed). Uncertainty is naturally greater at these time points. The seasonal component is computed from the imputed series in log domain; in the original scale, the seasonal is obtained by exponentiating and adjusting for meager values, as given by (13). The seasonal adjustment was adequate according to standard diagnostics (McElroy, 2012).

We also examined the seasonal adjustment revisions for the Avocado series. We begin by withholding five years of data from the end of the sample, and continue to add one quarter at a time. However, in order to enforce consistency with past analyses, we do not allow previously identified

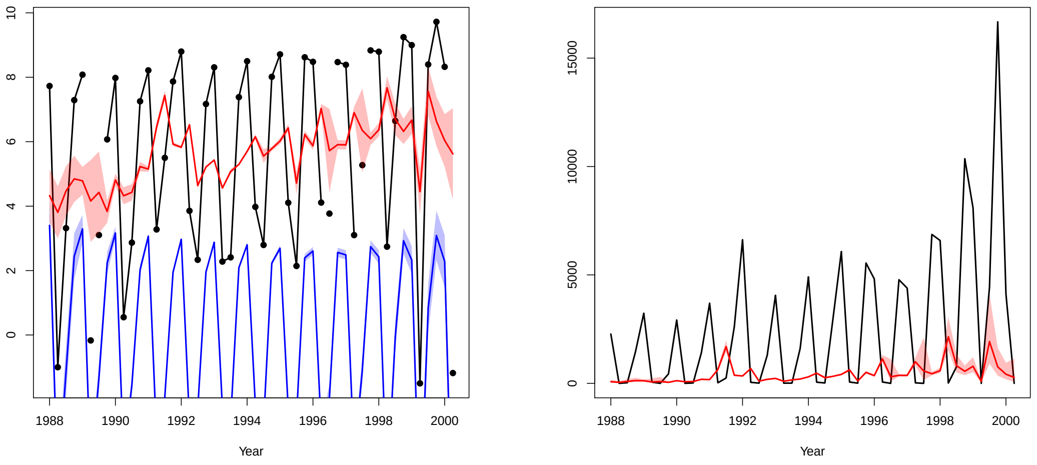

Figure 2: Left panel: Log scale of Avocado data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Right panel: data in original scale, with exponential of the X12 seasonal adjustment (red). Confidence intervals are shaded.

extremes to become non-extremes as additional data is added (although previously identified non-extremes can be rendered as an extreme as data is added, and moreover the imputations themselves can change). The model parameters are re-estimated, and hence the resulting seasonal adjustments can change as data is added; Figure 3 summarizes the behavior.

Apparently, the revisions are small for this example. If we reduce α to .01, we are being more demanding in our criterion for extreme behavior, and hence less extremes will be identified; as a result, the seasonal adjustments will tend to revise more, as they are impacted more greatly by influential values. Conversely, if we increase α to .10, then more extremes will be identified, with their values being imputed, and the seasonal adjustment will be more stable. In Figures 4 and 5, we display the revision results for $\alpha = .01$ and $\alpha = .10$.

5.2 Apple

We next consider the New Zealand Apple series. A quarterly SARIMA(1,1,1)(1,1,1) model was identified and fitted (to the logged, trimmed data). The differenced series $\{W_t\}$ followed the model

$$(1 + .441B)(1 - .373B^4)W_t = (1 + .409B)(1 - .572B^4)\epsilon_t,$$

with innovation variance $\exp\{-.534\}$. The lag 24 Ljung Box statistic has p-value .962, and the sign test had p-value .5. There were four negative values, which were automatically omitted, and then the sequential extreme-value testing procedure with $\alpha = .05$ was applied to the logged data. In addition, meager values were identified at 1990.Q4, 1995.Q4, and 1996.Q1. Having imputed replacements for these, the resulting extreme-value adjusted series (left panel of Figure 6) was seasonally adjusted via the X12 method, using a 3 x 5 seasonal and a Henderson 9 trend. In exponential scale, the final results are given in the right panel of Figure 6. The adjustment was deemed adequate.

5.3 Berry

Finally, we consider the New Zealand Berry series. A quarterly SARIMA(3,1,0)(0,1,1) model was identified and fitted (to the logged, trimmed data). The differenced series $\{W_t\}$ followed the model

$$(1 + .056B + .018B^2 + .216B^3)W_t = (1 + .309B^4)\epsilon_t,$$

with innovation variance $\exp\{-1.640\}$. The lag 24 Ljung Box statistic has p-value .931, and the sign test had p-value .093. This series was positive, so no initial omissions were inserted, but the sequential extreme-value testing procedure with $\alpha = .05$ detected five anomalies: three large extremes at 1989.Q3, 1992.Q3, and 2003.Q3, and two meager values at 1990.Q3 and 2000.Q3. Having imputed replacements for these, the resulting extreme-value adjusted series (left panel of Figure 7) was seasonally adjusted via the X12 method, using a 3 x 5 seasonal and a Henderson 9 trend. In exponential scale, the final results are given in the right panel of Figure 7.

This adjustment (Figure 7) did not have any obvious outliers, and the uncertainty is quite low compared to the other examples. The third quarter value for 2003 was a bit higher than was usual for the seasonal pattern, resulting in an apparent additive outlier in the non-seasonal. These types of effects, although located in the seasons more naturally associated with seasonality (third quarter), are anomalous to the regular calendrical pattern, and hence are typically associated with the irregular component, and not the seasonal. (E.g., having an unusually cold winter constitutes an effect located in the irregular, *not* the seasonal.)

6 Conclusion

In this paper we consider the problem of multiplicative seasonal adjustment for time series wherein negligible values may be present, possibly in addition to large extremes. This problem has been motivated by meager series produced by StatsNZ. We demonstrate that straightforward log transformations do not offer a feasible solution, and instead advocate a maximum entropy approach whereby extremes and meager values are replaced by their Gaussian conditional expectations. The algorithms, including computation of the Gaussian likelihood and calculation of imputed values, are fully discussed. Both the fitted model and the imputed data are determined concurrently, and X12 seasonal adjustments with uncertainty can be computed. This results in a multiplicative seasonal adjustment where we have assigned the meager values to the seasonal component.

This approach is similar in spirit to the method of X12, which tests sequentially the placement of an AO at each time point in the sample, although the new method treats missing values and extremes as random variables rather than as deterministic quantities. For any extreme-value adjustment procedure a decision must be rendered about which values are extremes, and how they are to be adjusted. We examine this through the lens of entropy maximization, and utilize quadratic statistics to sequentially test for batches of extremes.

To see the impact of the extreme-value adjustment upon the resulting seasonal adjustments, one can adjust the critical values of the extreme test statistic S_∂. A larger α increases the number of identified extremes, rendering the imputed series more regular, and hence providing a smoother seasonal adjustment. This behavior has been observed in simulation and on agricultural yield data. It also has an impact on revisions to seasonal adjustments, with lower values of α involving less extreme-value adjustment of the series, with a greater possibility for revision when new data is obtained.

Ultimately, we hope that identifying and correcting extremes will improve seasonal adjustment. The philosophy of filtering is motivated by the attempt to elicit latent dynamics in a time series, which are defined through recurring behavior (e.g., seasonal and cyclical movements) or long-term patterns (e.g., trends). Extreme values and anomalies fall outside the scope of these latent dynamics, and therefore should be omitted or attenuated before utilizing linear filters. Failure to do so will creates the danger that a single extreme value has an undue and spurious impact upon extracted latent dynamics.

An alternative approach involves taking the synthetic process described in Section 4 as the basis for modeling extremes (and the left-censoring process generating meager values). One could also adopt a mechanistic model of the accounting process that yields meager values. While being of potential academic interest, such efforts would require non-Gaussian models, ultimately resulting in non-linear signal extraction filters. Such non-linear filters would achieve much the same effect as excision and imputation of the extremes (c.f., Trimbur, 2010), though with the attractive feature of shrinking extremes (soft thresholding). For example, filtering based upon an absolute loss (instead of the usual squared loss, which gives rise to linear filters) results in signal extraction estimates that are more robust to outliers (Yamada and Jin, 2013).

While we believe such efforts have merit, for production purposes it is preferable to use a methodology that is agnostic about the mechanism that generates extreme values and meager values. It is essential to have a method that can be applied using fast linear algorithms. With tens of thousands of series to process, the reality at an NSO is that human analysts will not have time to examine each dataset. With the automatic modeling capabilities of X-12-ARIMA and TRAMO-SEATS, the modeling, extreme-value adjustment, and seasonal adjustment of time series data must be, is now, and will continue to be performed via automaton.

Appendix

Proof of Theorem 1. We first show that maximization of $H(\theta, \sigma^2)$ subject to the constraint $\sigma^2 = q(\theta)$ is identical to minimization of $W(\theta, \sigma^2)$. Defining the Lagrangian to be

$$J(\theta, \sigma^2, \lambda) = 2\, H(\theta, \sigma^2) - \lambda\, (q(\theta) - \sigma^2),$$

we find that the gradient is

$$\begin{aligned} \nabla_\theta J(\theta, \sigma^2, \lambda) &= \nabla_\theta \log \det \Gamma_\theta - \lambda\, \nabla_\theta q(\theta) \\ \nabla_{\sigma^2} J(\theta, \sigma^2, \lambda) &= (T - d)\, \sigma^{-2} + \lambda \\ \nabla_\lambda J(\theta, \sigma^2, \lambda) &= \sigma^2 - q(\theta). \end{aligned}$$

Setting this to zero, we find that $\lambda = -T\, \sigma^{-2}$, $\sigma^2 = q(\theta)$, and θ therefore optimizes $\mathcal{D}(\theta) = \log \det \Gamma_\theta + (T - d)\, \log q(\theta)$. The Hessian of J is given by

$$\nabla \nabla' J(\theta, \sigma^2, \lambda) = \left[\begin{array}{ccc} \nabla_\theta \nabla'_\theta \log \det \Gamma_\theta - \lambda\, \nabla_\theta \nabla'_\theta q(\theta) & 0 & -\nabla_\theta q(\theta) \\ 0 & -(T-d)\, \sigma^{-4} & 1 \\ -\nabla'_\theta q(\theta) & 1 & 0 \end{array} \right].$$

At a critical point, the upper left block is

$$\nabla_\theta \nabla'_\theta \log \det \Gamma_\theta + (T - d)\, \nabla_\theta \nabla'_\theta q(\theta)\, \sigma^{-2},$$

denoted A. By results on nested matrices, it follows that the eigenvalues of the Lagrangian's Hessian are given by the eigenvalues of A, as well as the negative value $-(T-d)\, \sigma^{-4}$, and a final eigenvalue equal to the Schur complement:

$$-\left[-\nabla'_\theta q(\theta), 1 \right] \left[\begin{array}{cc} A^{-1} & 0 \\ 0 & -\sigma^4/(T-d) \end{array} \right] \left[\begin{array}{c} -\nabla_\theta q(\theta) \\ 1 \end{array} \right] = \sigma^4/(T-d) - \nabla'_\theta q(\theta)\, A^{-1}\, \nabla_\theta q(\theta).$$

Below, we show that this is positive, and conclude that the Lagrangian is maximized with respect to σ^2. It is known that the Hessian of Kullback–Leibler divergence is positive definite, and the gradient of $W(\theta, \sigma^2)$ is

$$\begin{aligned} \nabla_\theta W(\theta, \sigma^2) &= \nabla_\theta \log \det \Gamma_\theta + (T - d)\, \nabla_\theta q(\theta)\, \sigma^{-2} \\ \nabla_{\sigma^2} W(\theta, \sigma^2) &= (T - d)\, \left(\sigma^{-2} - q(\theta)\, \sigma^{-4} \right). \end{aligned}$$

Setting to zero, we find that a critical point must satisfy $\sigma^2 = q(\theta)$, and plugging into the first equation, we must optimize $\mathcal{D}(\theta)$. The Hessian of $W(\theta, \sigma^2)$ evaluated at a critical point is

$$\nabla \nabla' W(\theta, \sigma^2) = \left[\begin{array}{cc} A & -(T-d)\, \nabla_\theta q(\theta)\, \sigma^{-4} \\ -(T-d)\, \nabla'_\theta q(\theta)\, \sigma^{-4} & T\, \sigma^{-4} \end{array} \right]$$

The eigenvalues of the Hessian consist of the positive eigenvalues of A and the Schur complement

$$(T-d)\,\sigma^{-4} - (T-d)^2\,\sigma^{-8}\,\nabla_\theta' q(\theta)\,A^{-1}\,\nabla_\theta q(\theta),$$

which must be positive; hence we conclude that $\nabla_\theta' q(\theta)\,A^{-1}\,\nabla_\theta q(\theta) < \sigma^4/(T-d)$. In summary, constrained maximization of $H(\theta,\sigma^2)$ and minimization of $W(\theta,\sigma^2)$ have the same critical points, namely $\sigma^2 = q(\theta)$ for θ minimizing $\mathcal{D}(\theta)$; these critical points must be minimizers of Kullback–Leibler divergence and maximizers of the Gaussian entropy. The function $\mathcal{D}(\theta)$ is just $\mathcal{W}(\theta, q(\theta))$, the concentration of Kullback–Leibler divergence with respect to innovation variance.

For the second part of the theorem, we subsitute some estimate of $q(\theta)$ to obtain an empirical entropy. Utilizing Q (5) as such an estimate – which introduces the parameter β through the mean vector η_β – yields the Gaussian divergence $\mathcal{L}$. The maximum entropy principle indicates that the minimizers $[\widehat{\beta}, \widehat{\theta}, \widehat{\sigma}^2]$ are maximum entropy estimators. □

Proof of Proposition 1. Using the notation of subsection 3.1, we have

$$\begin{aligned}
\mathbb{E}X_\partial &= [0\ 1_r]\,\Pi^{-1}\,\Lambda^{-1}\begin{bmatrix}\mu_* \\ \eta_\beta\end{bmatrix} \\
\mathrm{Cov}[X_\partial, X_\circ] &= [0\ 1_r]\,\Pi^{-1}\,\Lambda^{-1}\begin{bmatrix}\Sigma_* & 0 \\ 0 & \sigma^2\,\Gamma_\theta\end{bmatrix}\Lambda^\dagger\,\Pi^\dagger\,[1_m, 0]' \\
&= [0\ 1_r]\,\Pi^{-1}\,\Lambda^{-1}\begin{bmatrix}\Sigma_* & 0 \\ 0 & \sigma^2\,\Gamma_\theta\end{bmatrix}\begin{bmatrix}1_d & 0 \\ \underline{A} & \underline{B}\end{bmatrix}' Q^\dagger \\
\mathrm{Cov}[X_\partial, X_\circ]\,\mathrm{Cov}[X_\circ]^{-1} &= [0\ 1_r]\,\Pi^{-1}\,\Lambda^{-1}\begin{bmatrix}1_d & 0 \\ 0 & \Gamma_\theta\,C(\theta)\end{bmatrix} R.
\end{aligned}$$

Substituting the expression for $R\,X_\circ$ and simplifying yields the conditional expectation formula, noting that μ_* cancels out. The projection error is

$$X_\partial - \widehat{X_\partial} = [0\ 1_r]\,\Pi^{-1}\,\Lambda^{-1}\begin{bmatrix}0 \\ [1_{T-d} - \Gamma_\theta\,C(\theta)]\ (W - \eta_\beta)\end{bmatrix},$$

and computing the covariance yields the expression for $\Sigma_{\partial|\circ}$. □

References

[1] Ansley, C. and Kohn, R. (1985) Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *Ann. Statist.* **13**, 1286–1316.

[2] Bell, W. (1984) Signal extraction for nonstationary time series. *The Annals of Statistics* **12**, 646–664.

[3] Bell, W. and Hillmer, S. (1991) Initializing the Kalman filter for nonstationary time series models. *Journal of Time Series Analysis* **12**, 283–300.

[4] Brubacher, S.R. and Wilson, G.T. (1976) Interpolating time series with applications to the estimation of holiday effects on electricity demand. *Applied Statistics* **25**(2), 107–116.

[5] Chang, I., Tiao, G.C., and Chen, C. (1988) Estimation of time series parameters in the presence of outliers. *Technometrics* **30** (2), 193–204.

[6] Findley, D., Monsell, B., Bell, W., Otto, M., and Chen, B. (1998) New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economics Statistics* **16**, 127-177.

[7] Fox, A.J. (1972) Outliers in time series. *Journal of the Royal Statistical Society, Series B* **34**(3), 350–363.

[8] Golan, A., Judge, G., Miller, D., (1996) Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, New York.

[9] Gómez, V., Maravall, A., and Peña, D. (1999) Missing observations in ARIMA models: skipping approach versus additive outlier approach. *Journal of Econometrics* **88**, 341–363.

[10] Jaynes, E. T. (1957a) Information theory and statistical mechanics. *Physical Review, Series II* **106**(4), 620–630.

[11] Jaynes, E. T. (1957b) Information theory and statistical mechanics II. *Physical Review, Series II* **108**(2), 171–190.

[12] Ladiray, D. and Quenneville, B. (2001) *Seasonal adjustment with the X-11 method.* New York: Springer-Verlag, Vol. 158.

[13] Ljung, G. (1993) On outlier detection in time series. *Journal of the Royal Statistical Society, Series B* **55**(2), 559–567.

[14] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979) *Multivariate Analysis.* San Diego: Academic Press

[15] McElroy, T. (2008) Matrix Formulas for Nonstationary ARIMA Signal Extraction. *Econometric Theory* **24**, 1-22.

[16] McElroy, T. (2012) An Alternative Model-based Seasonal Adjustment that Reduces Over-Adjustment. *Taiwan Economic Forecast and Policy* **43**, 33–70.

[17] McElroy, T. (2016) On the Measurement and Treatment of Extremes in Time Series. *Extremes* **19**(3), 467–490.

[18] McElroy, T. and Monsell, B. (2015) Model estimation, prediction, and signal extraction for nonstationary stock and flow time series observed at mixed frequencies. *Journal of the American Statistical Association* **110**, 1284–1303.

[19] Park, S.Y. and Bera, A.K. (2009) Maximmum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics* **150**, 219–230.

[20] Peña, D. (1990) Influential observations in time series. *Journal of Business and Economics Statistics* **8**, 235–241.

[21] Trimbur, T. (2010) Stochastic level shifts and outliers and the dynamics of oil price movements. *International Journal of Forecasting* **26**, 162–179.

[22] U.S. Census Bureau (2011) *X-12-ARIMA Reference Manual.* `http://www.census.gov/ts/x12a/v03/x12adocV03.pdf`

[23] Yamada, H. and Jin, L. (2013) Japan's output gap estimation and ℓ_1 trend filtering. *Empirical Economics* **45**, 81–88.

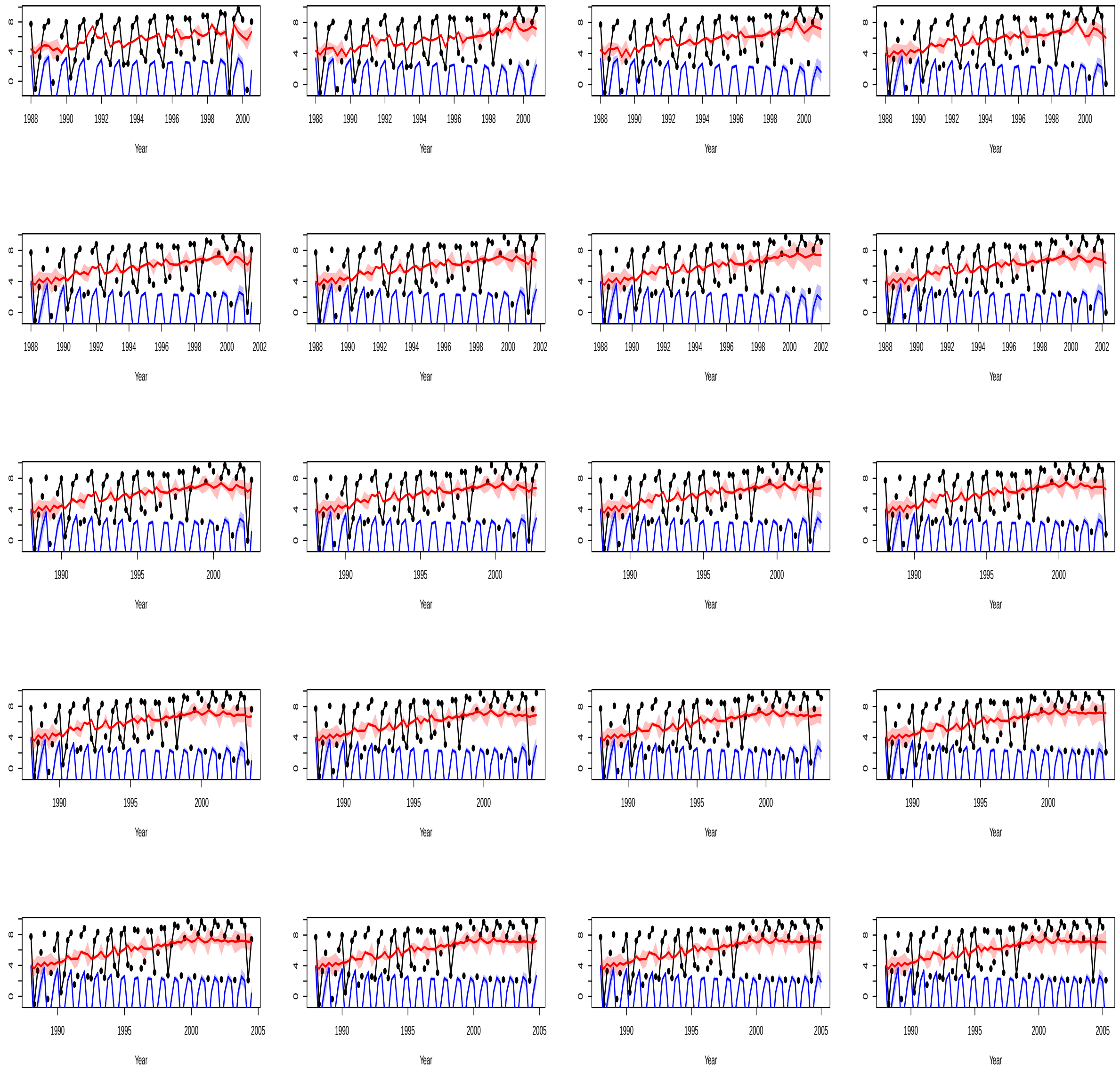

Figure 3: Log scale of Avocado data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Reading left to right is the addition of a new quarter, while reading top to bottom gives the addition of another year of data. These results correspond to $\alpha = .05$ in the extreme-value test statistic.

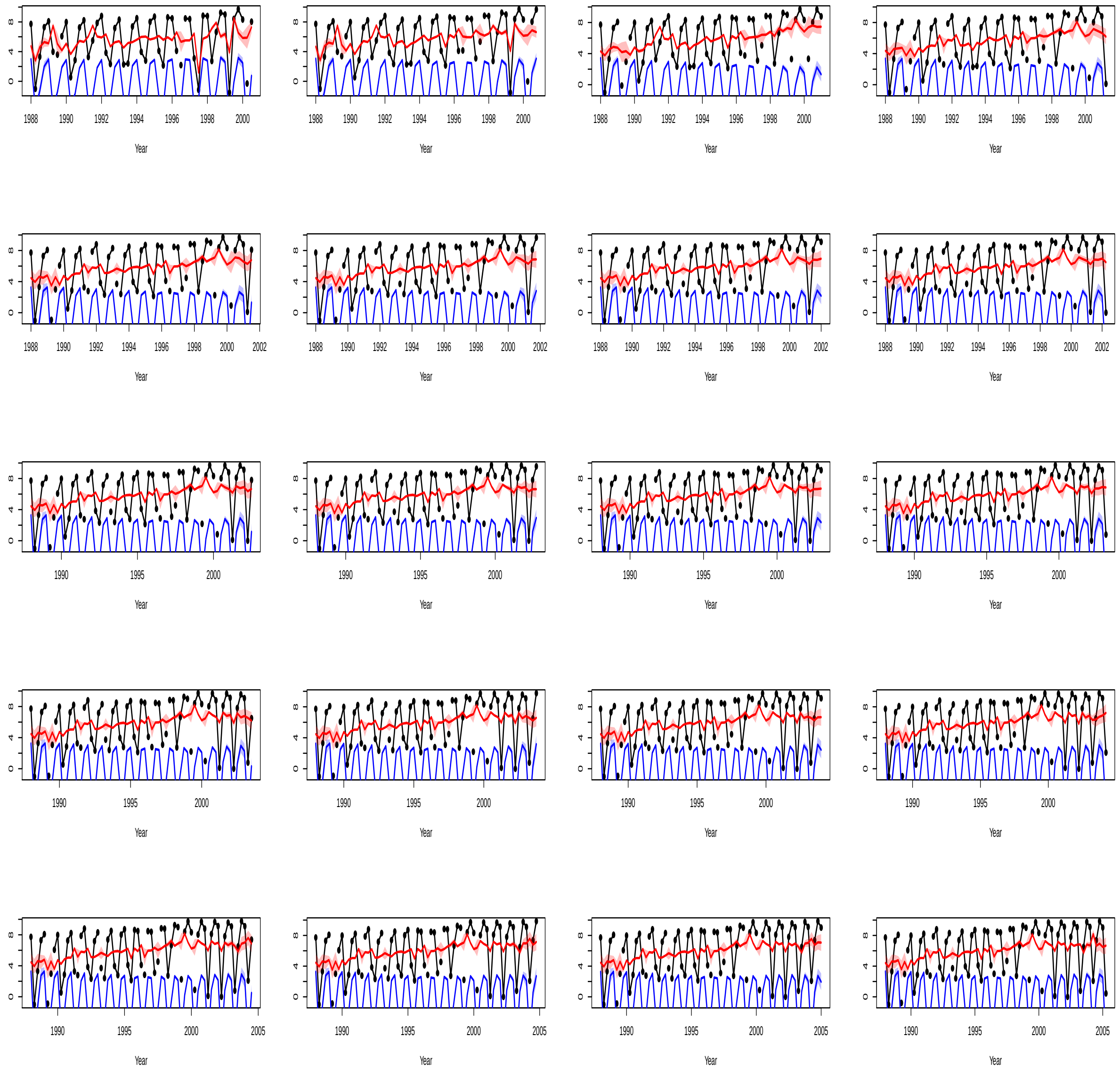

Figure 4: Log scale of Avocado data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Reading left to right is the addition of a new quarter, while reading top to bottom gives the addition of another year of data. These results correspond to $\alpha = .01$ in the extreme-value test statistic.

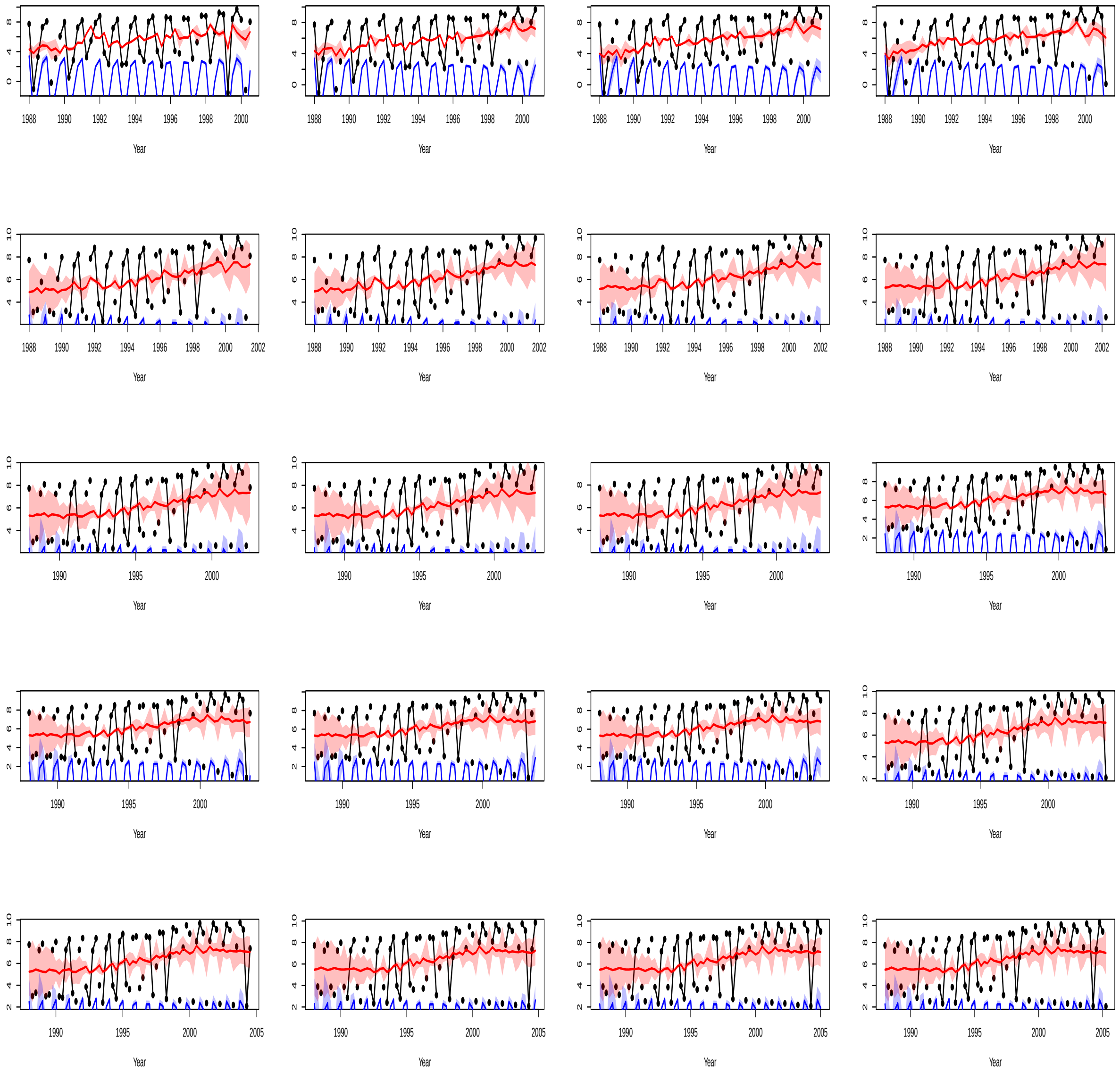

Figure 5: Log scale of Avocado data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Reading left to right is the addition of a new quarter, while reading top to bottom gives the addition of another year of data. These results correspond to $\alpha = .10$ in the extreme-value test statistic.

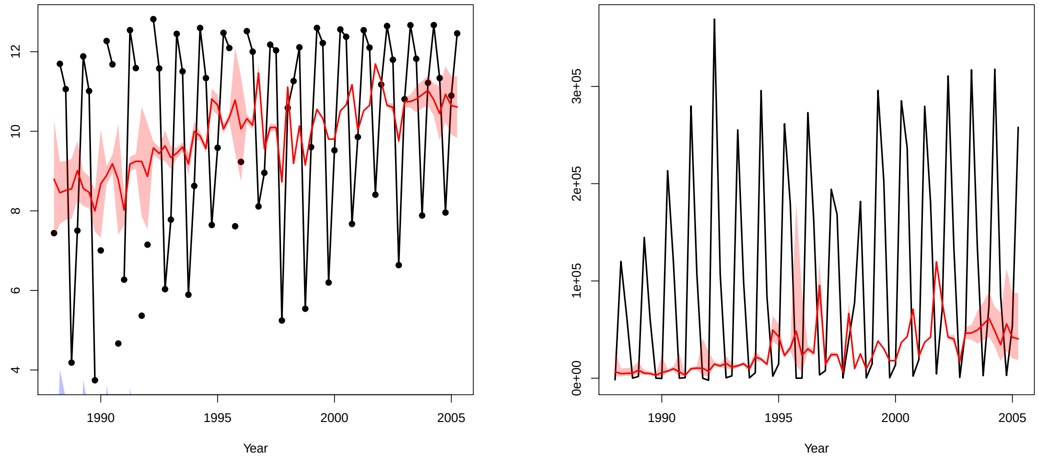

Figure 6: Left panel: Log scale of Apple data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Right panel: data in original scale, with exponential of the X12 seasonal adjustment (red). Confidence intervals are shaded.

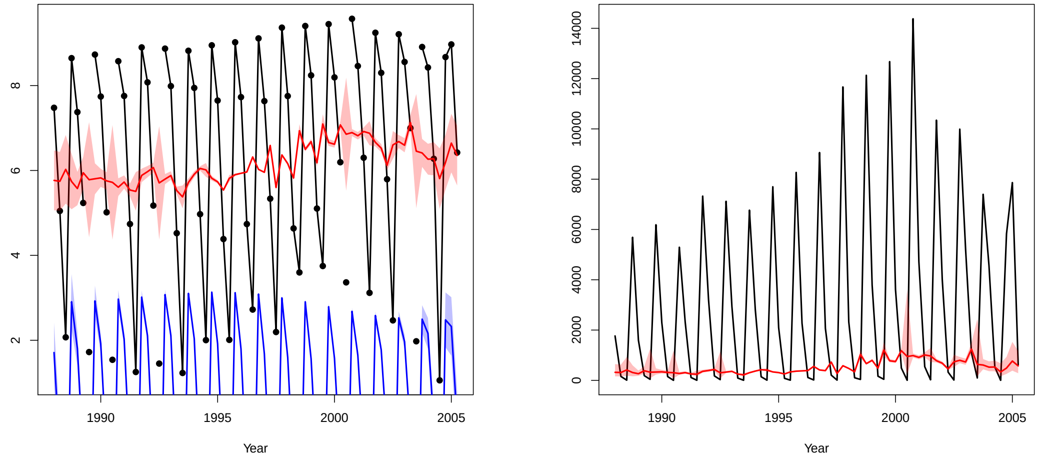

Figure 7: Left panel: Log scale of Berry data, with imputed values as black dots, with X12 seasonal adjustment (red), and with extracted seasonal (blue). Right panel: data in original scale, with exponential of the X12 seasonal adjustment (red). Confidence intervals are shaded.