Journal of Data Science **8**(2010), 555-577

Edition and Imputation of Multiple Time Series Data Generated by Repetitive Surveys

Victor M. Guerrero[1,2] and Blanca I. Gaspar[2,3]
[1]*Instituto Tecnológico Autónomo de México,* [2]*Instituto Nacional de Estadística y Geografía and* [3]*Banco de México*

Abstract: This paper considers the statistical problems of editing and imputing data of multiple time series generated by repetitive surveys. The case under study is that of the Survey of Cattle Slaughter in Mexico's Municipal Abattoirs. The proposed procedure consists of two phases; firstly the data of each abattoir are edited to correct them for gross inconsistencies. Secondly, the missing data are imputed by means of restricted forecasting. This method uses all the historical and current information available for the abattoir, as well as multiple time series models from which efficient estimates of the missing data are obtained. Some empirical examples are shown to illustrate the usefulness of the method in practice.

Key words: Compatibility tests, mean square error, missing data, restricted forecasts, VAR models.

1. Introduction

The National Institute of Statistics and Geography (INEGI) carries out the Survey of Cattle Slaughter in Mexico's Municipal Abattoirs (ESGRM for its name in Spanish). This repetitive survey captures monthly data for each abattoir in a questionnaire that asks questions about the slaughter of cattle for human consumption. Four species of animals are considered here: cattle, swine, sheep and goats. Even though INEGI puts a lot of effort to collect and publish trustworthy data, it is a fact that the quality of some statistical figures published by this official statistical agency can still be greatly improved. Such is the case of the data generated by the ESGRM, since this survey presents the typical problems of: (1) inconsistency of the collected data (the informant at the abattoir responded to the questionnaire, but the answers are not considered valid by some criteria used to verify the information), and (2) missing data (at least one of the variables lacks its value requested in the questionnaire). These problems create the necessity of applying statistical procedures for editing and imputing data. Such tasks should

be done for each questionnaire (at the abattoir level) to avoid the accumulation of errors when aggregating data of two or more abattoirs. Therefore, it is desirable to use editing and imputing procedures with solid statistical foundations, which can also be computationally automated for massive and repetitive application (in all the municipal abattoirs, month after month).

In this work we propose a statistical methodology that is supported by multiple time series models. These models take into account historical information on the variables under study as well as their possible interrelations. The following three basic variables of the ESGRM were considered relevant: (1) number of heads (number of animals that are introduced alive to the abattoir), (2) weight on the hoof (weight of the live animal when entering the abattoir) and (3) weight of the beef carcasses (weight of the slaughtered animal after taking out some of its parts, like its skin, its head and its offal).

There is a large body of literature dealing with the problem of missing data in different types of surveys. We refer the reader to such authoritative works as Little and Rubin (1987) and Schafer (1997) for tools designed to perform statistical analysis of incomplete multivariate observations and to Zhang (2003) for a review of multiple imputation methods in use nowadays. Here we just consider the issue of imputation without taking into account the subsequent analyses of the data, because INEGI is a national statistical agency in charge of collecting and publishing data for the general public and it does not necessarily analyze the data.

With regard to the missing data problem in a univariate time series setting, some influential works are those of Kohn and Ansley (1986) and Gómez *et al.* (1999), although some other works have appeared in the literature (e.g. Guerrero, 1994). All these works suggest building an Auto-Regressive Integrated Moving Average (ARIMA) model for the available data, then use all the data (observed before and after the missing ones) to get efficient estimates of the unobserved values. The problem in this context is also known as interpolation and it consists in essence of predicting the outcome of the unobserved variable by means of the expected value of the predictive distribution. A particular work that deals with edition and imputation, considered as tools for quality control of univariate time series data, is that of Caporello and Maravall (2002).

In the multiple time series case we found only a few proposals, even though the problem appeared in the literature as early as 1974. The solution proposed by Sargan and Drettakis (1974) was very thorough, but the authors accepted explicitly that their method was difficult to apply in practice. The proposal by Luceño (1997) is more general than the previous one, since it can be used with vector Auto-Regressive and Moving Average (ARMA) models, but it relies on Maximum Likelihood Estimation, so that it also becomes difficult for massive

application. Pfefferman and Nathan (2002) suggested another method, based on a multivariate state-space representation of the time series that relies also on Maximum Likelihood Estimation. The procedure suggested here differs from the already existing methods in that it is designed for a multiple time series that can be represented by a simple VAR (Vector Auto-Regressive) model that can be efficiently estimated by Ordinary Least Squares, equation by equation, a fact that simplifies considerably its practical implementation. It should be noticed that all the aforementioned methods aim to produce minimum Mean Square Error (MSE) estimates of the missing values. This fact differs from the usual approach used to impute values by drawing simulated values from a predictive distribution. An exception is Pfefferman and Nathan (2002) who did obtain the simulated values to represent uncertainty more appropriately. Thus, in our case we do not attempt to refer to the imputed values as simulated individual observations, but as estimated expected values. In the next section we describe the ESGRM and present some graphs that allow us to appreciate the typical dynamic behavior of the variables under study, along with some suggested transformations, which will serve mainly to edit the data. In the section after we present the VAR model and the restricted forecasting methodology that will be used to impute the missing data. Then, we present some aspects of the model building process and the estimation results, for the abattoir and animal species under consideration. Afterwards we present some results produced by the methodology, as much in the edition as in the imputation of data for the ESGRM. In that section we also show the results of some simulations carried out to verify the effectiveness of the method in practice. We conclude with some practical considerations.

2. Preliminary Data Analysis

The ESGRM covers the 31 States of Mexico where abattoirs are in operation. This study made use of data from the abattoirs of a State whose name is omitted for confidentiality reasons. Only the results of one abattoir will be used to illustrate the methodology, although the intention is to apply it to all the existing municipal abattoirs in the country (this number changes every year, there were 907 abattoirs in December, 2004 and 890 in December, 2007).

The sample period covers data from January of 1998 to December of 2003, since those were the historical data available in the more recent annual publication (see INEGI, 2004) at the beginning of the study. It should be stressed that we had data up to December, 2003 and that we required estimating old data to get a complete record of multiple time series in order to build a model and start the repetitive (monthly) application of our methodology. The joint dynamic behavior of the three variables for each abattoir of the State in consideration was represented by a VAR model. We searched for a generic model and found

a specification that provides reasonably valid results, in statistical terms, for all the abattoirs of the State.

2.1 Definition of concepts

The following concepts are used in the ESGRM. A municipal abattoir is the basic unit of observation, defined as the building where the slaughter of animals for human consumption takes place. The animal species considered are: cattle (includes bulls, oxen, cows, heifers and yearling calves), swine (includes pigs), sheep (includes lambs), and goats.

Two constructed variables: average weight on the hoof (weight on the hoof divided by number of heads) and yield of meat (ratio of weight of the beef carcass to weight on the hoof) were used to validate the collected data. We observed that these variables tend to stay near a constant value for several consecutive months. In Figure 1 we show an example of this phenomenon for an abattoir, called A for confidentiality of the data. The species slaughtered are cattle and swine.

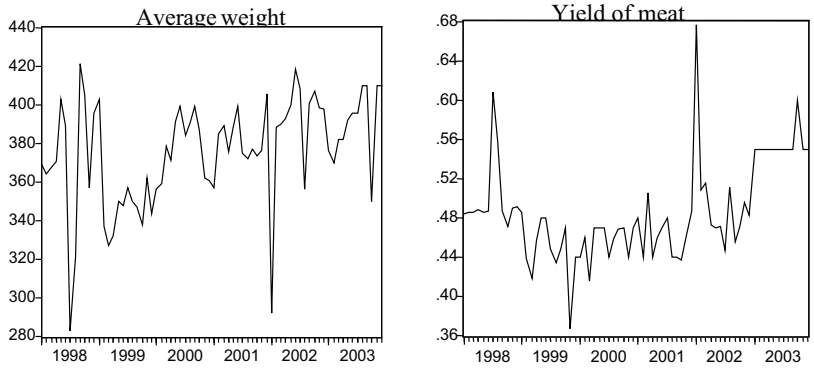

Figure 1: Average weight on the hoof and yield of meat (Cattle)

The basic behavior observed in these graphs is repeated for all the animal species and abattoirs in consideration. This fact led us to think that the informants tend to report numbers close to the average values or within certain bounds of the constructed variables. The data are usually reported with evident errors and the INEGI personnel edit some of those errors when capturing the data every month, since this is a repetitive survey. Even with this first validation of data, several inconsistencies remain unnoticed by the capture personnel. Besides, some new errors are introduced when capturing the data. Thus, in order to correct these errors in a systematic way, we propose to start the editing process by taking into account the permissible values of the constructed variables. The aggregated

figures for all the abattoirs in each Mexican State that appear in INEGI (2004) are official and definitive, nevertheless it was necessary to perform a preliminary analysis of the data pertaining to each and every abattoir under study to make sure that all the data were within the permissible limits. This was done for each and every variable. In fact, if a recorded observation of a variable fell outside those limits, it was replaced by a permissible value, while the observations of the remaining variables were not affected. This way we obtained a set of edited historical time series. Then, we applied some transformations that allowed us to see the natural variability of the data more clearly and proceeded to build multiple time series models, as described below. The transformation procedure must be applied month after month as new data arrive, in order to force them to be consistent with each other and with their historical records. Afterwards, we can use the imputation methodology to estimate missing data in the transformed scale by means of restricted forecasting. The restricted forecasts provide optimal estimates (in a statistical sense defined below) of the missing data. Finally, we retransform the estimates back to the original scale of the variables.

2.2 Transformation of variables

The transformations used are intended to complement the data edition in such a way that the transformed data satisfy the criteria that informants have been trying to apply routinely, although in an informal way. Therefore, transformation of the data is a fundamental part of the editing procedure. The variables are defined for month t as follows; NH_t denotes number of heads, WH_t weight on the hoof measured in kilograms, and WB_t weight of beef carcasses also measured in kilograms. Since $0 < NH_t$ for $t = 1, ..., N$, the proposed transformation becomes

$$-\infty < TNH_t = \log\left(\frac{NH_t}{NH_{t-1}}\right) < \infty$$

and the variable in differences that will appear in the VAR model is

$$DTNHT_t = \log(NH_t/NH_{t-1}) - \log(NH_{t-1}/NH_{t-2}) \approx r_t^{NH} - r_{t-1}^{NH}$$

with r_t^{NH} the relative growth rate of NH_t. The reason for using differences will be explained when considering the model building process.

For the average weight, it is well known that the following restriction must hold, $K1 \leq WH_t/NH_t \leq K2$ where $K1$ and $K2$ are some known structural constants (e.g. for cattle, $K1 = 250$ and $K2 = 550$). Then we know that $K1 < WH_t/NH_t + 0.1$ and $WH_t/NH_t < K2 + 0.1$. The constant 0.1 was added to get strict inequalities without affecting the dynamics of the series involved and its inclusion amounts to extending the original interval by 100 grams at each side.

The proposed transformation is now

$$-\infty < TWH_t = \log\left(\frac{WH_t/NH_t + 0.1 - K1}{K2 - WH_t/NH_t + 0.1}\right) < \infty$$

and the variable to use in the model turns out to be

$$\begin{aligned} DTWH_t &= \log\left(\frac{WH_t/NH_t + 0.1 - K1}{K2 - WH_t/NH_t + 0.1} \Big/ \frac{WH_{t-1}/NH_{t-1} + 0.1 - K1}{K2 - WH_{t-1}/NH_{t-1} + 0.1}\right) \\ &\approx r_t^{WH/NH-K1} - r_t^{K2-WH/NH} \end{aligned}$$

It is also known that the yield of meat must satisfy the restriction $K3 \leq WB_t/WH_t \leq K4$ for all t, with $K3$ and $K4$ some known positive structural constants (for example, for cattle $K3 = 0.5$ and $K4 = 0.55$). We decided to use the ratio WB/NH, rather than WB/WH, because NH usually has complete data, in contrast with WH that usually lacks some data values. To make this change, we know that

$$K1(WB_t/WH_t) \leq (WH_t/NH_t)(WB_t/NH_T) \leq K2(WB_t/WH_t)$$

from which we get $K1(K3) \leq WB_t/NH_t \leq K2(K4)$. Therefore, as in the previous case, the transformation proposed is

$$-\infty < TPC_t = \log\left(\frac{WB_t/NH_t + 0.01 - K1 \cdot K3}{K2 \cdot K4 - WB_t/NH_t + 0.01}\right) < \infty.$$

The constant 0.01 was chosen as in the previous case and its use implies that the interval $[K1 \cdot K3, K2 \cdot K4]$ is extended 10 grams at each side. Therefore, the variable to be used in the model is

$$DTWB_t \approx r_t^{WB/NH-K1\cdot K3} - r_t^{K2\cdot K4-WB/NH}$$

The structural constants $K1, K2, K3$ and $K4$ used here for transforming the data are usually employed for descriptive purposes (see Ruiz *et al.*, 2001). Since there is a great variety of animal species, classified by age and race, we decided to look for the most commonly used values in each Mexican State. We found some values on the following websites: Faculty of Veterinary Medicine at the National Autonomous University of Mexico (Veterinary, 2005, www.veterin.unam.mx), Secretariat of Agriculture (SAGARPA, 2005, www.sagarpa.gob.mx/ganaderito) and cattle dealer associations (Cattle dealers, 2005, www.mexicoganadero.com/limousin). The final decision of which values to use was made by asking for advice to the veterinary personnel that works at the municipal abattoirs in the State under consideration. Figure 2 shows the behavior of the time series in the original scale (NH, WH and WB), as well as that of the transformed variables (TNH, TWH

and TWB) and their first differences ($DTNH, DTWH$ and $DTWB$), for the abattoir under consideration and the two species of animals. There we can see that the original series have trend and seasonality, we also see that the transformation not only extends the numerical scale from $(0, \infty)$ to $(-\infty, \infty)$ but also helps to stabilize the trend in the data. Differencing ensures stationarity of the time series and makes the data fluctuate around zero.

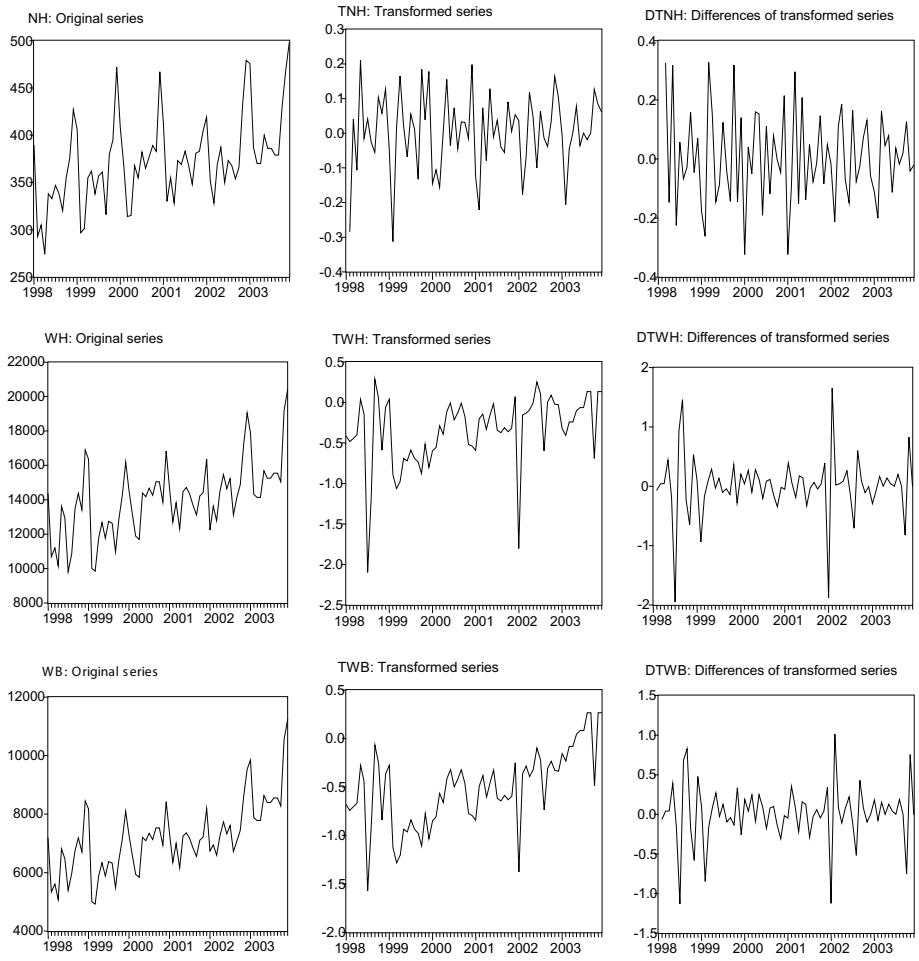

Figure 2: Original and transformed data (Cattle)

3. Statistical Methodology

The most important methodological aspect of this work is the use of restricted forecasting, supported by multiple time series models. Such models allow us to

capture the dynamics of the time series and here we will place special emphasis on their use for imputing missing data of the ESGRM.

3.1 VAR model

Let $\mathbf{Z}_t = (Z_{1t}, \ldots, Z_{kt})'$ be a k-dimensional multiple time series observed for $t = 1, \cdots, N$. The VAR representation for this time series is

$$\Pi(B)\mathbf{Z}_t = \Lambda\mathbf{D}_t + \mathbf{a}_t \tag{3.1}$$

where $\Pi(B)$ denotes a matrix polynomial of order $p < \infty$, in the backshift operator B, such that $BZ_t = Z_{t-1}$ for every Z and t, that is, $\Pi(B) = I_k - \Pi_1 1B - \cdots - \Pi_p B^p$ with I_k the identity matrix of order k and

$$\Pi_j = \begin{pmatrix} \pi_{j,11} & \pi_{j,12} & \cdots & \pi_{j,1k} \\ \pi_{j,21} & \pi_{j,22} & \cdots & \pi_{j,2k} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{j,k1} & \pi_{j,k1} & \cdots & \pi_{j,kk} \end{pmatrix}$$

for $j = 1, \cdots, p$. This model takes into account the deterministic vector of variables $\mathbf{D}_t = (D_{1t}, \ldots, D_{kt})'$ that may include constant levels, seasonal dummies and intervention variables whose effects on $\mathbf{Z}_t$ are captured by the parameter matrix $\Lambda\{\mathbf{a}_t\}$ is a zero-mean Gaussian white noise process, so that the $\mathbf{a}_t$'s are independent and identically distributed as $\mathbf{a}_t \sim N_k(\mathbf{0}_k, \Sigma_\mathbf{a})$ for $t = 1, \cdots, N$, where $\mathbf{0}_k$ is the zero vector and $\Sigma_\mathbf{a}$ is the contemporaneous error variance-covariance matrix, given by

$$\Sigma_\mathbf{a} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{pmatrix}$$

with $\sigma_{ij} = Cov(a_{it}, a_{jt})$ for $i, j = 1, ..., k$.

When the series $\{\mathbf{Z}_t\}$ is stationary, model (3.1) is well defined. Otherwise, for the model to be well defined, it is necessary to assume that the process started at a finite time point in the past, with fixed initial conditions. When the individual time series are all integrated of at least order one and they are not cointegrated, we can apply the difference operator $\bigtriangledown = 1 - B$ to each series in order to get an appropriate VAR representation for the new time series $\{\nabla\mathbf{Z}_t\}$. On the other hand, when a cointegration relationship exists among the variables (see Engle and Granger, 1987) we could work with the following VAR model in Error Correction (VEC) form

$$\Pi^*(B) = \nabla\mathbf{Z}_t = \Lambda\mathbf{D}_t - \Pi(1)\mathbf{Z}_{t-1} + \mathbf{a}_t. \tag{3.2}$$

This model arises from the equation $\Pi(B) = \Pi(1)B + \Pi^*(B)\nabla$. In this case, the matrix polynomial $\Pi^*(B)$ is of order $p-1$, whereas the matrix $\Pi(1)$ is defined in accordance with the cointegration relationships that exist among the variables.

Let us notice that $\Pi^*(B)\nabla\mathbf{Z}_t$ captures the short-run relations in $\mathbf{Z}_t$, while $\Pi(1)\mathbf{Z}_{t-1}$ represents the long-run relations and all the elements of equation (3.2) are stationary. The VEC model is related to economic theory since it allows interpreting the results and making inferences about both short-run and long-run economic relations. However, it should be noticed that the information captured by the VAR and the VEC models is exactly the same, even though their representations are formally different. In the present case, we only want to get one-step-ahead forecasts from the model and we are not interested in the long-run relations that may exist among the variables. Hence, in what follows we will use the VAR form (so that no cointegration analysis is required) and the restricted forecasting methodology will only be presented for that model.

3.2 Restricted forecasts

Here we will assume that the model and its parameters are known, so that we do not consider at this time such issues as specification, estimation and validation of the model. In practice however, those issues have to be faced as in the illustrative application shown below. The vector $\mathbf{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_N')'$ has the observations of $\mathbf{Z}_t = (TNH_t, TWH_t, TWB_t)'$ for $t = 1, \ldots, N$, while $\mathbf{Z}_{N+1}$ is the vector of future values to be forecast, with origin at time N. Since the stationary series is $\nabla\mathbf{Z}_t = (DTNH_t, DTWH_t, DTWB_t)'$ the VAR model to be used becomes

$$\Pi(B)\nabla\mathbf{Z}_t = \Lambda\mathbf{D}_t + \mathbf{a}_t.$$

Then, the minimum Mean Square Error (MSE) linear forecast of $\nabla\mathbf{Z}_{N+1}$ is its conditional expectation given all the historical information, that is,

$$E(\nabla\mathbf{Z}_{N+1}|\mathbf{Z}) = \Lambda\mathbf{D}_{N+1} + \Pi_1 E(\nabla\mathbf{Z}_N|\mathbf{Z}) + \cdots + \Pi_p E(\nabla\mathbf{Z}_{N+1-p}|\mathbf{Z}).$$

The corresponding forecast error is given by

$$\nabla\mathbf{Z}_{N+1} - E(\nabla\mathbf{Z}_{N+1}|\mathbf{Z}) = \mathbf{a}_{N+1} \tag{3.3}$$

and its MSE matrix is

$$MSE[E(\nabla\mathbf{Z}_{N+1}|\mathbf{Z})] = Var(\mathbf{a}_{N+1}|\mathbf{Z}) = \Sigma_{\mathbf{a}}.$$

Now, suppose that we also know the vector of observations $\mathbf{Y} = (Y_1, \ldots, Y_M)'$ that imposes $M \geq 0$ linearly independent restrictions on the future values of the

vector $\mathbf{Z}$. Such restrictions come from an external source to the time series model and can be expressed as

$$\mathbf{Y} = C\nabla\mathbf{Z}_{N+1} + \mathbf{u} \tag{3.4}$$

where $\mathbf{u} = (u_1, \ldots, u_M)'$ is a random vector distributed as $N(\mathbf{0}_M, \Sigma_\mathbf{u})$. The $M \times k$ matrix C is known and has rank $M \leq k$. We will show below some particular forms of C for the imputing problem faced by the ESGRM. In those cases, the data $Y_1, \ldots, Y_M$ are the observed values of the variables at time $N+1$.

Expression (3.4) can be deemed as a set of stochastic linear restrictions to be imposed on $\nabla\mathbf{Z}_{N+1}$. We assume $E(\mathbf{u}|\mathbf{Z}) = \mathbf{0}$, in such a way that the restrictions are unbiased and their uncertainty is linked to the variance $Var(\mathbf{u}|\mathbf{Z}) = \Sigma_\mathbf{u}$, which in the present application is $\Sigma_\mathbf{u} = \mathbf{0}$. If $E(\mathbf{a}_{N+1}\mathbf{u}'|\mathbf{Z}) = 0$ we get the following optimal restricted forecast of $\nabla\mathbf{Z}_{N+1}$ given $\mathbf{Z}$ and $\mathbf{Y}$,

$$\nabla\hat{\mathbf{Z}}_{N+1} = E(\nabla\mathbf{Z}_{N+1}|\mathbf{Z}) + A[\mathbf{Y} - CE(\nabla\mathbf{Z}_{N+1}|\mathbf{Z})] \tag{3.5}$$

with

$$A = \Sigma_\mathbf{a}C'(C\Sigma_\mathbf{a}C')^{-1} \tag{3.6}$$

and

$$MSE(\nabla\hat{\mathbf{Z}}_{N+1}) = (\mathbf{I}_k - AC)\Sigma_\mathbf{a}. \tag{3.7}$$

This result is shown in the Appendix for the general case $\Sigma_\mathbf{u} \neq \mathbf{0}$ (a different proof, based on signal extraction techniques, can be found in Pankratz, 1989).

It should be stressed that the linear combination (3.5) is optimal in the sense of having minimum MSE within the class of linear predictors for $\nabla\mathbf{Z}_{N+1}$. Further, the restricted forecast has lower MSE than that of the unrestricted forecast because

$$MSE[E(\nabla\mathbf{Z}_{N+1}|\mathbf{Z})] = MSE(\nabla\hat{\mathbf{Z}}_{N+1}) + AC\Sigma_\mathbf{a}$$

where $AC\Sigma_\mathbf{a} = \Sigma_\mathbf{a}C'(C\Sigma_\mathbf{a}C')^{-1}C\Sigma_\mathbf{a}$ is a positive semi definite matrix.

3.3 Forecasts in the original scale

To obtain the forecasts of NH_{N+1}, WH_{N+1} and WB_{N+1} from the forecasts of $\nabla\mathbf{Z}_{N+1} = (DTNH_{N+1}, DTWH_{N+1}, DTWB_{N+1})'$, we first obtain the forecasts in transformed levels by means of $Z_{N+1} = Z_N + DZ_{N+1}$ for $Z = TNH, TWH, TWB$. Then we go back to the units of the original variables as follows; for NH we know that $\exp(TNH_{N+1}) = NH_{N+1}/NH_N$, so that $NH_{N+1} = NH_N \cdot \exp(TNH_{N+1})$. For WH we have

$$\frac{WH_{N+1}}{NH_{N+1}} = (K2 + 0.1)\exp(TWH_{N+1}) + K1 - 0.1 - \frac{WH_{N+1}}{NH_{N+1}}\exp(TWH_{N+1})$$

hence,

$$WH_{N+1} = NH_{N+1}[(K2+0.1)\exp(TWH_{N+1}) + K1 - 0.1]/[1+\exp(TWH_{N+1})].$$

Similarly, for WB we get

$$WB_{N+1} = \frac{NH_{N+1}[(K2 \cdot K4 + 0.01)\exp(TWB_{N+1}) + K1 \cdot K3 - 0.01]}{[1+\exp(TWB_{N+1})]}.$$

When the value of WH_{N+1} is known we should use it to get the forecast of WB_{N+1} that satisfies the true restriction $K3 \leq WB_{N+1}/WH_{N+1} \leq K4$. To do that, instead of the previous expression we should use

$$WB_{N+1} = WH_{N+1}[(K4+0.01)\exp(TWB_{N+1}) + K3 - 0.01]/[1+\exp(TWB_{N+1})].$$

It is worth noticing that the log transformation is nonlinear, so that backtransforming the forecasts produced by the model in the transformed scale induces bias on the forecasts in the original scale. That happens because the forecast represents a median value in the original scale and the analyst usually expects the forecast to represent a mean value. A correction for bias may be applied, as indicated in Guerrero (1993) or, as we prefer in this case, the uncorrected forecasts in the original scale should be interpreted as median values.

3.4 Compatibility test

Since the optimal forecast implies combining information from two different sources, we should be aware of the possibility of combining contradictory information. Thus, we propose to judge the validity of this combination empirically. We say that the extra-model information $\mathbf{Y}$, is compatible with the information provided by the model, $CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})$, if the distance between those vectors is close to zero. We define the random vector of differences

$$\mathbf{d} = \mathbf{Y} - CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z}) = C\mathbf{a}_{N+1}$$

which is distributed as $N(\mathbf{0}_M, C\Sigma_{\mathbf{a}}C')$. Therefore, the distance that takes into account the variability of $\mathbf{d}$ yields the statistic $K = \mathbf{d}'(C\Sigma_{\mathbf{a}}C')^{-1}\mathbf{d} \sim \chi^2_M$.

It follows that $\mathbf{Y} - CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})$ belongs in the compatibility region if

$$K_{cadc} = [\mathbf{Y} - CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})]'(C\Sigma_{\mathbf{a}}C')^{-1}[\mathbf{Y} - CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})] \leq \chi^2_M(\alpha) \quad (3.8)$$

with $\chi^2_M(\alpha)$ the upper α percentage point of the Chi-square distribution with M degrees of freedom. Equivalently, $\mathbf{Y}$ is incompatible with $CE(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})$, at the $100\alpha\%$, significance level, if (3.8) does not hold. This decision rule is based on the assumption that all the model parameters are known. Then, even when the

parameters are consistently estimated, the rule is only asymptotically valid. In case of incompatibility we may conclude that the observed data at time $N+1$ are atypical.

3.5 Restricted forecasts for the ESGRM

In the case of the ESGRM, the extra-model information is the observed data at time $t = N+1$ and the unrestricted VAR forecasts are given by

$$E(\nabla \mathbf{Z}_{N+1}|\mathbf{Z}) = (E(DTNH_{N+1}|\mathbf{Z}), E(DTWH_{N+1}|\mathbf{Z}, E(DTWB_{N+1}|\mathbf{Z})).$$

The pattern followed by the missing data gives raise to eight different cases: zero observations missing, one observation missing, (NH, WH or WB), two observations missing (NH and WH, NH and WB or WH and WB) or the three observations missing. These cases can be expressed in terms of the arrays $C, \mathbf{Y}$ and

$$\nabla \hat{\mathbf{Z}}_{N+1} = (\widehat{DTNH}_{N+1}, \widehat{DTWH}_{N+1}, \widehat{DTWB}_{N+1})'$$

that appear in expressions (3.5)-(3.7) and (3.8). For instance, when no data are missing at $t = N+1$, we have $C = \mathbf{I}_3$ and $\mathbf{Y} = \nabla \mathbf{Z}_{N+1}$ so that the restricted forecasts are given by $\nabla \hat{\mathbf{Z}}_{N+1} = \mathbf{Y}$, with $\widehat{MSE}(\nabla \hat{\mathbf{Z}}_{N+1}) = 0$. Besides, the statistic $K_{calc} = [\mathbf{Y} - E(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})]' \hat{\Sigma}_{\mathbf{a}} [\mathbf{Y} - E(\nabla \mathbf{Z}_{N+1}|\mathbf{Z})]$ allows us to validate the joint compatibility of the three newly arrived data with their unrestricted forecasts, by comparing its value against a Chi-square distribution with 3 degrees of freedom.

Another example of how the arrays are specified to obtain the restricted forecasts is when $DTWH_{N+1}$ and $DTWB_{N+1}$ are missing. Then $C = (1, 0, 0), \mathbf{Y} = DTNH_{N+1}$ and

$$\nabla \hat{\mathbf{Z}}_{N+1} = \begin{pmatrix} DTNH_{N+1} \\ E(DEWH_{N+1}|\mathbf{Z}) + \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{11}}[DTNH_{N+1} - E(DTNH_{N+1}|\mathbf{Z})] \\ E(DEWH_{N+1}|\mathbf{Z}) + \frac{\hat{\sigma}_{13}}{\hat{\sigma}_{11}}[DTNH_{N+1} - E(DTNH_{N+1}|\mathbf{Z})] \end{pmatrix}$$

with

$$\widehat{MSE} \begin{pmatrix} \widehat{DTWH}_{N+1} \\ \widehat{DTWB}_{N+1} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_{22} - \hat{\sigma}_{12}^2/\hat{\sigma}_{11} & \hat{\sigma}_{23} - \hat{\sigma}_{12}\hat{\sigma}_{13}/\hat{\sigma}_{11} \\ \hat{\sigma}_{23} - \hat{\sigma}_{12}\hat{\sigma}_{13}/\hat{\sigma}_{11} & \hat{\sigma}_{33} - \hat{\sigma}_{13}^2/\hat{\sigma}_{11} \end{pmatrix},$$

together with the compatibility statistic $K_\alpha = [DTNH_{N+1} - E(DTNH_{N+1}|\mathbf{Z})]^2 / \hat{\sigma}_{11}$. We should notice in these expressions that the restricted forecasts for data actually observed yield exactly those observations, so that the restrictions imposed are satisfied exactly. Furthermore, the estimated MSE of the restricted forecasts are smaller than those for the unrestricted forecasts.

4. Building the VAR Model

The VAR model was built from adjusted data (edited as indicated previously, with no allowance for outliers) following standard procedures. That is, we first decided the degree of differencing that renders each individual series stationary. To that end we applied the variate difference method (see Anderson, 1976) on the transformed series. That is, out of those series with successive degrees of differencing, the one with minimum standard deviation should be used. Of course, we could have used a unit root test, implying a case-by-case analysis for each species. That would have made the proposed method difficult to apply massively (for all the abattoirs in Mexico).

4.1 Degree of differencing

The results of applying the variate difference method are shown in Table 1. There we see that the transformed series requires at most one difference to become stationary. This pattern was observed for all the species at every abattoir under study. Thus, by applying one difference to the series we will achieve stationarity, although we may run into overdifferencing, but we did not consider that to be as serious a problem as that of underdifferencing, because the model is only going to be used to produce one-step-ahead forecasts. We refer the reader to Sánchez and Peña (2001) for an argument that favors overdifferencing to underdifferencing a time series, when using an autoregressive model to produce forecasts. Let us recall that we were looking for a generic transformation to stationarity that could be applied to all the series, in order for the procedure to be computationally automated for massive and repetitive application (so that it could be used as a canned package by the capture personnel).

Table 1: Standard deviation for successive degrees of differencing

Variable	Species	Degree of differencing		
		0	1	2
TNH	Cattle	0.11	0.15	0.26
	Swine	0.15	0.23	0.42
TWH	Cattle	0.43	0.52	0.84
	Swine	3.05	3.22	5.27
TWB	Cattle	0.40	0.36	0.58
	Swine	3.27	3.39	5.54

4.2 Likelihood ratio tests

Once stationarity was achieved, we had to decide the VAR order for each species at every abattoir. Thus, we applied a likelihood ratio testing scheme (see Lütkepohl, 2005, pp. 128-144) that allowed us to test sequentially the hypotheses H_0 order p vs. H_A: order $p-1$, starting with the value $p=5$. Some examples of the tests results are shown in Table 2. We found a generic specification with $p=3$, since that order yielded the first significant value of χ^2 in most cases, when we reduced the number of lags.

Table 2: Likelihood ratio test results

p (lags)	Cattle		Swine	
	χ^2	Signif.	χ^2	Signif.
5	13.28	0.15	11.18	0.26
4	4.99	0.84	4.94	0.84
3	19.48	0.02	21.39	0.01
2	37.36	0.00	21.94	0.01

In all cases, we used the same model specification

$$\nabla \mathbf{Z}_t = \Lambda \mathbf{D}_t + \Pi_1 \nabla \mathbf{Z}_{t-1} + \cdots + \Pi_p \nabla \mathbf{Z}_{t-p} + \mathbf{a}_t$$

where the vector of deterministic variables $\mathbf{D}_t$ contains seasonal dummies.

4.3 Model estimation

The resulting VAR model is given by the following three equations

$$\begin{aligned} DTZ_t &= \pi_{1,11} DTNH_{t-1} + \cdots + \pi_{3,11} DTNH_{t-3} + \pi_{1,21} DTWH_{t-1} + \cdots \\ &\quad + \pi_{3,21} DTWH_{t-3} + \pi_{1,31} DTWB_{t-1} + \cdots + \pi_{3,31} DTWB_{t-3} \\ &\quad + \sum_{i=1}^{12} \lambda_i D_{it} + a_{DTZ,t} \end{aligned}$$

where Z stands for NH, WH and WB, for $t = 6, \ldots, N$. Besides, the seasonal dummies are centered (so that the sum of the $D_{i,t}$ values within each year is zero, with $i = 1, \ldots, 12$) that is,

$$\begin{aligned} D_{j,t} &= 11/12 \quad \text{if } t = \text{the } i\text{-th month} \\ &= -1/12 \quad \text{otherwise,} \end{aligned}$$

$t = 1, 2, \ldots, 12$.

Ordinary Least Squares was applied to each equation separately because that method produces efficient joint estimates of all the model parameters, as it is shown in Lütkepohl (2005, pp. 71-72). Therefore, parameter estimation was also carried out easily (we employed the statistical package RATS, version 5, available at http://www.estima.com). Some typical estimation results appear in Tables 3 and 4. There we see that seasonality has significant effects to explain each variable for every species. The adjusted coefficient of determination reaches values around 75% for DTNH, but it is sensibly smaller for DTWH and DTWB (around 35%). Each variable is explained by itself, among others, as shown by the F statistics used to test the null hypothesis of no significant effects on the variable to be explained.

Table 3: Estimation results of the VAR(3) model: Cattle

Variable to	Explanatory variables (F-statistics)				$\bar{R}^2$
be explained	DTNH	DTWH	DTWB	Seasonality	
DTNH	24.25***	1.67	1.74	***	0.80
DTWH	1.20	0.87	0.27	***	0.29
DTWB	1.02	0.86	0.45	**	0.29

(***) 1% Significant, (**) 5% Significant, (*) 10% Significant, (–) Not significant at 10%

Table 4: Estimation results of the VAR(3) model: Swine

Variable to	Explanatory variables (F-statistics)				$\bar{R}^2$
be explained	DTNH	DTWH	DTWB	Seasonality	
DTNH	18.59***	0.98	0.90	***	0.74
DTWH	0.24	3.55**	0.43	–	0.15
DTWB	0.20	0.88	3.82**	**	0.21

(***) 1% Significant, (**) 5% Significant, (*) 10% Significant, (–) Not significant at 10%

4.4 Model validation

Empirical adequacy of the model was checked first by means of visual inspection of residual plots as those shown in Figure 3. There we see that no outliers are present and no serious violation of the homoscedasticity assumption is evident. It should be mentioned that the extreme observations for DTWH (in 1998:07 and 2002:01) as well as those for DTWB (in 1998:07, 1999:02 and 2002:01) were adjusted during the edition procedure of the historical record.

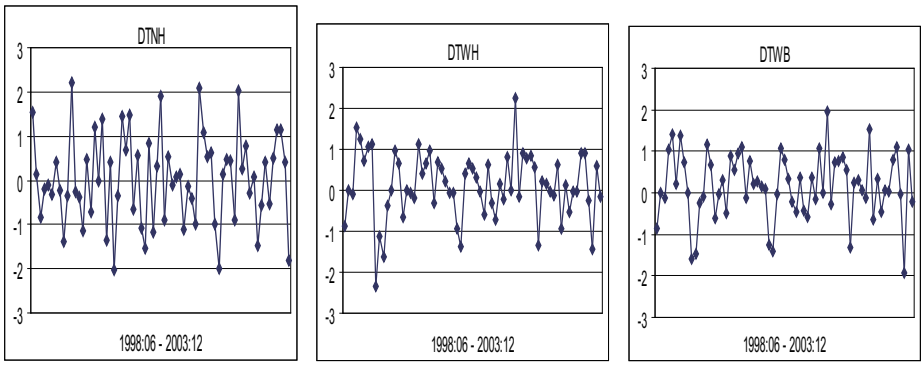

Figure 3: Standardized residuals of the VAR model (Cattle)

We also calculated the Ljung-Box statistic to check for zero autocorrelation and the Jarque-Bera statistic for normality. The corresponding values for cattle appear in Table 5 and none of those calculated statistics show evidence of inadequacy. Therefore, we concluded that the estimated model was reasonably supported by the data at hand.

Table 5: Ljung-Box and Jarque-Bera statistics for residuals: Cattle

Series	Ljung-Box (24 lags)		Jarque-Bera	
	Statistic	p-value*	Statistic	p-value
DTNH	23.1	0.34	0.69	0.71
DTWH	21.4	0.43	3.22	0.20
DTWB	18.1	0.64	1.74	0.42

*By comparing against a Chi-square distribution with 21 degrees of freedom.

5. Application of the Method

This section is devoted to show some empirical results produced by the proposed methodology. These examples are shown only for illustrative purposes of the kind of results produced by the method in practice. It is important to appreciate how the methodology is to be applied separately to data from each abattoir on a monthly basis, once the corresponding VAR model was built from the edited historical time series:

1. Locate any recorded values that are obviously incorrect, for example by checking that the average weight per animal before slaughtering lies in the

range $(K1, K2)$, where $K1$ and $K2$ are species-specific constants. Replace these with adjusted values that satisfy the corresponding edit rule.

2. Test that each month's vector of observations is compatible with the previous month's one-step-ahead forecast, by computing the compatibility statistic. If the test statistic exceeds some percentage point of the appropriate Chi-square distribution, conclude that the data were recorded incorrectly and replace them with the model-based forecasts. Otherwise, keep the data as recorded.

5.1 Edition of data

In Table 6 we show an example of an abattoir without missing data or gross inconsistencies of the observations, as compared with their one-step-ahead forecasts, during each of the months under consideration. The observed data are not significantly different from the unrestricted forecasts since the largest calculated compatibility statistic became $K_{calc} = 4.0$ which is not significant at the 10% level when compared against a Chi-square distribution with 3 degrees of freedom. Thus, by applying the edition procedure in this case we changed the observations of WB forcing them to satisfy the edit rule imposed by $K2$ and $K4$, without changing any observations of NH or WH. All the observations on WB were changed because they were too low (it is well known that some informants tend to lower the values for this variable, trying to hide the actual yield of meat).

Table 6: Application of the editing procedure in real time: Cattle

Year	Data			Compat.	Unrestricted forecasts			Restricted forecasts		
2004	NH	WH	WB	stat (sig.)	NH	WH	WB	NH	WH	WB
Apr	376	142880	61296	2.5 (.29)	395	162708	87720	376	142880	71440
May	398	159200	73232	1.9 (.39)	428	172684	94435	398	159200	79600
Jun	379	151600	69736	0.7 (.69)	384	161968	87396	379	151600	75800

Structural constants: $K1 = 250, K2 = 550, K3 = 0.5, K4 = 0.55$

5.2 Edition and imputation of missing or inconsistent data

In Table 7 we show two examples where both edition and imputation took place. In order to apply the imputation procedure we decided to fix 1% as the cut-off point for the significance level of the compatibility statistic. The first example corresponds to a gross inconsistency in abattoir A1, found in the observed data of swine in August (compatibility statistic $K_{calc} = 9.9$). All the WB data were

changed essentially by the edition rule, while the WH datum was estimated by restricted forecasting since it was considered too high by the automatic procedure (as compared with its corresponding unrestricted forecast).

Table 7: Application of the imputation method in real time: Swine

Year	Data			Compat.	Unrestricted forecasts			Restricted forecasts		
2004	NH	WH	WB	stat (sig.)	NH	WH	WB	NH	WH	WB
Jun	140	13300	6384	0.7 (.72)	150	15076	10861	140	13300	9310
Jul	160	15200	7296	3.1 (.22)	136	13495	9904	160	15200	10640
Aug	158	18960	9101	9.9 (.01)	139	13631	9524	158	18170	12705

5.3 Simulation

A small simulation study was carried out to validate the usefulness of the method to approximate the true values of the missing or inconsistent data. The experiment was intended to reproduce the most frequent situations that occur in practice, these are shown in Table 8, where the notation employed is: O = observed datum, $-$ = datum 25% lower than its actual value and $+$ = datum 25% higher than its actual value.

Table 8: Experimental design for the simulation study

Experi.	Values of			Experi.	Values of			Experi.	Values of		
run	NH	WH	WB	run	NH	WH	WB	run	NH	WH	WB
1	O	O	O	11	O	NA	–	21	–	–	–
				12	O	NA	+	22	–	–	NA
2	NA	O	O	13	O	NA	NA	23	–	NA	–
3	–	O	O					24	–	NA	O
4	+	O	O	14	NA	–	–	25	–	NA	NA
5	O	NA	O	15	NA	–	O				
6	O	–	O	16	NA	+	O	26	+	+	+
7	O	+	O	17	NA	+	+	27	+	+	NA
8	O	O	NA	18	NA	NA	–	28	+	NA	+
9	O	O	–	19	NA	NA	O	29	+	NA	O
10	O	O	+	20	NA	NA	+	30	+	NA	NA

Even without a formal statistical analysis of the experimental results, in Table 9 we can clearly see that the estimated values produced by restricted forecasting, either to replace inconsistent data or to impute missing data, are very reasonable. The largest discrepancies are: 36.8% for NH, 38.6% for WH and -52.8% for

WB, indicating that WB is the most sensitive variable. Furthermore, the values of the compatibility statistics do not lead us to declare incompatibility between observed data (NH = 492, WH = 196800, WB = 88560) and unrestricted forecasts ($NH = 463, WH = 166645, WB = 95247$). When the data are set in accordance with the experimental design incompatibility arises at the 5% significance level, in most cases.

Table 9: Simulation results for Cattle (January, 2004)

Restrictions			Compat	Restricted forecasts			Observ.-Estim. in %		
NH	WH	WB	stat. (sig.)	NH	WH	WB	NH	WH	WB
492	196800	88560	4.0 (.13)	492	196800	108240	0.0	0.0	-22.2
NA	196800	88560	4.0 (.13)	492	196800	108240	0.0	0.0	-22.2
369	196800	88560	88.3 (.00)	369	196800	104183	25.0	0.0	-17.6
615	196800	88560	27.9 (.00)	615	196800	108240	-25.0	0.0	-22.2
492	NA	88560	400.0 (.00)	492	177120	88560	0.0	10.0	0.0
492	147600	88560	8.4 (.01)	492	147600	81180	0.0	25.0	8.3
492	246000	88560	37.3 (.00)	492	246000	129761	0.0	-25.0	-46.5
492	196800	NA	4.0 (.13)	492	196800	108240	0.0	0.0	-22.2
492	196800	66420	4.0 (.13)	492	196800	108240	0.0	0.0	-22.2
492	196800	110700	4.0 (.13)	492	196800	108240	0.0	0.0	-22.2
492	NA	66420	400.0 (.00)	492	120764	66420	0.0	38.6	25.0
492	NA	110700	400.0 (.00)	492	221400	110700	0.0	-12.5	-25.0
492	NA	NA	1.2 (.27)	492	174736	96105	0.0	11.2	-8.5
NA	1 147600	66420	18.0 (.00)	369	147600	81180	25.0	25.0	8.3
NA	147600	88560	18.0 (.00)	369	147600	81180	25.0	25.0	8.3
NA	246000	88560	32.6 (.00)	615	246000	135300	-25.0	-25.0	-52.8
NA	246000	110700	32.6 (.00)	615	246000	135300	-25.0	-25.0	-52.8
NA	NA	66420	510.3 (.00)	311	132840	66420	36.8	32.5	25.0
NA	NA	88560	424.1 (.00)	414	177120	88560	15.9	10.0	0.0
NA	NA	110700	396.6 (.00)	518	221400	110700	-5.3	-12.5	-25.0
369	147600	66420	18.0 (.00)	369	147600	81180	25.0	25.0	8.3
369	147600	NA	18.0 (.00)	369	147600	81180	25.0	25.0	8.3
369	NA	66420	451.9 (.00)	369	132840	66420	25.0	32.5	25.0
369	1 NA	88560	451.9 (.00)	369	177120	88560	25.0	10.0	0.0
369	NA	NA	17.3 (.00)	369	139623	76793	25.0	29.1	13.3
615	246000	110700	32.6 (.00)	615	246000	135300	-25.0	-25.0	-52.8
615	246000	NA	32.6 (.00)	615	246000	135300	-25.0	-25.0	-52.8
615	NA	110700	398.9 (.00)	615	221400	110700	-25.0	-12.5	-25.0
615	NA	88560	398.9 (.00)	615	177120	88560	-25.0	10.0	0.0
615	NA	NA	27.1 (.00)	615	208221	114522	-25.0	-5.8	-29.3

Unrestricted forecasts: $NH = 463, WH = 166645, WB = 95247$

6. Final Considerations

The proposed procedure consists of two phases. Edition is carried out first, in order to produce valid data that can be used as input when estimating a VAR model for the multiple time series under study. Imputation is then applied to estimate missing or inconsistent data. Edition makes use of some known structural constants that define the permissible range for two constructed variables. These constants are employed to transform the data in such a way as to get valid data that can be used when building models for the imputation phase. A VAR model is used to get restricted forecasts which provide statistically efficient estimates of the missing data. This model has to be reestimated each month as new data arrive.

An added benefit of using explicit statistical models for editing and imputing data, rather than numerical algorithms, is that we can measure uncertainty of the imputed values and also make statistical inference, as that provided by the compatibility statistic. Nevertheless, we should be aware that this procedure is applied at the abattoir level, for data of each animal species, so that a measure of uncertainty for the aggregated data at the State level is something to be done in the future (this measure could be useful for performing statistical analysis of the ESGRM data). Finally, in order to generalize the use of the VAR model here employed to other Mexican States we require searching for generic models that provide statistically valid results for the abattoirs in those States.

Acknowledgements

Victor M. Guerrero gratefully acknowledges support from INEGI to carry out this work. He also thanks Asociación Mexicana de Cultura, A.C. for support granted through a professorship on Time Series Analysis and Forecasting in Econometrics at ITAM. The input provided by Brad McBride helped to improve a previous version of this paper.

Appendix. Proof of the restricted forecasting formulas

Equations (3.3) and (3.4) can be written as the following system

$$\begin{pmatrix} E(\nabla \mathbf{Z}_{N+1}|\mathbf{Z}) \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_k \\ C \end{pmatrix} \nabla \mathbf{Z}_{N+1} + \begin{pmatrix} -\mathbf{a}_{N+1} \\ \mathbf{u} \end{pmatrix}$$

with

$$E\begin{pmatrix} -\mathbf{a}_{N+1} \\ \mathbf{u} \end{pmatrix} |\mathbf{Z}\Big) = \mathbf{0}_{k+M} \text{ and } Var\begin{pmatrix} -\mathbf{a}_{N+1} \\ \mathbf{u} \end{pmatrix} |\mathbf{Z}\Big) = \begin{pmatrix} \Sigma_{\mathbf{a}} & 0 \\ 0 & \Sigma_{\mathbf{u}} \end{pmatrix}.$$

Application of Generalized Least Squares yields the following minimum MSE linear predictor

$$\begin{aligned}\hat{\mathbf{Z}}_{N+1} &= [(\mathbf{I}_k, C')\begin{pmatrix}\Sigma_{\mathbf{a}}^{-1} & 0\\ 0 & \Sigma_{\mathbf{u}}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I}_k\\ C\end{pmatrix}]^{-1}(\mathbf{I}_k, C')\begin{pmatrix}\Sigma_{\mathbf{a}}^{-1} & 0\\ 0 & \Sigma_{\mathbf{u}}^{-1}\end{pmatrix}\\ &\quad\times\begin{pmatrix}E(\mathbf{Z}_{N+1}|\mathbf{Z})\\ \mathbf{Y}\end{pmatrix}\\ &= (\Sigma_{\mathbf{a}}^{-1}+C'\Sigma_{\mathbf{u}}^{-1}C)^{-1}[\Sigma_{\mathbf{a}}^{-1}E(\mathbf{Z}_{N+1}|\mathbf{Z})+C'\Sigma_{\mathbf{u}}^{-1}\mathbf{Y}],\end{aligned}$$

with $MSE(\hat{\mathbf{Z}}_{N+1}) = (\Sigma_{\mathbf{a}}^{-1}+C'\Sigma_{\mathbf{u}}^{-1}C)^{-1}$. The Matrix Inversion Lemma (see Harvey, 1981, p. 118) leads us to

$$(\Sigma_{\mathbf{a}}^{-1}+C'\Sigma_{\mathbf{u}}^{-1}C)^{-1} = \Sigma_{\mathbf{a}} - \Sigma_{\mathbf{a}}C'(\Sigma_{\mathbf{u}}C\Sigma_{\mathbf{a}}C')^{-1}C\Sigma_{\mathbf{a}} = (\mathbf{I}_k - A_{\mathbf{u}}C)\Sigma_{\mathbf{a}}$$

and

$$(\Sigma_{\mathbf{a}}^{-1}+C'\Sigma_{\mathbf{u}}^{-1}C)^{-1}C'\Sigma_{\mathbf{u}}^{-1} = \Sigma_{\mathbf{a}}C'[\Sigma_{\mathbf{u}}^{-1} - \Sigma_{\mathbf{u}}^{-1}C(\Sigma_{\mathbf{a}}^{-1}C'\Sigma_{\mathbf{u}}^{-1}C)^{-1}] = A_{\mathbf{u}}.$$

Hence,

$$\begin{aligned}\hat{\mathbf{Z}}_{N+1} &= (\mathbf{I}_k - A_{\mathbf{u}}C)E(\mathbf{Z}_{N+1}|\mathbf{Z}) + A_{\mathbf{u}}\mathbf{Y}\\ &= E(\mathbf{Z}_{N+1}|\mathbf{Z}) + A_{\mathbf{u}}[\mathbf{Y} - CE(\mathbf{Z}_{N+1}|\mathbf{Z})]\end{aligned}$$

with $MSE(\hat{\mathbf{Z}}_{N+1}) = (\mathbf{I}_k - A_{\mathbf{u}}C)\Sigma_{\mathbf{a}}$.

References

Anderson, O. D. (1976). *Time Series Analysis and Forecasting. The Box-Jenkins approach*. Butterworth & Co.

Caporello, G. and Maravall, A. (2002). A tool for quality control of time series data. Program TERROR. Available at the website of Banco de España, http://www.bde.es/servicio/software/econom.htm

Cattle dealers (2005) http:// www.mexicoganadero.com/limousin

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction representation, estimation and testing. *Econometrica* **55**, 251-276.

Gómez, V., Maravall A. and Peña, D. (1999). Missing Observations in ARIMA Models: Skipping Approach versus Additive Outlier Approach. *Journal of Econometrics* **88**, 341-364.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting* **12**, 37-48.

Guerrero, V. M. (1994). Restricted forecasts of missing observations in univariate time series. *Estadistica* **46**, 1-23.

Harvey, A. (1981) *Time Series Models*. Philip Allan Publishers.

INEGI (2004). Publicación anual de la Estadística de Sacrificio de Ganado en Rastros Municipales por Entidad Federativa 1998-2003. http://www.inegi.gob.mx/

Kohn, R. and Ansley, C. F. (1986). Estimation, Prediction and Interpolation for ARIMA Models with Missing Data. *Journal of the American Statistical Association* **81**, 751-761.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley.

Luceño, A. (1997) Estimation of missing values in possibly partially nonstationary vector time series. *Biometrika* **84**, 495-499.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Pankratz, A. (1989). Time series forecasts and extra-model information. *Journal of Forecasting* **8**, 75-83.

Pfefferman, D. and Nathan, G. (2002). Imputation for Wave Nonresponse: Existing Methods and a Time Series Approach. In *Survey Nonresponse* (Edited by Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. J. A., 417-429). John Wiley.

Ruíz F. A., *et al.* (2001). Diagnóstico del sector pecuario. In: México Rural: Políticas para su reconstrucción. Universidad Autónoma Chapingo, Chapingo, México.

SAGARPA (2005). www.sagarpa.gob.mx/ganaderito

Sánchez, I. and Peña, D. (2001). Properties of predictors in overdifferenced nearly nonstationary autoregression. *Journal of Time Series Analysis* **22**, 45-66.

Sargan, J. D. and Drettakis, E. G. (1974). Missing data in an autoregressive model. *International Economic Review* **15**, 39-58.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.

Veterinary (2005). www.veterin.unam.mx

Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review* **71**, 581-592.

Received December 3, 2008; accepted May 6, 2009.

Victor M. Guerrero
Department of Statistics
Instituto Tecnológico Autónomo de México (ITAM)
Río Hondo 1, Col. Progreso-Tizapán, México 01080, D.F., MEXICO
guerrero@itam.mx

Blanca I. Gaspar
Banco de México
Calzada Legaria
No. 691, Col Irrigación, México 11500, D.F., MEXICO
bgaspar@banxico.org.mx