ARTICLE IN PRESS
International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Testing time series data compatibility for benchmarking

Benoît Quenneville *, Christian Gagné

Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6

ARTICLE INFO

Keywords:
Autoregressive process
Chi-square distribution
Likelihood ratio test
Score test
Wald test

ABSTRACT

Compatibility testing determines whether two series, say a sub-annual and an annual series, both of which are subject to sampling errors, can be considered suitable for benchmarking. We derive statistical tests and discuss the issues with their implementation. The results are illustrated using the artificial series from Denton (1971) and two empirical examples. A practical way of implementing the tests is also presented.

1. Introduction

Benchmarking is done when two time series measuring the same variable at different frequencies are both subject to measurement errors. Benchmarking combines information from the two series to get a better estimate of the target variable. For example, monthly surveys of business revenue should add up to an equivalent annual survey. However, they will not typically add up, due to measurement errors. Benchmarking would produce a monthly series which will add up to the annual figures. For further details, see Cholette and Dagum (1994) for benchmarking time series using autocorrelated sampling errors, and Dagum and Cholette (2006) for a recent book on benchmarking and temporal distribution methods.

Before benchmarking two time series, it might be advisable to test whether these series should be benchmarked in the first place. Two series are compatible if they are jointly likely to have been observed, considering the distributional assumptions made on the error process. Guerrero (1990) proposes a test for compatibility in the context of temporal disaggregation, which can be cast as a form of benchmarking. The test determines whether the observed aggregates are likely to have been observed under a given ARIMA model, where, in practice, the model is determined from observed sub-annual data. In turn, this series is used to temporally disaggregate the original series. The test is suitable when choosing one sub-annual series among many candidates, but this is not what we are considering in this paper. We are also looking for methods which can be implemented in a computer package, such as Statistics Canada's in-house SAS® Proc Benchmarking (Latendresse, Djona, & Fortier, 2007), with as few inputs from the user as possible. Furthermore, our primary concern is the presence of measurement errors which cause discrepancies and therefore generate the need for benchmarking. The tests proposed here are designed for quality testing when we already know which series are involved. These tests are based on the observed discrepancies between the two series and determine whether the observed discrepancies are as expected. If the answer is no, then benchmarking should not be applied. A list of possible conceptual, operational and methodological differences to investigate is provided by Brisebois and Yung (2007).

Section 2 of the paper states the model and hypotheses to be tested. Section 3 defines some statistics based on the observed discrepancies. Section 4 derives the test statistics assuming a full knowledge of the covariance matrices of the errors. It includes a discussion of benchmarking using signal extraction methods, which requires additional knowledge of the data generating process. In practice, the covariance matrices of the errors may not be fully known or available, in which case some additional assumptions are required. We discuss the issues facing the implementation of the tests when only the sub-annual series and the

* Corresponding author.
E-mail addresses: Benoit.Quenneville@statcan.gc.ca (B. Quenneville), Christian.Gagne@statcan.gc.ca (C. Gagné).

0169-2070/$ – see front matter Crown Copyright © 2011 Published by Elsevier B.V. on behalf of International Institute of Forecasters. All rights reserved.
doi:10.1016/j.ijforecast.2011.10.001

benchmarks are provided. Section 5 presents two simpler statistics which do not depend on the covariance structure, although their distributions do. The two simpler statistics can be used as indices of quality. Section 6 provides two real examples in which one is believed to be likely compatible and the is other not. These two examples illustrate how the compatibility tests can be used in practice. Section 7 presents a summary of the issues discussed and concludes the paper.

2. Model and hypotheses

We observe two time series, $s = (s_1, \ldots, s_T)$ and $a = (a_1, \ldots, a_B)$. For example, s might represent a monthly or quarterly series, while a represents an annual series of benchmarks. We assume that benchmarks occur at regular intervals and that the period which each of them covers is of length m ($m = 4$ for quarterly data and $m = 12$ for monthly data). Only the observations within the span of the benchmarks are needed for compatibility testing. Both s and a contain possible measurement errors. The mathematical model is

$$s = \eta + e, \quad e \sim N\left[0, V(e)\right],$$
$$a = \alpha + \epsilon, \quad \epsilon \sim N\left[0, V(\epsilon)\right], \tag{1}$$

with $\mathrm{Cov}(e, \epsilon) = 0$.

The two vectors of interest are $\eta = (\eta_1, \ldots, \eta_T)$ and $\alpha = (\alpha_1, \ldots, \alpha_B)$, which are related by the benchmarking relationship $J\eta = \alpha$, where $J = I(B) \otimes 1'_m$; 1_m is a vector of ones of length m; $I(B)$ is an identity matrix of order B; and $\otimes$ is the Kronecker product. For example, with 2 years of quarterly data: $\eta = (\eta_1, \ldots, \eta_8)$, $\alpha = (\alpha_1, \alpha_2)$, and

$$J = I(2) \otimes 1'_4 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The benchmarking relationship $J\eta = \alpha$ means $\eta_1 + \eta_2 + \eta_3 + \eta_4 = \alpha_1$ and $\eta_5 + \eta_6 + \eta_7 + \eta_8 = \alpha_2$.

Benchmarking is about estimating η and α under the hypothesis $J\eta - \alpha = 0$; however, before benchmarking the series, it may be appropriate to test whether this should be done. The compatibility test is about the null hypothesis

$$H_0 : J\eta - \alpha = 0 \tag{2}$$

against the alternative

$$H_1 : J\eta - \alpha \neq 0. \tag{3}$$

There may be clearly understood sources of measurement errors, such as stable under-coverage of the sub-annual survey, in which case a bias adjustment may be useful before proceeding with the compatibility test H_0 against H_1. Therefore, we also need to consider the alternative hypothesis of significant bias

$$H_2 : J\eta - \alpha = k1_B \neq 0, \tag{4}$$

where 1_B is a vector of ones of length B. Obviously, alternative hypotheses such as a linear or spline trend for the bias are also possible; however, this paper will restrict its discussion to the case of a constant bias, as described by the hypothesis H_2.

3. Observed discrepancies and proportional discrepancies

The vector of the observed discrepancies between the sub-annual and annual series is

$$d = Js - a \tag{5}$$
$$= J\eta - \alpha + Je - \epsilon. \tag{6}$$

The sample mean of the discrepancies, denoted by $\bar{d}$, is a consistent estimator of the expected discrepancy $E\,(J\eta - \alpha)$. The optimal estimator which depends on $V(e)$ and $V(\epsilon)$ could be derived (see, for example, Cholette & Dagum, 1994, for the case where $V(\epsilon) = 0$ and $V(e)$ is a covariance matrix of a sample of size T from an ARMA process), but $\bar{d}$ requires fewer assumptions.

The vector of proportional discrepancies is

$$\tilde{d} = D(a^{-1})d, \tag{7}$$

where $D(a^{-1})$ is a diagonal matrix with vector a^{-1} on the main diagonal and 0 elsewhere. The sample mean of the proportional discrepancies, denoted by $\bar{\tilde{d}}$, estimates the average proportional discrepancy $E\left[D(\alpha^{-1})\,(J\eta - \alpha)\right]$, where $D(\alpha^{-1})$ is the diagonal matrix with vector α^{-1} on the main diagonal and 0 elsewhere.

4. Test statistics

Assume initially that the sub-annual series is free of coverage bias. The vector d of the observed discrepancies between the sub-annual and annual series is centered on a known distribution under H_0:

$$d \sim N\left[0, JV(e)J' + V(\epsilon)\right], \tag{8}$$

which immediately leads to the following test:

$$T_0 = d'\left[JV(e)J' + V(\epsilon)\right]^{-1} d \sim \chi^2_B. \tag{9}$$

T_0 can only be applied if the covariance matrices $V(e)$ and $V(\epsilon)$ are known. T_0 asks whether the observed discrepancies are of the expected scale. If they are not, then benchmarking should not be done.

Most of this paper is about how to implement T_0 in a computer package, such as Proc Benchmarking, when only the series s and the benchmarks a are provided. Additional assumptions will be required, and we investigate a few possible implementation issues.

4.1. *Estimation and benchmarking using model-based methods*

It may be appropriate at this point to discuss benchmarking estimation methods and the use of model-based methods. As in the work of Cholette and Dagum (1994), the η_ts are constants and the measurement error model (1) can be viewed as a regression model. The estimate of η obtained by generalized least squares is

$$\hat{\eta} = a + V(e)J'\left[JV(e)J' + V(\epsilon)\right]^{-1}(a - Js)\,. \tag{10}$$

The prediction error covariance matrix is

$$\mathrm{V}(\hat{\eta}) = \mathrm{E}\left(\hat{\eta} - \eta\right)\left(\hat{\eta} - \eta\right)'$$
$$= V(e) - V(e)J'\left[JV(e)J' + V(\epsilon)\right]^{-1} JV(e), \tag{11}$$

with the property $\mathrm{V}(\hat{\eta}) \leq V(e)$.

If η_t is assumed to be a stochastic process, then benchmarking using signal extraction methods such as those of Chen, Cholette, and Dagum (1997), Durbin and Quenneville (1997) and Hillmer and Trabelsi (1987), may provide better predictors. For these methods, additional