Données manquantes et prévisions : méthodes à imputation variable

Antonio ANSELMI [1], Paola Maddalena CHIODINI [2] et Flavio VERRECCHIA [3]

En littérature, dans le contexte des séries chronologiques, pour l'imputation des données manquantes, on se réfère à des statistiques appliquées à tous les termes de la série analysée spécifique (e.g. moyenne arithmétique), en obtenant une constante d'imputation généralement convenable pour quelques séries spécifiques. Si les séries sont en nombre n (n ⊚ ∞), il est impossible de trouver une fonction unique pour les n constantes d'imputation des données manquantes.

En abandonnant les « anciennes méthodes » (Schafer, Graham 2002), l'objectif devient celui d'évaluer les nouvelles propositions de méthodes d'imputation des données manquantes pour les bases de données hiérarchiques achevées pour la mise en pratique des modèles de séries chronologiques (Rubin 1996 ; Chiodini, Verrecchia 2008 ; Anselmi, Chiodini, Verrecchia 2008). Les méthodes proposées dans cet ouvrage sont inspirées de la théorie des échantillons où il faut trouver la meilleure solution au problème des données manquantes. En particulier, nous allons tout d'abord examiner les méthodes – ou des séquences de méthodes - d'imputation qui aident à reconstruire les données manquantes en tenant compte de la variabilité naturelle du phénomène étudié (Rubin 1987, 1996 ; Hergoz et Rubin, 1983 ; Rubin et Shenker, 1986). Une première question à développer dans ce domaine concerne la vérification de la validité de l'hypothèse de normalité sur les distributions transformées (e.g. logarithme), et, le cas échéant, en utilisant des variables mélangées. Un deuxième point d'intérêt concerne l'ancrage macro-régional pour l'estimation des paramètres des distributions qui, dans certains cas, peut conduire à des distorsions dans l'imputation des données (voir, par exemple, BiCRA-PSG in : Eusepi, Cepparulo, Verrecchia 2007).

Enfin, on présentera des applications produites avec SAS Forecast Server pour les différentes méthodes d'imputation, qui permettront de comparer les modèles (automatiquement sélectionnés) en partant de la base des données traitées avec différents types d'imputation.

Keywords : Données Manquantes, Imputation Stochastique, Séries Chronologiques, Prévision Hiérarchique.

[1] SAS Institute

[2] Université de Milano-Bicocca

[3] Président, EseC, Economic Statistics e-Center, web: www.economicstatistics.eu - flavio.verrecchia@gmail.com

Bibliographie :

AA.VV. (2004), Handling missing data: applications to environmental analysis, WIT Press, UK

Anselmi A., Chiodini P.M., Verrecchia F. (2008) ESeC-Rubin Missing Value Interpretation for a Regional Bottom-Up Hierarchical Forecasting, ESeC Working Paper [ESeC_WP002_V20080926], Handle [RePEc:est:wpaper:002], Online [http://www.economicstatistics.eu/wp]

Bailar B.A. and Bailar J.C. III (1978) Comparison of two procedures for imputing missing survey values. In Proceedings of the Survey Research Methods Section, American Statistical Association, Washington, D.C., 462-467.

Buck S.F. (1960) "A method of estimation of missing values in multivariate data suitable for use with an electronic computer", Journal of the Royal Statistical Society B, 22, 302-306

Chiodini P.M., Verrecchia F. (2008) "Imputazione dei dati mancanti in basi dati economico-sociali per il forecasting regionale: il metodo ESeC-Rubin". In SAS Business Analytics Gallery 2008. Roma. Online [http://www.economicstatistics.eu/poster].

Eusepi G., Cepparulo A., Verrecchia F. (2007) "Bilevel Comparative Regional Analysis - Performances in Structural Grid.", ESeC. Working paper ESeC_WP001.

Herzog T.N. and Rubin D.B. (1983) Using multiple imputation to handle nonresponse in sample surveys. In *Incomplete Data in Sample Surveys*, Vol. 2, W.G. Madow, I. Olkin and D.B. Rubin, Eds., Academic Press, New York, 210-245.

Little R.J.A., Rubin D.B. (2002), Statistical analysis with missing data, John Wiley & Sons, New York

Martini M. (2001) "Numeri indice per il confronto nel tempo e nello spazio", CUSL, Milano.

Rubin D.B. (1987) Multiple imputation for Nonresponse in Surveys. John Wiley & Sons, New York.

Rubin D.B. (1996) Multiple imputation after 18+ year. *J. Am. Stat. Assoc.*, 91, 507-510.

Rubin D.B. and Schenker N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Stat. Assoc.*, 81, 366-374.

SAS for Forecasting Time Series, Second Edition by John Brocklebank and David Dickey Copyright(c) 2003 by SAS Institute Inc., Cary, NC, USA ISBN 1-59047-182-2

SAS Institute Inc. (2006) SAS Forecast Studio 1.4: User's Guide. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2007) SAS Forecast Server 1.4: Administrator's Guide. Cary, NC: SAS Institute Inc.

SAS Forecast Server. Cary, NC: SAS Institute Inc. Online:

[http://www.sas.com/technologies/analytics/forecasting/forecastserver/factsheet.pdf]

Schafer J.L. and Graham J.W. (2002), ``Missing Data: Our View of the State of the Art'', Psychological Methods, 7 , 147-177

Verrecchia F. (2008) The Generalised Index Numbers, Journal of ESeC Short Papers, Italy, 1 (1): 9-12.

Verrecchia F. (2008) "Previsione e selezione automatica dei modelli per serie storiche regionali: metodo bi-fase a conciliazione esterna". In SAS Business Analytics Gallery 2008. Roma. Online [http://www.economicstatistics.eu/poster].

Verrecchia F., Chiodini P.M., Coin D., Facchinetti S., Nai Ruscone M. (2008) Bayesian Approach for Nonresponse, in: SSBS08 (Sample Surveys and Bayesian Statistics) - Satellite conference to the RSS 2008 conference, Southampton, UK (26-29 August 2008). Online [http://www.s3ri.soton.ac.uk/ssbs08/programme.php].