DONNEES MANQUANTES ET PREVISIONS: METHODES A IMPUTATION VARIABLE[1]

Antonio ANSELMI (*), Paola Maddalena CHIODINI (**),Flavio VERRECCHIA(***)

(*) SAS Institute, Customer Support
(**) University of Milano-Bicocca, Département de Statistique
(***) ESeC – Economic Statistics e-Center

Introduction

En littérature, dans un contexte de séries temporelles, pour l'imputation des données manquantes on se réfère à statistiques appliquées à tous les termes de la spécifique série analysée (e.g. moyenne arithmétique), en obtenant une constante d'imputation généralement convenable pour quelque spécifiques séries. Si les séries sont n ($n \rightarrow \infty$), il est impossible trouver une fonction unique pour les n costante d'imputation des données manquantes.
En abandonnant les «anciennes méthodes» (Schafer, Graham 2002), l'objectif devien celui d'évaluer les nouvelles propositions de méthodes d'imputation des données manquantes pour les bases de données hiérarchique achevé pour la mise en pratique des modèles pour séries temporelles (Rubin 1996; Chiodini, Verrecchia 2008; Anselmi, Chiodini, Verrecchia 2008). Les méthodes proposées dans cet ouvrage sont inspirés à 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 aide à reconstruire les données manquantes en tenant compte de la naturelle variabilité du phénomène à l'étude (Rubin 1987, 1996; Hergoz e Rubin, 1983; Rubin e Shenker, 1986). Une première question à se développer dans ce domaine concerne la vèrification de la validité de l'hypothèse de normalité sur les distributions transformés (e.g. logarithme. Un deuxième point d'intérêt concerne l'ancrage macro-régional pour l'estimation des paramètres des distributions que, 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 different méthodes d'imputation, qui permetront de comparer les modèles (automatiquement sélectionnés) en partant de la base de données traitée avec différents typologies d'imputation.

1. Méthodes

1.1. Echantillonnage et traitement des données manquantes

Dans le cadre des enquêtes par sondage (recencement ou non), de données longitudinales ainsi que d'autres applications dans le domaine des applications statistiques sur des données réelles, les données manquantes représentent un problème de grande importance. Les raisons pour lesquelles il y a des données manquantes sont plusieurs, d'intérêt pour l'objectif du papier est leur classement en vue de définir la méthode la plus appropriée pour la reconstruction des données manquantes.

Little et Rubin (2002) ont fourni trois définitions de donnée manquante qui sont devenues de référence en littérature:

[1] Le travail est de responsabilité de tous les auteurs. En particulier, les paragraphes 1.4 et 2.2.1 ont été traités par Anselmi, le paragraphe 1.1 a été traités par Chiodini, les paragraphes 1.2, 2.1, 2.2.2 ont eté traités par Verrecchia. Tandis que les paragraphes 1.3, 3 et 4 ont été traités par Chiodini et Verrecchia.
Le travail a le support de SAS Institute. Correspondance concernant ce papier doit être adressée à F. Verrecchia, ESeC (via Matteotti, 15/A – 20090 Assago (MI), Italie, Email: flavio.verrecchia@gmail.com, Web: www.economicstatistics.eu) ou à P.M. Chiodini (paola.chiodini@unimib.it).

- donnée manquantes par hasard (MAR *missing at random*);
- donnée manquantes complètement par hasard (MCAR *missing completely at random*);
- donnée manquantes non ignorable.

Dans le premier cas, nous supposons que la probabilité que la donnée soit manquante dépend des données observées et non pas des données manquantes. Dans ce cas, les valeurs manquantes peuvent être reconstruite par l'exploitation du lien existant avec les données observées.

Dans le second cas, nous supposons que l'élément manquant est indépendant des autres données observées, pour mieux dire la donnée manquante est indépendante des valeurs de toutes les variables observé ou non. En fait, on peut supposer que l'absence d'une donnée est imputable à des causes purement accidentelle.

Le dernier cas se réfère à la situation plus complexe, parce que on suppose que les données manquantes ne dépendent pas de facteurs accidentels, mais ils n'est pas possible les reconstruire par des relations fonctionnelles qui peuvent les lier à d'autres variables de l'ensemble de données.

On comprend immédiatement que le traitement des données manquantes est un thème délicat qui se prête à solutions différentes. En littérature en trouve plusieurs méthodes d'imputation, dont certains sont idéalement regroupés parmi les "vieilles méthodes" (Schafer, Graham 2002). Parmi les différentes solutions proposées en littérature on a l'élimination de l'information partielle ou fragmentaire. Cette méthode peut être problématique lorsque les données manquantes sont nombreuses a cause de la réduction de la base de données (i.e. échantillon trop petit).

Par conséquent, il est certainement plus efficace la reconstruction des données manquantes, mais pour procéder il faut comprendre avant tout (si possible) quel est le mécanisme qui a conduit à la non registration d'une partie des données.

On va introduire, ci-dessous, les principales méthodes d'imputation (AAVV 2004) généralement utilisés dans les recencement. En premier, on parlerà des méthodes d'imputation simple.

Méthode de substitution

L'élément manquant est remplacé par l'utilisation d'une unité initialement non présente dans l'échantillon, mais qui est similaire à celle manquante (par exemple non-répondants dans les enquêtes sur la popolation). Il est clair que les données rassemblé avec cette méthode il faut les traiter comme des données imputées.

Méthode de la moyenne conditionnelle

On remplace la donnée manquante par la moyenne des informations observée (la moyenne de l'échantillon) ou par la mode (données catégoriques). On obtien une constante pour toutes les données manquantes. La technique est simple mais pas totalement satisfaisant dans le cas de MCAR, les paramètres (par exemple, variance, corrélation, etc.) sont touchés par distorsion.

Méthode de la moyenne conditionnelle (ou de régression)

Introduite par Buck (1960) et reprise par Little et al. (2002) la méthode consiste dans l'estimation de la moyenne et de la matrice de covariance sur la base des informations complètes de l'échantillon. Ensuite, ces estimations sont utilisées, afin de déterminer la ligne des moindres carrés (modèle de régression multiple), qui permet d'estimer les valeurs manquantes. Cette approche, aussi, présente un problème de sous-estimation de la variance comme dans le cas de la moyenne non conditionnelle, mais non si remarquable.

Méthode de régression stochastique

Une proposition alternative à la méthode juste décrite est celle de la régression stochastique. L'idée est simple, il est proposé de remplacer les données manquantes en utilisant toujours les informations obtenues par le « bias » de la régression, mais à la variable de réponse est ajoutée une erreur de variance égale à l'estimation de la variance résiduelle. Cela permettra d'introduire un élément de

hasard à la valeur estimée par la régression. Dans le cas habituel de données distribuées en fonction de la loi normale, les résidus se distribuent en fonction de la même loi avec moyenne nul et variance correspondant au résiduelle de régression.

Méthode de l'imputation hot-deck

C'est une méthode largement utilisée (voir Bailar et al. 1978). On prend l'information disponible pour estimer une valeur manquante. Cette fois, cependant, on utilise une donnée observée semblable à celle manquante. Pour chaque élément manquant on identifi un cas observé similaire. Pour cet objectif, les variables sont regroupées en catégories (classes d'imputation).

Si on ne parviens pas à établir la correspondance on réduit l'ensemble des variables réduisant aussi le nombre de catégories. Sinon, on peut définir une distance à partir des variables observées à la fois pour les répondants et pour les non-répondants. La données utilisées à la place de l'élément manquant est celle qui a une distance mimal avec l'élément. En pratique on a:

- Hot-Deck avec ajustement de cellules: les cellules d'ajustement sont construites par les variables catégorielles. Les données manquantes d'une cellule peuvent être remplacées par les données de la même cellule.
- Nearest Neighbour Hot-Deck: on défini une mesure pour la distance entre les unités, on choisi entre les données existantes la plus proche de l'absente et donc on la remplace.

Il semble évident que le principal problème de cette méthode est la définition des paramètres qui, dans certains contextes peut être « naturelle » (par exemple, dans le cas d'observations spatial on peut utiliser la métrique euclidienne), dans d'autre cas non (e.g. séries temporelles).

Nearest-Neighbour interpolation

Cette méthode est normalement utilisée lorsque les données sont spatial. La méthode consiste dans l'organisation des données collectées dans une matrice et de remplacer les données manquantes avec la moyenne des valeurs plus proches. Une extension de cette approche c'est donner des poids: plus petite est la distance entre la donnée observées et celle mancante, majeurs est le poids qui aura en moyenne. De cette façon, les données les plus éloignés de l'élément manquant peuvent être utilisé, mais avec un poids faible liée à la distance. Même dans ce cas il se pose la question de la définition d'une métrique.

Méthode d'imputation Cold-Deck

La donnée manquante est remplacé par une valeur obtenue à partir d'une source de données externe (par exemple, une enquête précédente).

Imputation multiple

La méthode d'imputation multiple arrive a resoudre la plus grande partie des problèmes des méthodes d'imputation précédemment introduit. En fait, l'imputation avec une valeur constante ne peut jamais reproduire la variabilité typique d'un phénomène. Cette tendance conduit à une sous-estimation de la variabilité. Avec la méthode d'imputation multiple, cependant, on cherche à remédier à ce problème avec la reconstruction des données manquantes obtenu comme synthèse de différentes valeurs échantillonnées à partir d'une distribution normale caractérisée par des indices de position et de dispersion typique du phénomène observé (la littérature suggère l'utilisation de cinq valeurs).

1.2. Les nombre indices

Dans le domaine de la statistique-économique la comparaison temporelle des agrégats socio-économiques est de grand intérêt, parce que les statistiques des unités élémentaires qui composent ces agrégats évoluent dans le temps.

D'après les résultats de la théorie axiomatique (voir Martini 2001), de la définition de indice généralisé gIN (voir Verrecchia 2007, 2008), et de la définition générale de nombre indice élémentaire [1], on défini les nombres indice élémentaire a base fixe [2] et mobile [3]. Etant donner que le principal objectif de l'article est l'imputation des données manquantes, la définition de taux moyen de variation [4] sera utile pour la reconstruction des missing.

Définition 1. nombre indice élémentaire

Soit (Ω, *F*) un espace mesurable, Ω un ensemble appelé espace échantillon, *F* une σ- algèbre sur Ω, $\wp$ une probabilité sur (Ω, *F*). Soit le vecteur $\mathbf{E}_{\cdot} \in \Re^{+Z}$ le vecteur des éléments de base pour T+1 situations (avec t = 0, 1, 2, …, T) et $\mathbf{I}^{\mathbf{c}}_{[\mathbf{b}\cap\mathbf{t}]}$ la fonction indicatrice des co-présents dans deux situations b et t (où t désigne la situation rapporté et b la situation base). Soit **e.** et $\mathbf{i}^{\mathbf{c}}_{[\mathbf{b}\cap\mathbf{t}]}$ les déterminations des variables aleatories **E.** e $\mathbf{I}^{\mathbf{c}}_{[\mathbf{b}\cap\mathbf{t}]}$. Soit ($\mathbf{E}_{\mathbf{b}}$, $\mathbf{E}_{\mathbf{t}}$, $\mathbf{I}^{\mathbf{c}}_{[\mathbf{b}\cap\mathbf{t}]}$) un vecteur aléatoire défini sur (Ω, *F*, $\wp$) à valeurs en $\Re^{+2Z} \times \{0,1\}$ (avec $\Re^{+2Z} = (((\Re^{+})^{Z})^{2})$).

Une application mesurable non négative définie sur *F* exprimée par le rapport

$$f(\mathbf{E}_{\mathbf{b}},\mathbf{E}_{\mathbf{t}}, \mathbf{I}^{\mathbf{c}}_{[\mathbf{b}\cap\mathbf{t}]}) : F \rightarrow [0;\infty) \qquad [1]$$

qui transforme les variables aleatorires des situations élémentaires en nombre $\in \Re^{+Z}$ est dite nombre indice élémentaire sur (Ω, *F*).

Définition 2. nombre indice élémentaire à base fixe

Avec b = 0 la [1] est un nombre indice élémentaire à base fixe.

Définition 3. nombre indice élémentaire à base mobile

Avec b = t-1le [1] est un nombre indice élémentaire à base mobile

Les indices élémentaires sont de nombre pures indépendant de la mesure et ont certaines propriétés d'intérêt. Soit Z le numéro de unité élémentaires observées en T+1 situations. On assume b = t-1 et soit ${}_{t-1}f_{t,[z]} = f(\mathbf{E}_{\mathbf{t-1,z}},\mathbf{E}_{\mathbf{t,z}}, \mathbf{I}^{\mathbf{c}}_{[\mathbf{t-1}\cap\mathbf{t}],\mathbf{z}})$ le z-ième indice de la z-ième unité élémentaire observée (avec z = 1, 2, ..., Z) on peut vérifier que:

i) ${}_{t-1}f_{t,[z]} \cdot {}_{t}f_{t-1,[z]} = 1$ ou ${}_{t-1}f_{t,[z]} = 1 / {}_{t}f_{t-1,[z]}$;

ii) ${}_{t-1}f_{t,[z]} \cdot {}_{t}f_{t+n,[z]} = {}_{t-1}f_{t+n,[z]}$;

iii) ${}_{t-1}f_{t,[z=r\cdot s]} = {}_{t-1}f_{t,[z=r]} \cdot {}_{t-1}f_{t,[z=s]}$ ou ${}_{t-1}f_{t,[z=r/s]} = {}_{t-1}f_{t,[z=r]} / {}_{t-1}f_{t,[z=s]}$.

Il est clair que la ii. permet de enchaîner une série d'indices a basés mobile et de passer à un indice à base fixe :

${}_{0}f_{1,[z]} \cdot {}_{1}f_{2,[z]} \cdot \ldots \cdot {}_{t-1}f_{t,[z]} \cdot {}_{t}f_{t+1,[z]} = {}_{0}f_{t+1,[z]}$

Il convient de observer que, en référence au signifié économique[2] des indices, pour mesurer uniquemenent l'intensité des facteurs liés à la dynamique temporelle, il est généralement nécessaire

[2] Il n'est pas object du travail la vérification du respect des requis. On utiliserait, donc, les series cronologiques de Eurostat, même si on ne peut pas exclure la nécessité d'une reconstruction des series ici utilisées avec une finalité examplificatrice.

de répondre à certains requis. Si ces requis ne sont pas satisfaits (par exemple, modification des critères et des méthodes de recueil des données), on peut reconstruire les séries en utilisant la chaîne d'indices ou rendre techniquement comparable les termes de la série à l'étude.

Définition 4. Indice *CPGR (Compound Periodical Growth Rate)*

La variation moyenne de période entre t et b, définie par le suivant expression géométrique ${}_b f_{t,[z]}^{(t-b)^{-1}}$, est appelé taux moyen de variation (et si annuel CAGR).

Soit les ***e.** les éléments mancants construits par la déf. [4]. On peut vérifier que:

i) $\mathbf{*e}_{b+1,[z]} = \mathbf{e}_{b[z]} \cdot {}_b f_{t,[z]}^{(t-b)^{-1}}$

ou:

$\mathbf{*e}_{t-1,[z]} \cdot {}_b f_{t,[z]}^{(t-b)^{-1}} = \mathbf{e}_{t,[z]}$ avec

$\mathbf{*e}_{t-1,[z]} = \mathbf{*e}_{t-2,[z]} \cdot {}_b f_{t,[z]}^{(t-b)^{-1}}$ avec

... avec

$\mathbf{*e}_{b+1,[z]} = \mathbf{e}_{b[z]} \cdot {}_b f_{t,[z]}^{(t-b)^{-1}}$

et que, si $\mathbf{e}_{b,[z]} = \mathbf{e}_{t,[z]}$ allor

ii) ${}_b f_{t,[z]}^{(t-b)^{-1}} = 1$ pour $\forall$ t, b

iii) $\mathbf{e}_{t,[z]} = \mathbf{*e}_{t-1,[z]} = \mathbf{*e}_{t-2,[z]} = \ldots = \mathbf{*e}_{b+1,[z]} = \mathbf{e}_{b,[z]}$.

À partir de iii. On comprend que dans le cas de séries stationnaires en moyenne l'application du CPGR est équivalent à l'application de la moyenne.

1.3. L'approche ESeC-Rubin

ESeC-Rubin est liée au contexte spatio-temporel dans lequel on peut supposer liens dimensionals, dans le temps et dans l'espace, entre les données (Fig. 1). Dans le cas de differentes séries hétérogènes à analyser en même temps, si le remplacement des données manquantes à partir d'une seule fonction (par exemple, moyenne temporel) est souvent insuffisante, il est vrait aussi qu'il serait pratiquement impossible de choisir la meilleure fonction pour toutes les differentes séries temporelles à analyser. Non seulement si sur les séries reconstruites on souhaite appliquer les modèles de prévision, l'attribution d'une fonction constante pour chaque élément manquant pourrait conduire à des problèmes (par exemple intervalles de confiance).

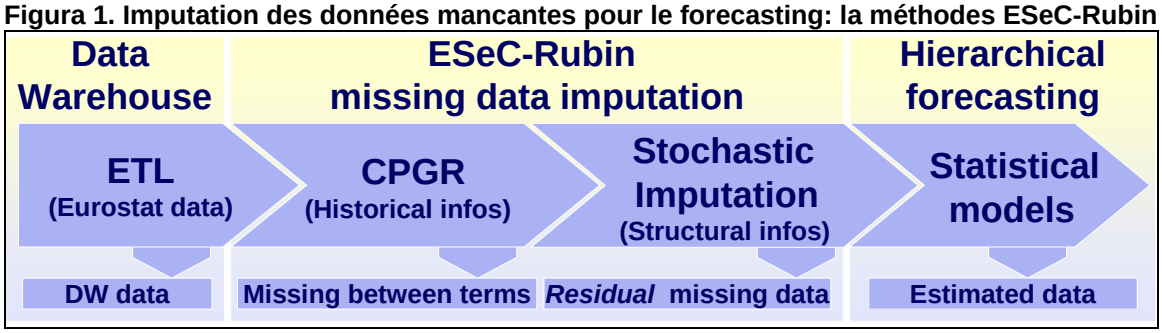
ESeC-Rubin est basée sur l'utilisation séquentielle des méthodes d'imputation des données en utilisant les relations temporelles et spatiales et du mécanisme de la générations des données. La méthode qui est proposée dans ce travail est structuré comme suit:

- **information temporal**. On utilise le temps comme liens avec l'application du CPGR pour les données mancantes entre deux observations disponible.
- **numéros purs**. Les indices a base mobile sont calculés sur la base des déterminations de chaque series (la méthode est donc applicable indépendamment du facteur d'échelle);
- **information hierarchical**. Compte tenu de l'information résultant de la dynamique du groupe ou de la hiérarchie de appartenance (par exemple territoriale). L'hypothèse est que les paramètres empiriques pour chaque t (temp) considérée, soit les même des distribution à partir des quelles les données ont été produites (par exemple le calcul de la moyenne et l'écart des indices à base mobiles au moment t, des déterminations régional du groupe consideré)

- on procéde à l'extraction aléatoire, à partir des t distributions supposer normales[3], afin de imputer les indices à base mobiles manquants (ou l'application de la MI de Rubin);
- des contraintes peuvent être imposées, par exemple, sur le signe ou en termes d'écarts par rapport à la moyenne, afin de ne pas donner trop de variabilité aux données imputées;
- application des indices à base mobile aux données observées pour obtenir les données mancantes.

Cette méthode simple et intuitif utilise la structure temporelle et l'information disponible, en permetant de spécifier des modèles de la réalité qui, par example, ne produisent pas des estimations indéfiniment constante dans le cas limite d'une unique donnée disponible.

Figura 1. Imputation des données mancantes pour le forecasting: la méthodes ESeC-Rubin



Source: Chiodini M.P., Verrecchia F., 2008

1.4. Hierarchical forecasting

Une stratégie de prévisions sur base de données hiérarchique à besoin de l'analyse des différentes hiérarchies et des différents niveaux.

La définition de la hiérarchie d'analyse pose un premier problème, si elle n'est pas unique, mais elle est le résultat de l'intersection de plusieurs hiérarchies, par example une hiérarchie qui comprend à la fois un split géographique et de produit. On pourrait diviser le totale en premier pour unités géographiques et ensuite par produit, ou vice-versa pour produit en premier et ensuite pour géographie, ou même faire un mix des deux hiérarchies (par exemple, avant par pays, puis par produit, et ensuite pour détail de chaque nation). Une fois defini la hiérarchie, il y a un deuxième problème: la réconciliation de la prévision. L'approche traditionnelle est une approche « Bottom-up », où les prévisions sont produites au niveau le plus bas et ensuite regroupées. Les problèmes inhérents à cette technique sont liées à l'absence de l'intervalle de confiance et à la perte de précision des données agrégées. SAS Forecast Server étend l'approche traditionnelle, avec la production des estimations statistiques pour chaque niveau de la hiérarchie et avec une réconciliation automatique, ainsi que la prévision au niveau le plus élevé correspond avec la somme des prévisions générées au niveau plus bas de la hiérarchie. Ainsi est assurée la cohérence des prévisions à tous les niveaux. Plus précisément, la réconciliation peut être obtenue en utilisant trois techniques différentes:

- Bottom-up: les prévisions sont produites au niveau le plus bas de la hiérarchie et la prédiction du niveau plus élevé est obtenu par agrégation. On estime également un modèle pour le niveau supérieur, on peut évaluer les différences entre les estimations et les statistiques de prévision réconcilier afin de repérer les tendances anomales.
- Top-down: les estimations obtenues au niveau le plus élevé de la hiérarchie sont partagées entre les sous-niveaux en fonction des différents méthodes (la plus répandue prévoit une répartition proportionnelle pour les estimations statistiques générer à ce niveau). Dans ce cas aussi, il est utile pouvoir comparer les prévisions avec les statistiques réconciliées.

[3] Dans l'application sur un groupe homogène, grâce à la transformation au logarithme, on verifie la forme distributive.

- Middle-out: les prévisions sont produites à un niveau intermédiaire de la hiérarchie et la réconciliation a lieu à partir de ce niveau vers le haut (bottom-up) et vers le bas (top-down) en même temps.

En général, donc, un système hiérarchique détermine une prolifération du nombre de séries qui doit être analysées et modélisées simultanément. Il est donc essentiel de mettre en place des mécanismes pour la diagnostic des séries, pour la recherche et spécification des modèles et pour la gestion des résultats.

2. Une application sur un groupe hiérarchique administratif

2.1. Imputation des données: Moyenne, CPGR et Stochastic-imputation

L'imputation de la moyenne, considéré partie des "vieilles méthodes", ou d'autres fonctions calculé sur toutes la serie cronologiques pose problèmes quand l'imputation est relative, au même temps, à des milliers de séries temporelles. L'application présentée ici porte sur le chômage en Europe (source: Eurostat), où les données enregistrées comprennent un grand nombre de valeurs manquantes (par exemple, voir tableau 1). Dans le cas de séries non-stationnaires en moyenne l'imputation de la moyenne est une contradiction en termes (et avec autre fonctionels on aurait des contradictions avec d'autres séries). Avec l'imputation automatique de la moyenne, on peut observer des cas limite comme pour le Tirol (tableau 2): avec la moyenne (3200 unités) on attribue données au-dessus du 50% des extrêmes de l'interval 1999-2003.

Avec la CPGR les données manquantes d'une série sont imputés selon un méthode qui se fonde sur le taux de croissance annuel moyen entre deux périodes disponibles. Par exemple, dans le cas du l'Tirol (tableau 3), le taux de chômage 15-24 ans on a des valeurs manquantes entre 2000 et 2002. Avec le CPGR les données d'imputation sont prochaines (et interne) aux termes extrêmes observés, ainsi que non égale à une constante. Le CPGR utilisé pour l'imputation entre les termes de la série, cependant, n'est pas applicable lorsque la valeur est manquant au début ou à la fin de la série observée.

L'objectif de la Stochastic-imputation - dans cette section utilisé pour chaque typologie de données - est d'imputer données mancantes au début ou à la fin de la série observée. La variabilité des données imputées est nécessaire à la phase de spécification des modèles. Par exemple, pour la région Kärnten l'imputation de la moyenne porte à une spécification d'une constante quand les données observé ne sont pas stationnaires en moyenne (Tab. 8 e Fig. 4).

ESeC-Rubin se propose d'agir sur le cas non imputable avec la CPGR en utilisant des techniques différentes pour les différents typologies de données (c'est-à-dire CPGR, Stochastic-imputation, voir Anselmi, Chiodini, Verrecchia; 2008).

Table 1. Chômage (en centaines), 1999-2006

			1999	2000	2001	2002	2003	2004	2005	2006
CHÔMAGE TOTAL 15-24										
Austria	AT33	Tirol	20				21	39	44	36
Belgium	BE31	Prov. Brabant Wallon	28				29	30	35	32
Belgium	BE34	Prov. Luxembourg		25					28	30
Spain	ES23	La Rioja	26	26					24	24
Netherlands	NL23	Flevoland	16	17		24	30	41	36	34
Netherlands	NL34	Zeeland	27	21				17	19	
CHÔMAGE FÉMININ										
Austria	AT11	Burgenland	26		24	27	26	37	44	37
Austria	AT34	Vorarlberg	29		22	20	38	40	54	47
Spain	ES63	Ciudad Autónoma de Ceuta	35	45					34	34
Spain	ES64	Ciudad Autónoma de Melilla	34	31				30	23	21
Portugal	PT15	Algarve	50			57	65	55	69	65
CHÔMAGE TOTAL										
Spain	ES63	Ciudad Autónoma de Ceuta	77	79			26	29	64	62
Spain	ES64	Ciudad Autónoma de Melilla	54	55			22	48	36	35

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour Austria, Belgium, Spain, Portugal, Netherlands avec donée manquantes compris entre termes des series.

Table 2. Chômage (en centaines), avec imputation de la moyenne, 1999-2006

			1999	2000	2001	2002	2003	2004	2005	2006	MEDIA
CHÔMAGE TOTAL 15-24											
Austria	AT33	Tirol	20	32	32	32	21	39	44	36	32
Belgium	BE31	Prov. Brabant Wallon	28	31	31	31	29	30	35	32	31
Belgium	BE34	Prov. Luxembourg	28	25	28	28	28	28	28	30	28
Spain	ES23	La Rioja	26	26	25	25	25	25	24	24	25
Netherlands	NL23	Flevoland	16	17	28	24	30	41	36	34	28
Netherlands	NL34	Zeeland	27	21	21	21	21	17	19	21	21
CHÔMAGE FÉMININ											
Austria	AT11	Burgenland	26	32	24	27	26	37	44	37	32
Austria	AT34	Vorarlberg	29	36	22	20	38	40	54	47	36
Spain	ES63	Ciudad Autónoma de Ceuta	35	45	37	37	37	37	34	34	37
Spain	ES64	Ciudad Autónoma de Melilla	34	31	28	28	28	30	23	21	28
Portugal	PT15	Algarve	50	60	60	57	65	55	69	65	60
CHÔMAGE TOTAL											
Spain	ES63	Ciudad Autónoma de Ceuta	77	79	56	56	26	29	64	62	56
Spain	ES64	Ciudad Autónoma de Melilla	54	55	42	42	22	48	36	35	42

Source: Elaborations ESeC sur donnée Eurostat. **Note**: **1**. Pour Austria, Belgium, Spain, Portugal, Netherlands avec donée manquantes compris entre termes des series.

Table 3. Chômage (en centaines),avec imputation CPGR, 1999-2006

			1999	2000	2001	2002	2003	2004	2005	2006
CHÔMAGE TOTAL 15-24										
Austria	AT33	Tirol	20	20	20	21	21	39	44	36
Belgium	BE31	Prov. Brabant Wallon	28	28	28	29	29	30	35	32
Belgium	BE34	Prov. Luxembourg		25	26	26	27	27	28	30
Spain	ES23	La Rioja	26	26	26	25	25	24	24	24
Netherlands	NL23	Flevoland	16	17	20	24	30	41	36	34
Netherlands	NL34	Zeeland	27	21	20	19	18	17	19	
CHÔMAGE FÉMININ										
Austria	AT11	Burgenland	26	25	24	27	26	37	44	37
Austria	AT34	Vorarlberg	29	25	22	20	38	40	54	47
Spain	ES63	Ciudad Autónoma de Ceuta	35	45	43	40	38	36	34	34
Spain	ES64	Ciudad Autónoma de Melilla	34	31	31	30	30	30	23	21
Portugal	PT15	Algarve	50	52	55	57	65	55	69	65
CHÔMAGE TOTAL										
Spain	ES63	Ciudad Autónoma de Ceuta	77	79	55	38	26	29	64	62
Spain	ES64	Ciudad Autónoma de Melilla	54	55	41	30	22	48	36	35

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour Austria, Belgium, Spain, Portugal, Netherlands avec données manquantes compris entre termes des series.

Table 4. Chômage 15-24 (en centaines), Austria, 1999-2006

		1999	2000	2001	2002	2003	2004	2005	2006
AT12	Niederösterreich	36	37	51	53	58	91	94	86
AT13	Wien	60	59	70	87	112	151	189	183
AT21	Kärnten					24	33	40	
AT22	Steiermark	43	48	48	42	44	60	74	63
AT31	Oberösterreich	51	49	50	51	59	89	76	64
AT32	Salzburg				20				
AT33	Tirol	20				21	39	44	36
AT34	Vorarlberg							31	

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour at11 - Burgenland (A) on a pas données

Table 5. Chômage 15-24 (indices à base mobile), Austria, 2000-2006

		Indice 00 (1999 = 100%)	Indice 01 (2000 = 100%)	Indice 02 (2001 = 100%)	Indice 03 (2002 = 100%)	Indice 04 (2003 = 100%)	Indice 05 (2004 = 100%)	Indice 06 (2005 = 100%)
AT12	Niederösterreich	102.8%	137.8%	103.9%	109.4%	156.9%	103.3%	91.5%
AT13	Wien	98.3%	118.6%	124.3%	128.7%	134.8%	125.2%	96.8%
AT21	Kärnten					137.5%	121.2%	
AT22	Steiermark	111.6%	100.0%	87.5%	104.8%	136.4%	123.3%	85.1%
AT31	Oberösterreich	96.1%	102.0%	102.0%	115.7%	150.8%	85.4%	84.2%
AT32	Salzburg							
AT33	Tirol					185.7%	112.8%	81.8%
AT34	Vorarlberg							

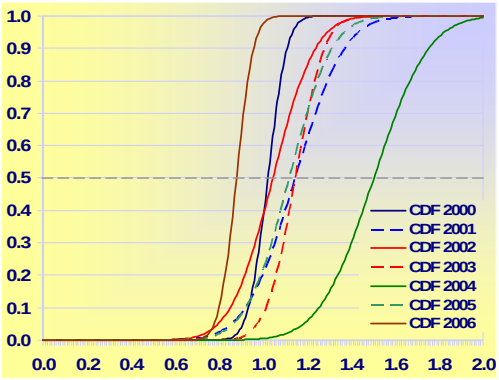
Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour at11 - Burgenland (A) on a pas données

Table 5. Chômage 15-24 (statistiques sur les indices), Austria, 2000-2006

	1999	2000	2001	2002	2003	2004	2005	2006
observations non manquantes		4	4	4	4	6	6	5
minimum		96%	100%	88%	105%	135%	85%	82%
maximum		112%	138%	124%	129%	186%	125%	97%
médiane		101%	110%	103%	113%	144%	117%	85%
moyenne		102%	115%	104%	115%	150%	112%	88%
écart quadratique moyen		0.0687	0.1758	0.1513	0.1040	0.1947	0.1530	0.0614
Variance		0.0047	0.0309	0.0229	0.0108	0.0379	0.0234	0.0038

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour at11 - Burgenland (A) on a pas données

Figure 2. CDF teoriques des indices à base mobile du Chômage 15-24, Austria, 2000-2006



Source: Elaborations ESeC

Table 7. Nombres casuels - distributions normales (Moyenne et écart quadratique moyen des indices nazionals), Austria, 2000-2006

		Indices 00 (1999 = 100%)	Indices 01 (2000 = 100%)	Indices 02 (2001 = 100%)	Indices 03 (2002 = 100%)	Indices 04 (2003 = 100%)	Indices 05 (2004 = 100%)	Indices 06 (2005 = 100%)
AT12	**Niederösterreich**	**102.8%**	**137.8%**	**103.9%**	**109.4%**	**156.9%**	**103.3%**	**91.5%**
AT13	**Wien**	**98.3%**	**118.6%**	**124.3%**	**128.7%**	**134.8%**	**125.2%**	**96.8%**
AT21	**Kärnten**	115.4%	112.0%	98.9%	**91.5%**	**137.5%**	**121.2%**	85.8%
AT22	**Steiermark**	**111.6%**	**100.0%**	**87.5%**	**104.8%**	**136.4%**	**123.3%**	**85.1%**
AT31	**Oberösterreich**	**96.1%**	**102.0%**	**102.0%**	**115.7%**	**150.8%**	**85.4%**	**84.2%**
AT32	**Salzburg**	105.5%	141.5%	**118.3%**	117.2%	147.2%	104.7%	90.6%
AT33	**Tirol**	**111.9%**	87.6%	104.8%	**110.4%**	**185.7%**	**112.8%**	**81.8%**
AT34	**Vorarlberg**	106.7%	149.5%	83.3%	102.4%	134.7%	**119.4%**	89.9%

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour at11 - Burgenland (A) on a pas données.

Table 8. Chômage 15-24 (en centaines) – interpretation des missing value: Stochastic-Imputation, Austria, 1999-2006

		1999	2000	2001	2002	2003	2004	2005	2006
AT12	**Niederösterreich**	**36**	**37**	**51**	**53**	**58**	**91**	**94**	**86**
AT13	**Wien**	**60**	**59**	**70**	**87**	**112**	**151**	**189**	**183**
AT21	**Kärnten**	21	24	27	26	**24**	**33**	**40**	34
AT22	**Steiermark**	**43**	**48**	**48**	**42**	**44**	**60**	**74**	**63**
AT31	**Oberösterreich**	**51**	**49**	**50**	**51**	**59**	**89**	**76**	**64**
AT32	**Salzburg**	11	12	17	**20**	23	34	36	33
AT33	**Tirol**	**20**	22	20	21	**21**	**39**	**44**	**36**
AT34	**Vorarlberg**	14	15	23	19	19	26	**31**	28
AT	**Austria**	256	266	306	319	360	523	584	527

Source: Elaborations ESeC sur données Eurostat. **Note**: **1**. Pour at11 - Burgenland (A) on a pas données.

2.2. Prévisions basées sur l'imputation des données manquantes

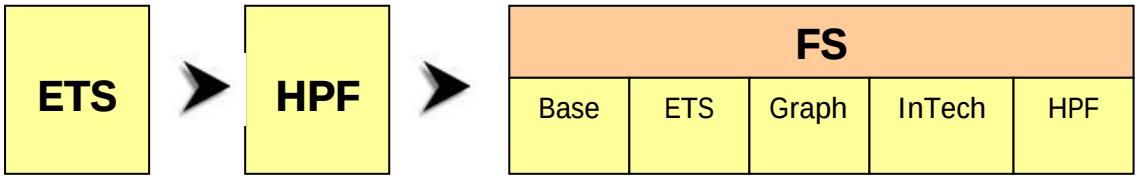
2.2.1. Système d'information pour la sélection automatique des modèles

SAS Forecast Server (figure 3) produit automatiquement des prévisions statistiques pour toutes les séries d'enquêtes object d'analise. Si on spécifie une hiérarchie, qui est également automatiquement réconciliée par le système en fonction des options qu'on définit. L'outil construit un modèle en choisissant parmi les familles Exponential Smoothing, ARIMAX, etc. Si on a des variables

indépendantes ou des variable événement, le système détermine automatiquement les variables qui ont une corrélation avec la variable dépendante, identifiant quelle est la fonction de transfert la plus approprié.

En plus des modèles automatiques, on peut inclure une liste des modèles personnalisés qui sont pris en compte dans le processus de sélection du meilleur modèle parmi les candidats. Le modèle final utilisé pour la prévision d'une série régionale est ensuite sélectionnées sur la base des statistiques de bonté d'adaptation (par exemple MAPE).

Figure 3. Evolution des instruments de prévision et modules qui composent la Plate-forme de Forecast Server



Source: SAS.

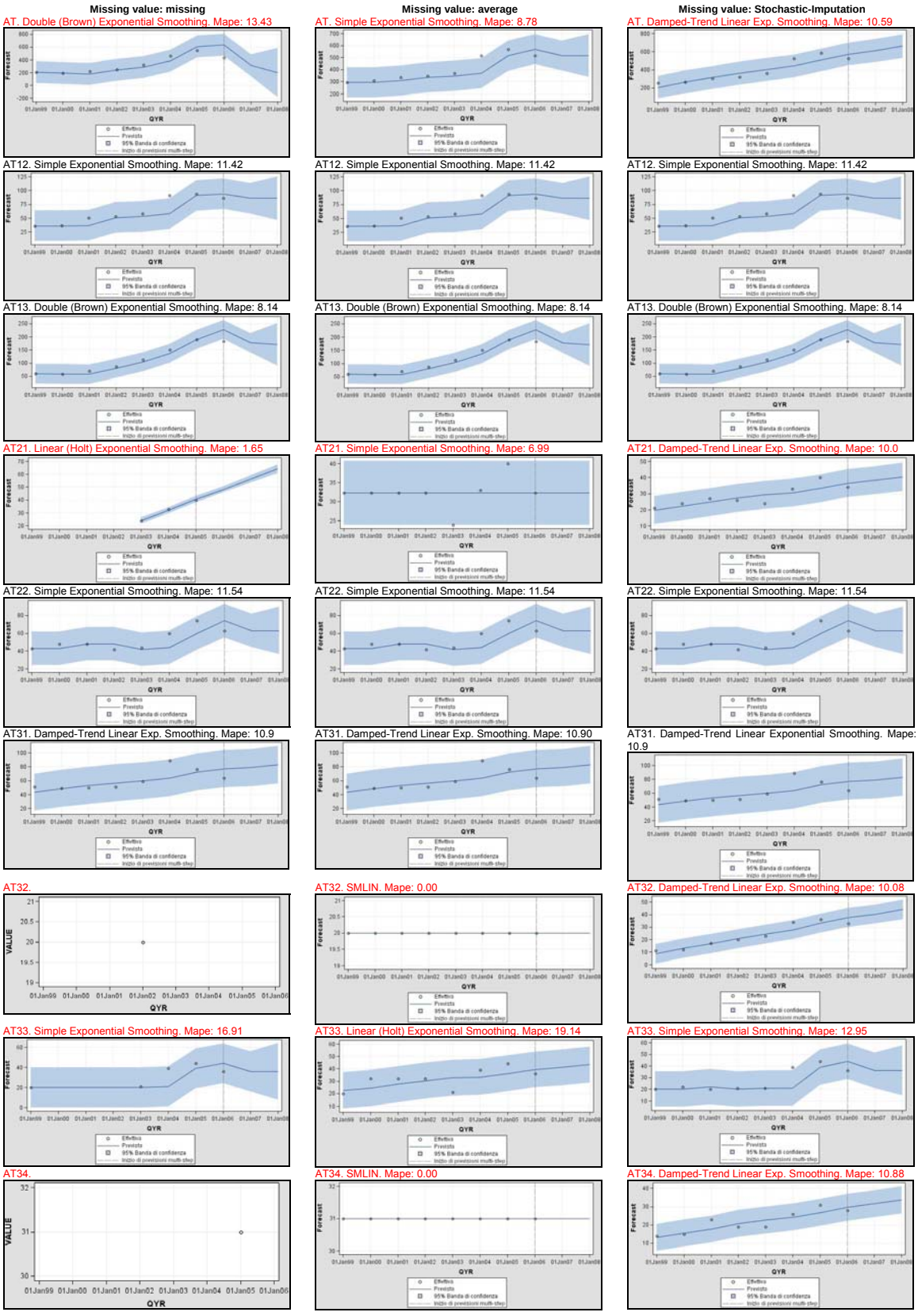
2.2.2. La spécification des modèles sur bases de données traitées avec différentes méthodes d'imputation

On va présenter une application d'imputation stochastique, avec le but de comparer des différences dans la sélection automatique des modèles à partir de la base de données observées. À ce fin, nous allons utiliser la plate-forme SAS Forecast pour la spécification des modèles pour les séries temporelles sur le chômage de l'Union européenne:
- sans aucune méthode de l'imputation des valeurs manquantes;
- données mancantes imputées par la moyenne;
- données mancantes imputées avec imputation stochastique.

Les résultats relatifs aux régions autrichiennes sont observées graphiquement en Figure 4. Tout d'abord, on peut noter que l'imputation stochastique permet de spécifier des modèles de réalité qui ne produisent pas des estimations indéfiniment costante, même dans le cas limite d'une seule donnée pour unité territoriale observée. D'intérêt les résultats en termes de modélisation et de bandes de confiance. En particulier:

- les statistiques de Fit pour la sélection automatique des modèles régionaux sont toujours calculables sur base de données traitées avec imputation stochastique – dans cette perspective les modèles stationnaire en moyenne ou on utilise seulement une donnée et l'imputation d'une constante ne sont pas considérées bonne synthèse de la réalité, même si le MAPE est égale à zéro (comme dans le cas du chômage 15-24 ans dans les régions autrichiennes du Vorarlberg et de Salzbourg);
- en outre les bandes de confiance sont toujours représentables avec l'imputation stochastique;
- les modèles automatiquement sélectionnés sont non banal seulment avec l'imputation stochastique. Emblématique, le cas du chômage (15-24 ans) de la région autrichienne de Kärnten, où la première et la dernière données de la série sont manquantes: l'imputation de la moyenne conduit à la spécification d'un modèle constant même en présence d'un série non-stationnaire en moyenne.

Figure 4. Chômage 15-24 (en centaines), Austria, previsions 2007-2008



Source: Elaborations ESeC sur données Eurostat.
Note: **1**. Pour at11 - Burgenland (A) on a pas données. **2**. at Austria; at12 Niederösterreich; at13 Wien; at21 Kärnten; at22 Steiermark; at31 Oberösterreich; at32 Salzburg; at33 Tirol; at34 Vorarlberg.

3. Une application sur un groupe homogène en termes de structure et de performances

Le travail est complété par une analyse effectuée sur les données régionales de l'Europe (source: Eurostat). Les régions européennes sont d'abord regroupées en groupes structurel et puis de performance. Pour l'analyse, nous avons utilisé 80 régions à base occupational limité (d'un point de vue structurel) et à participation et situation économique intermédiaire (voir BiCRA-PSG en: Eusepi, Cepparulo, Verrecchia 2007). Six régions autrichiennes sont présentes dans ce groupe (Basse-Autriche, Styrie, Haute, Carinthie, Tyrol, Vorarlberg), dont les derniers trois, comme vue en précédence, avec des données manquantes. Avant de procéder avec l'application de la méthode pour la reconstruction des donnée manquantes, on à appliquée une transformation logarithmique à les séries des numéros indices à base mobile en portant une correction en termes de distribution (distribution lisse).
En Figure 5, on peut observer les distributions des données traitées. Il semble évident que les distributions ont des formes que l'on peut raisonnablement supposer normal, à la lumière de l'évolution du Q-Q plot e des indicateurs (Skewness, Kurtosis).
Dans le Tableau 11 on peut voir le sommaires des statistiques pour les séries temporelles (2000-2006) tandis que dans les tableaux 12 et 13, respectivement, ils peuvent être observées les séries temporelles des numéros l'indice à base mobiles reconstruites et leur transformation en données en unitè de mesure original. On peut immédiatement vérifier que la méthode proposée pour la reconstruction des séries temporelles permet d'obtenir des données ayant une variabilité naturelle.
Le MI de Rubin, également applicables à l'analyse des données, et qui peut porter les valeurs imputées à une variation autour de la moyenne pour le groupe concerné, peut être d'intérêt dans la selection des modèles pour le prévision.

Table. 9 Régions à base occupational limité (classification structurel)

	N.	Degré de urbanisation (habitant par km2)	Employés (unité)	Spécialisation tertiaire (%)
Intervalle	**183**	**< 1.639**	**< 1.720.600**	**43% - 80%**

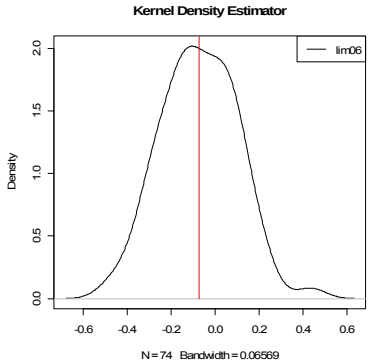
Source: Eusepi, Cepparulo, Verrecchia 2007.

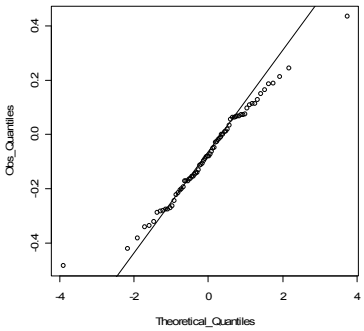
Table. 10 Régions à base occupational limité (classification basè sur les performances)

	N.	PIB par habitant (EU =100%)	Productivité (EU =100%)	Taux d'employ 15-64 (%)	Taux d'employ 15-64, Fémmin (%)	Taux d'employ 55-64 (%)	Taux d'employ 55-64, Fémmin (%)	Taux de chômage de long période (%)
Régions à participation intermédiaire et situation économique optimal	2	< 217.3%	< 164.3%	< 69.1%	< 61.9%	< 41.1%	< 29.4%	< 1.6%
Régions à participation intermédiaire et bonne situation économique	24	< 138.9%	< 133.1%	< 76.7%	< 70.6%	< 49.4%	< 38.4%	< 3.5%
Régions à forte participation au marchè du travail et situation économique intermédiaire	33	< 149.3%	< 115.1%	< 78.6%	< 76.1%	< 74.1%	< 72.8%	< 1.9%
Régions à participation et situation économique intermédiaire	80	< 139.8%	< 119.0%	< 71.5%	< 65%.0	< 49.3%	< 44.9%	< 6.4%
Régions à forte participation au marchè du travail et situation économique minimal	2	< 76.8%	< 55.1%	< 74.2%	< 70.8%	< 64.7%	< 57.1%	< 1.3%
Régions à participation intermédiaire et situation économique minimal	29	< 83.6%	< 93.3%	< 68.9%	< 61.5%	< 55.4%	< 44.1%	< 16.0%
Régions à participation et situation économique minimal	13	< 76.6%	< 105.8%	< 53.2%	< 44.9%	< 40.6%	< 33.0	< 22.3%
Régions à base occupational limité	183	48%-217%	54%-164%	40.1%-78.6%	24%-6.1%	21.6%-74.1%	11.4%-72.8%	0.1% - 22.3%
UE (15 pays)	213	100% (23428€)	100% (52405€)	64.3%	56.0%	41.7%	32.2%	3.3%

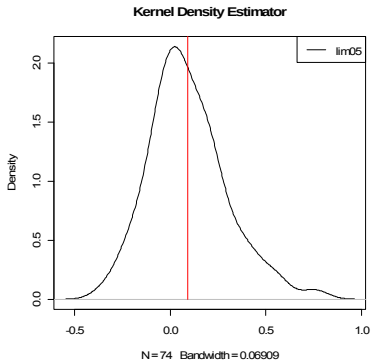
Source: Elaborations ESeC sur données Eurostat. **Note**: 1. Régions à participation et situation économique intermédiaire: Burgenland (AT11), Niederösterreich (AT12), Kärnten (AT21), Steiermark (AT22), Oberösterreich (AT31), Tirol (AT33), Vorarlberg (AT34), Prov. Liège (BE33), Prov. Luxembourg (B) (BE34), Prov. Namur (BE35), Karlsruhe (DE12), Freiburg (DE13), Tübingen (DE14), Niederbayern (DE22), Oberpfalz (DE23), Oberfranken (DE24), Mittelfranken (DE25), Unterfranken (DE26), Schwaben (DE27), Bremen (DE50), Gießen (DE72), Kassel (DE73), Braunschweig (DE91), Hannover (DE92), Lüneburg (DE93), Weser-Ems (DE94), Münster (DEA3), Detmold (DEA4), Amsberg (DEA5), Koblenz (DEB1), Trier (DEB2), Rheinhessen-Pfalz (DEB3), Saarland (DEC0), Schleswig-Holstein (DEF0), Cantabria (ES13), Pais Vasco (ES21), Comunidad Foral de Navarra (ES22), La Rioja (ES23), Aragón (ES24), Castilla y León (ES41), Illes Balears (ES53), Región de Murcia (ES62), Canarias (ES) (ES70), Itä-Suomi (FI13), Länsi-Suomi (FI19), Pohjois-Suomi (FI1A), Champagne-Ardenne (FR21), Picardie (FR22), Haute-Normandie (FR23), Centre (FR24), Basse-Normandie (FR25), Bourgogne (FR26), Nord - Pas-de-Calais (FR30), Lorraine (FR41), Alsace (FR42), Franche-Comté (FR43), Pays de la Loire (FR51), Bretagne (FR52), Poitou-Charentes (FR53), Aquitaine (FR61), Midi-Pyrénées (FR62), Limousin (FR63), Auvergne (FR72), Languedoc-Roussillon (FR81), Sterea Ellada (GR24), Attiki (GR30), Notio Aigaio (GR42), Border, Midlands and Western (IE01), Valle d'Aosta/Vallée d'Aoste (ITC2), Umbria (ITE2), Marche (ITE3), Abruzzo (ITF1), Lisboa (PT17), Tees Valley and Durham (UKC1), Northumberland, Tyne and Wear (UKC2), East Riding and North Lincolnshire (UKE1), South Yorkshire (UKE3), West Wales and The Valleys (UKL1), South Western Scotland (UKM3), Northern Ireland (UKN0).

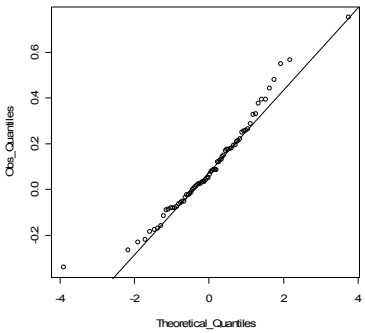
Figure 5. Distribution des logarithme du Chômage 15-24, pour les régions à participation et situation économique intermédiaire (KDE, QQplot, Skewness, Kurtosis), 2000-2006



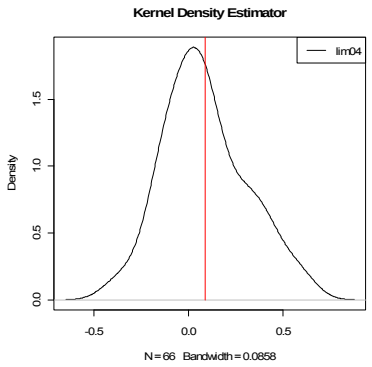


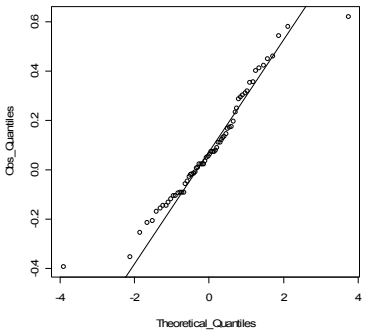
2006	
N:	74
Skewness:	0.101325
Kurtosis:	2.971209



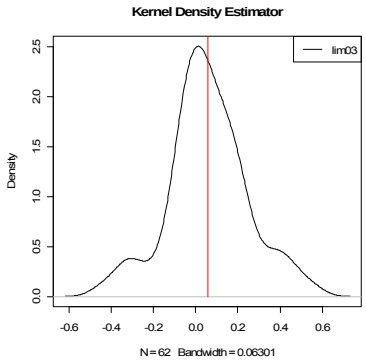


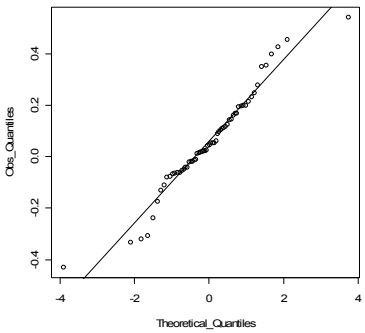
2005	
N:	74
Skewness:	0.668547
Kurtosis:	3.747535



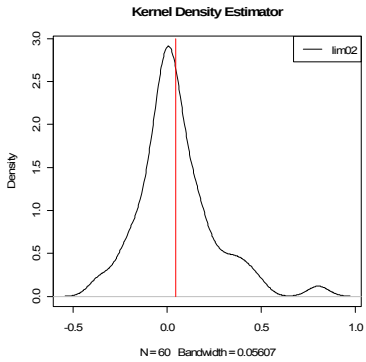


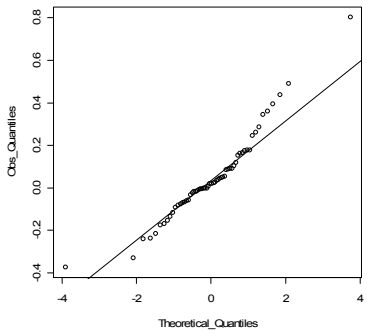
2004	
N:	66
Skewness:	0.374558
Kurtosis:	2.789564



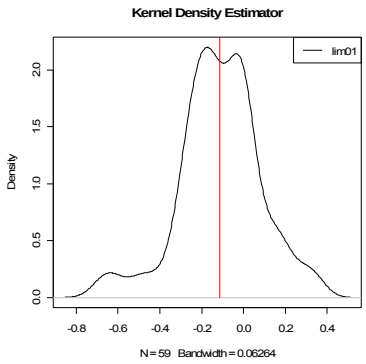


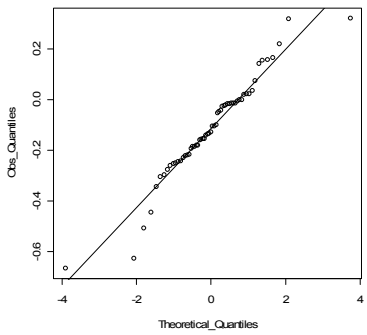
2003	
N:	62
Skewness:	0.034826
Kurtosis:	3.569263



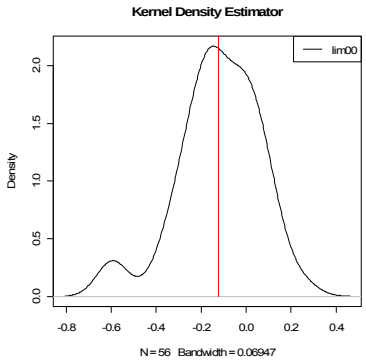


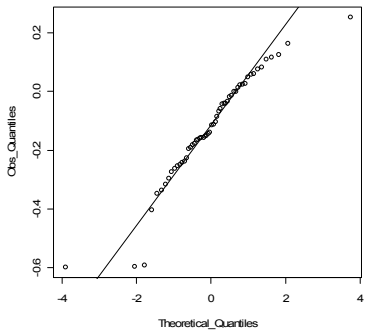
2002	
N:	60
Skewness:	1.008929
Kurtosis:	5.433043





2001	
N:	59
Skewness:	-0.365568
Kurtosis:	3.963497





2000	
N:	56
Skewness:	-0.660807
Kurtosis:	3.641830

Source: Elaborations ESeC sur données Eurostat.

Table 11. Chômage 15-24 (statistiques sur les indices), régions à participation et situation économique intermédiaire, 2000-2006

	1999	2000	2001	2002	2003	2004	2005	2006
observations non manquantes		56	59	60	62	66	74	74
minimum		55%	51%	69%	65%	68%	71%	62%
maximum		129%	138%	223%	172%	186%	213%	155%
médiane		88%	88%	102%	105%	106%	106%	92%
moyenne		90%	91%	107%	108%	112%	112%	94%
écart quadratique moyenne		15%	17%	24%	21%	26%	25%	17%
variance		0.0239	0.0300	0.0596	0.0425	0.0667	0.0614	0.0274

Source: Elaborations ESeC sur données Eurostat.

Table 12. Nombres casuels - distributions normales (Moyenne et écart quadratique moyen des indices du groupe), Régions de l'Autriche, 2000-2006

		Indices 00 (1999 = 100%)	Indices 01 (2000 = 100%)	Indices 02 (2001 = 100%)	Indices 03 (2002 = 100%)	Indices 04 (2003 = 100%)	Indices 05 (2004 = 100%)	Indices 06 (2005 = 100%)
AT21	**Kärnten**	86.8%	89.5%	98.8%	72.8%	**137.5%**	**121.2%**	87.5%
AT33	**Tirol**	126.9%	92.5%	110.4%	124.1%	**185.7%**	**112.8%**	**81.8%**
AT34	**Vorarlberg**	77.7%	156.4%	101.7%	103.9%	123.3%	92.9%	97.6%

Source: Elaborations ESeC sur données Eurostat.

Table 13. Chômage 15-24 (en centaines) – interpretation des missing value: Stochastic-Imputation, Régions de l'Autriche, 1999-2006

		1999	2000	2001	2002	2003	2004	2005	2006
AT21	**Kärnten**	43	37	33	33	**24**	**33**	**40**	35
AT33	**Tirol**	**20**	25	23	26	**21**	**39**	**44**	**36**
AT34	**Vorarlberg**	21	16	26	26	27	33	**31**	30

Source: Elaborations ESeC sur données Eurostat.

4. Conclusions

L'application de ESeC-Rubin et la comparaisons avec les "vieilles méthodes" (considé le cas de la moyenne conditionnelle), montre clairement les améliorations possibles résultant d'une approche multidisciplinaire qui développe les propositions méthodologiques de la théorie des échantillons et celle des numéros indice dans le cadre des séries temporelles économiques. Il s'agit évidemment d'un premier emploi avec un benchmark et des cas d'analyse empirique. La transformation logarithmique des données, l'étude sur les hypothèses de normalité, l'ancrage national et de groupe (voir BiCRA-PSG en Eusepi, Cepparulo, Verrecchia 2007) sont traités dans le papier. L'évolution naturelle des applications se rapporte à l'imputation multiple (Rubin 1996), et à la vérification de la robustesse de la méthode tant en termes d'imputations que de prévision (voir, par exemple, Verrecchia, Chiodini, Coin, Facchinetti, Nai Rusconi 2008).

Bibliographie

[1] AA.VV., «Handling missing data: applications to environmental analysis»,
[2] Anselmi A., Chiodini P.M., Verrecchia F., «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],
[3] Bailar B.A. and Bailar J.C. III, «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,
[4] Buck S.F., «A method of estimation of missing values in multivariate data suitable for use with an electronic computer», *Journal of the Royal Statistical Society B*, 1960, vol. 22, n° 2, pp 302-306,
[5] Chiodini P.M., Verrecchia F., «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],
[6] Eusepi G., Cepparulo A., Verrecchia F., «Bilevel Comparative Regional Analysis - Performances in Structural Grid.», ESeC. Working paper ESeC_WP001,
[7] Herzog T.N. and Rubin D.B., «Using multiple imputation to handle nonresponse in sample surveys.» In Incomplete Data in Sample Surveys, Vol. 2,
[8] Little R.J.A., Rubin D.B., «Statistical analysis with missing data»,
[9] Martini M., «Numeri indice per il confronto nel tempo e nello spazio»,
[10] Rubin D.B., «Multiple imputation for Nonresponse in Surveys»
[11] Rubin D.B., «Multiple imputation after 18+ year». *Journal of the American Statistical Association*, June 1996, Vol. 91, n°. 434, pp 507-510.
[12] Rubin D.B. and Schenker N., «Multiple imputation for interval estimation from simple random samples with ignorable nonresponse». *Journal of the American Statistical Association*, June 1996, Vol. 81, n° 394, pp 366-374.
[13] «SAS for Forecasting Time Series», by John Brocklebank and David Dickey Copyright(c) 2003,
[14] SAS Institute Inc.,«SAS Forecast Studio 1.4: User's Guide»,
[15] SAS Institute Inc., «SAS Forecast Server 1.4: Administrator's Guide»,
[16] «SAS Forecast Server». Cary, NC: SAS Institute Inc. Online:
[http://www.sas.com/technologies/analytics/forecasting/forecastserver/factsheet.pdf],
[17] Schafer J.L. and Graham J.W., «Missing Data: Our View of the State of the Art», *Psychological Methods*, 2002, Vol. 7, n° 2, pp 147-177,
[18] Verrecchia F.,«The Generalised Index Numbers», *Journal of ESeC Short Papers*, 2008, Vol.1, n°1, pp 9-12,
[19] Verrecchia F., «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],
[20] Verrecchia F., Chiodini P.M., Coin D., Facchinetti S., Nai Ruscone M., «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].