Données manquantes et prévisions: Méthodes à imputation variable

Antonio ANSELMI
SAS Institute, Customer Support

Paola M. Chiodini
Université of Milano-Bicocca

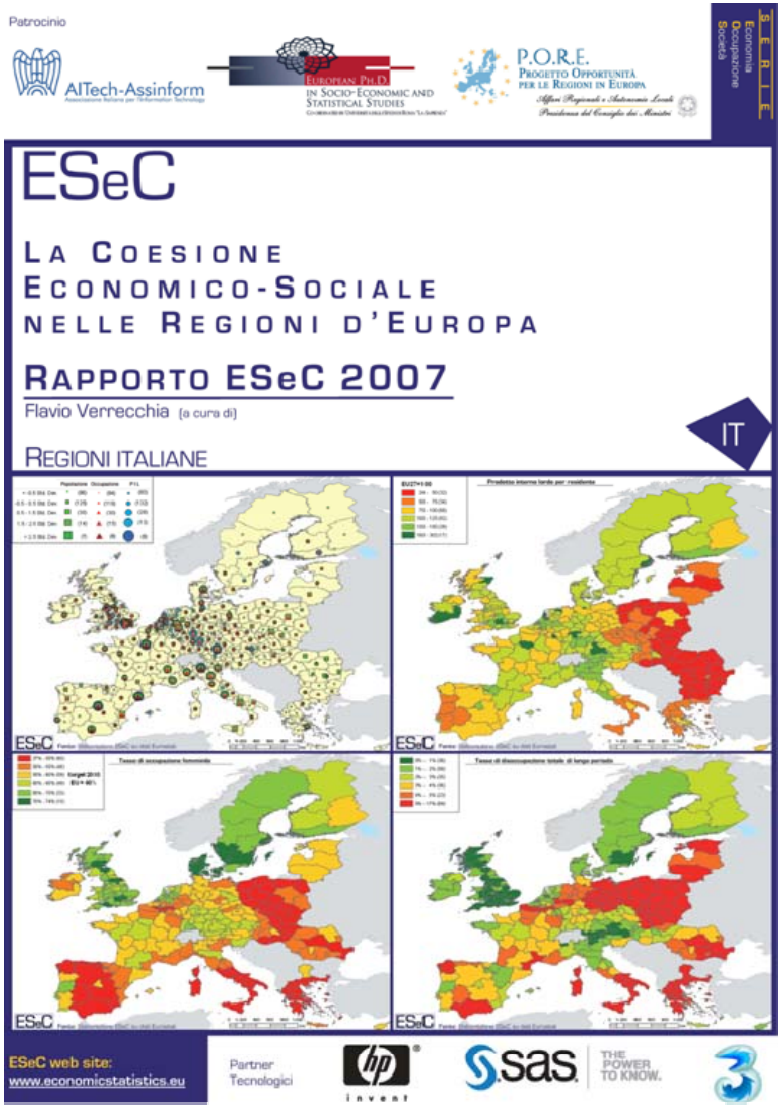
Flavio Verrecchia
ESeC

JMS - Journées de Méthodologie Statistique
Paris, le 24 mars 2009

Introduction

- Dans un contexte de séries temporelles, pour l'imputation des données manquantes on se réfère à la moyenne arithmétique. Si les séries sont n ($n \rightarrow \infty$), il est impossible trouver une fonction unique pour les n constant d'imputation des données manquantes.
- Le but devient celui d'évaluer les nouvelles propositions de méthodes d'imputation des données manquantes pour bases de données hiérarchique achevés pour la mise en pratique des modèles pour séries temporelles.
- En particulier nous allons présenter des:
  - **méthodes d'imputation qui aide à reconstruire les données manquantes en tenant compte de la naturelle variabilité du phénomène à l'étude.**
  - **applications***:*
    - Imputation de la moyenne, imputation CAGR, et imputation stochastique
    - SAS Forecast Server, utilisé pour la spécification des modèles (automatiquement sélectionnés), permettra de comparer les modèles en partant des données traitées avec différents typologies d'imputation.

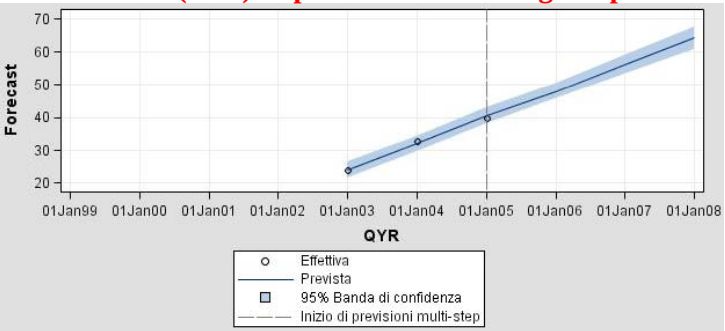
Objectif: Report sur les régions d'Europe avec prévisions 2008



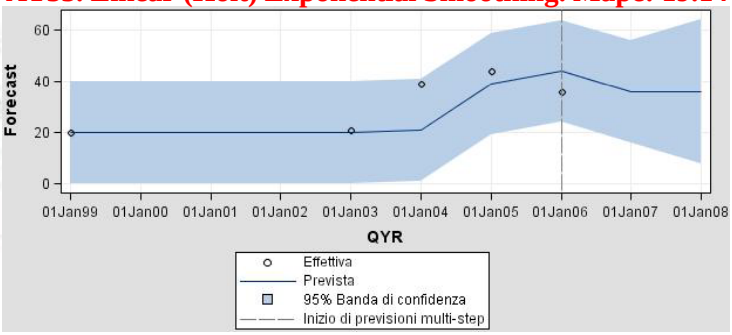
Problème de prévision: unique function de replacement des missing

Missing value: missing

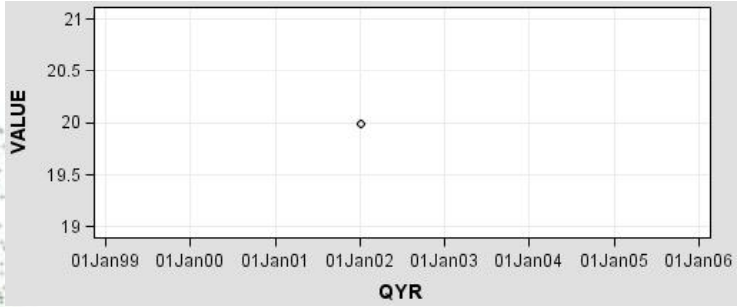
AT21. Linear (Holt) Exponential Smoothing. Mape: 1.65



AT33. Linear (Holt) Exponential Smoothing. Mape: 19.14

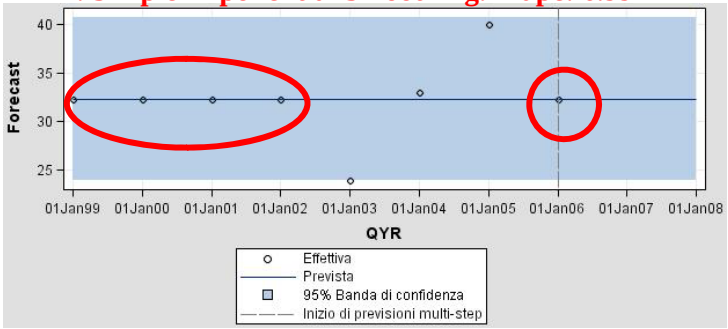


AT32.

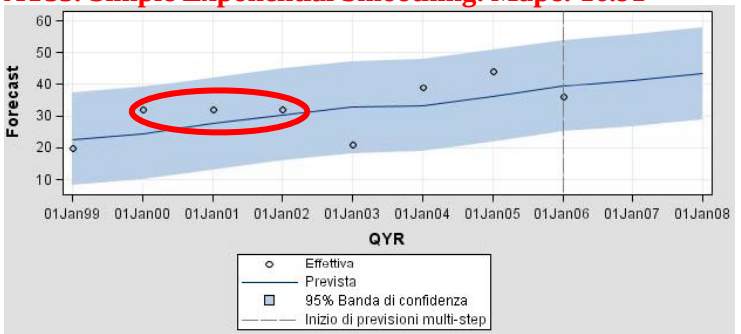


Missing value: moyenne

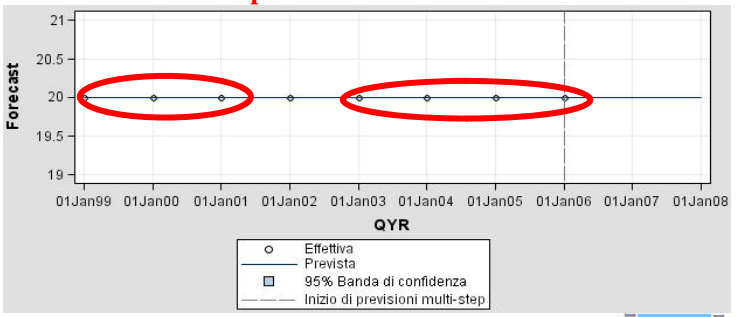
AT21. Simple Exponential Smoothing. Mape: 6.99



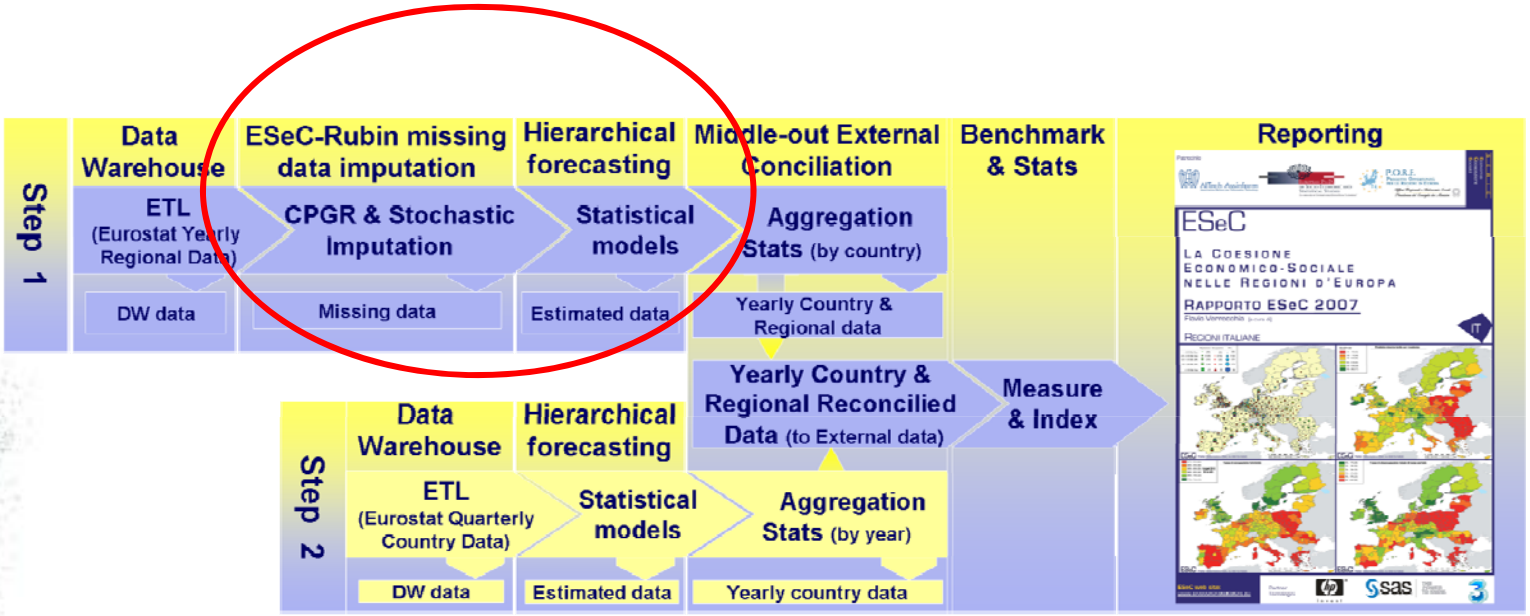
AT33. Simple Exponential Smoothing. Mape: 16.91



AT32. SMLIN. Mape: 0.00



Solution



Méthodes

- Echantillonnage et traitement des données manquantes
- Méthodes d'imputation
- Nombres Indice
- Hierarchical forecasting

Echantillonnage et traitement des données manquantes

Donnée manquantes par hasard (MAR *missing at random*)

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.

Donnée manquantes complètement par hasard (MCAR *missing completely at random*)

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.

Donnée manquantes que l'on peut pas ignorer

Est 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.

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

Méthodes d'imputation

Vieilles méthodes d'imputation	Méthodes à imputation variable (pour données hiérarchiques)
• Substitution	
• Moyenne non conditionnelle	• Imputation stochastique
• Moyenne conditionnelle (ou de régression)	• Imputation multiple
• Régression stochastique	
• Hot-deck	
• Nearest-Neighbour interpolation	
• Cold-Deck	

Méthodes d'imputation: substitution

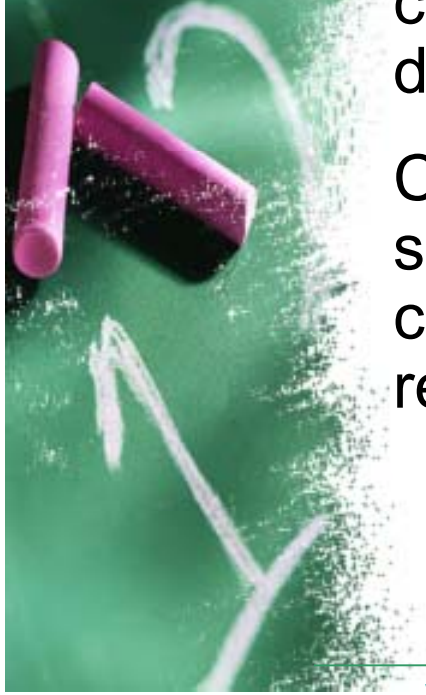
L'élément manquant est remplacé par l'utilisation d'une unité initialement non présente dans l'échantillon, mais qui est similaire à celle manquante (par exemple non-répondants dans les enquêtes sur la population).

Il est clair que les données rassemblé avec cette méthode il faut les traiter comme des données imputées.

Méthodes d'imputation: moyenne non conditionnelle

On remplace la donnée manquante par la moyenne des informations observée (la moyenne de l'échantillon) ou par la mode (données catégoriques).

On obtient 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.



Méthodes d'imputation: 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.

Méthodes d'imputation: régression stochastique

On remplace les données manquantes en utilisant toujours les informations obtenues par le « biais » 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.

Méthodes d'imputation: hot-deck

C'est une méthode largement utilisée (voir Bailar et al. 1978). Pour chaque élément manquant on identifie un cas observé similaire. Pour cet objectif, les variables sont regroupées en 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 donnée utilisée à la place de l'élément manquant est celle qui a une distance minimal 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 peuvent 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).

Méthodes d'imputation: 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 observée et celle manquante, 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).

Méthodes d'imputation: Imputation multiple

La méthode de imputation multiple arrive a résoudre la plus grande partie des problèmes des méthodes diimputation précédemment introduit.

En fait, l'imputation avec une valeur constante ne peut jamais reproduire la variabilité typique d'un phénomène et 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).

Nombres Indice: Nombre indice élémentaire

Déf. 1. 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 **E.** $\in \Re^{+z}$ le vecteur des éléments de base pour T+1 situations (avec t = 0, 1, 2, …, T) et $\mathbf{I^c_{[b\cap t]}}$ 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 $\mathbf{i^c_{[b\cap t]}}$ les déterminations des variables aléatories **E.** e $\mathbf{I^c_{[b\cap t]}}$. Soit ($\mathbf{E_b}$, $\mathbf{E_t}$, $\mathbf{Ic_{[b\cap t]}}$) 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

$f(\mathbf{E_b}, \mathbf{E_t}, \mathbf{I^c_{[b\cap t]}}) : F \rightarrow [0;\infty)$

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

Nombres Indice: indice élémentaire à base fixe et mobile

Déf. 2. Avec b = 0 la $f(\mathbf{E_b},\mathbf{E_t}, \mathbf{I^c_{[b\cap t]}})$: $F \rightarrow [0;\infty)$ est un nombre indice élémentaire à base fixe.

Déf. 3. Avec b = t-1la $f(\mathbf{E_b},\mathbf{E_t}, \mathbf{I^c_{[b\cap t]}}) : F \rightarrow [0;\infty)$ est un nombre indice élémentaire à base mobile

Nombres Indice:
CPGR (Compound Periodical Growth Rate)

Déf. 4. 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 manquants 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.

Hierarchical forecasting

Une stratégie de prévisions sur base de données hiérarchiques 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.

Une fois definit 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. 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 réconciliés.
- **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 doivent 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.

Les applications

- App. 1 - Application de la moyenne et de la CPGR
- App. 2 - Application sur un groupe hiérarchique administratif
- App. 3 - Prévisions basées sur l'imputation des données manquantes
- App. 4 - Application sur un groupe homogène en termes de structure et de performances

App. 1 - Application de la moyenne et de la CPGR

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

		1999	2000	2001	2002	2003	2004	2005	2006	μ
Austria	Tirol	20	32	32	32	21	39	44	36	32
Belgium	P. Brabant Wallon	28	31	31	31	29	30	35	32	31
Belgium	P. Luxembourg	28	25	28	28	28	28	28	30	28
Spain	La Rioja	26	26	25	25	25	25	24	24	25
Netherlands	Flevoland	16	17	28	24	30	41	36	34	28
Netherlands	Zeeland	27	21	21	21	21	17	19	21	21

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

		1999	2000	2001	2002	2003	2004	2005	2006
Austria	Tirol	20	20	20	21	21	39	44	36
Belgium	P. Brabant Wallon	28	28	28	29	29	30	35	32
Belgium	P. Luxembourg		25	26	26	27	27	28	30
Spain	La Rioja	26	26	26	25	25	24	24	24
Netherlands	Flevoland	16	17	20	24	30	41	36	34
Netherlands	Zeeland	27	21	20	19	18	17	19	

App. 1 - Application sur un groupe hiérarchique administratif: l'approche ESeC-Rubin

ESeC-Rubin est basée sur l'utilisation séquentielle des méthodes d'imputation des données en utilisant les relations temporelles et spatiales et le 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 manquantes entre deux observations disponible.

- **numéros purs**. Les indices a base mobile sont calculés sur la base des déterminations de chaque séries (la méthode est donc applicable indépendamment du facteur d'échelle);

- **information hiérarchique**. 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 (temps) 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 considérée)

- on procède à l'extraction aléatoire, à partir des t distributions supposer normales, 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 manquantes.

App. 1 - Application sur un groupe hiérarchique administratif

Chômage 15-24 (en centaines), Autriche, 1999-2006

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

Chômage 15-24 (indices à base mobile), Autriche, 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%)
Niederösterreich	102.8%	137.8%	103.9%	109.4%	156.9%	103.3%	91.5%
Wien	98.3%	118.6%	124.3%	128.7%	134.8%	125.2%	96.8%
Kärnten					137.5%	121.2%	
Steiermark	111.6%	100.0%	87.5%	104.8%	136.4%	123.3%	85.1%
Oberösterreich	96.1%	102.0%	102.0%	115.7%	150.8%	85.4%	84.2%
Salzburg							
Tirol					185.7%	112.8%	81.8%
Vorarlberg							

App. 1 - Une application sur un groupe hiérarchique administratif Chômage 15-24

Nombres casuels - distributions normales
(Moyenne et écart quadratique moyen des indices nationales), 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%)
Niederösterreich	102.8%	137.8%	103.9%	109.4%	156.9%	103.3%	91.5%
Wien	98.3%	118.6%	124.3%	128.7%	134.8%	125.2%	96.8%
Kärnten	115.4%	112.0%	98.9%	91.5%	137.5%	121.2%	85.8%
Steiermark	111.6%	100.0%	87.5%	104.8%	136.4%	123.3%	85.1%
Oberösterreich	96.1%	102.0%	102.0%	115.7%	150.8%	85.4%	84.2%
Salzburg	105.5%	141.5%	**118.3%**	117.2%	147.2%	104.7%	90.6%
Tirol	**111.9%**	87.6%	104.8%	110.4%	185.7%	112.8%	81.8%
Vorarlberg	106.7%	149.5%	83.3%	102.4%	134.7%	**119.4%**	89.9%

Interprétation des missing value: Imputation stochastique (en centaines), Autriche, 1999-06

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

App. 2 - Prévisions basées sur l'imputation des données manquantes

SAS Forecast Server - système d'information pour la sélection automatique des modèles **- produit automatiquement des prévisions statistiques pour toutes les séries d'enquêtes objet d'analyse.**

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.

Le modèle final utilisé pour la prévision d'une série régionale est sélectionnées sur la base des statistiques de bonté d'adaptation (par exemple MAPE).

App. 2 - Spécification des modèles sur bases de données traitées avec différentes méthodes d'imputation

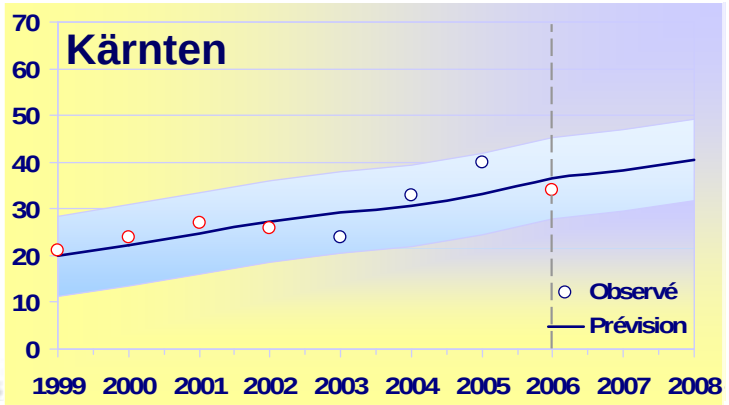
L'application, a le but de comparer les différences dans la sélection automatique des modèles à partir de la base de données observées.

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:

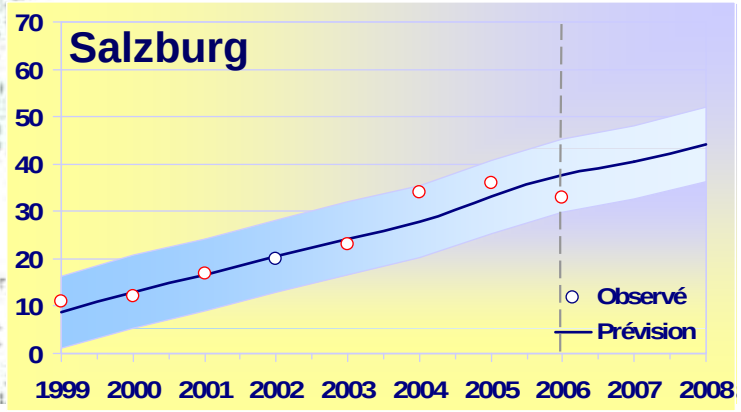
- sans **aucune méthode d'imputation**;
- données manquantes imputées par la **moyenne**;
- **imputation stochastique**.

App. 2 - Spécification des modèles sur bases de données traitées avec différentes méthodes d'imputation: Chômage 15-24 (en centaines), 1999-2008

Imputation: stochastique

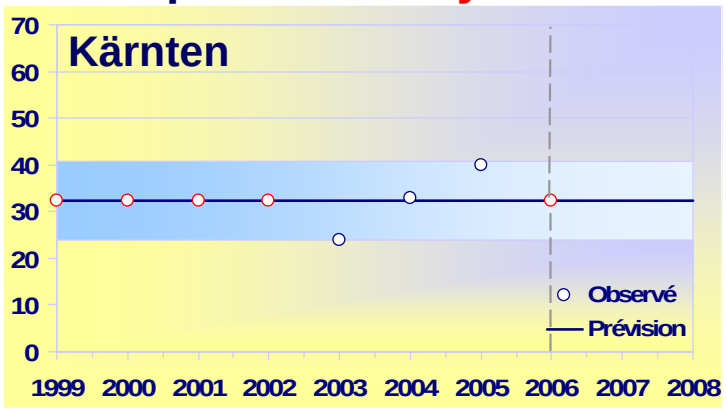


Modèle:Damped-Trend Linear Exponential Smoothing

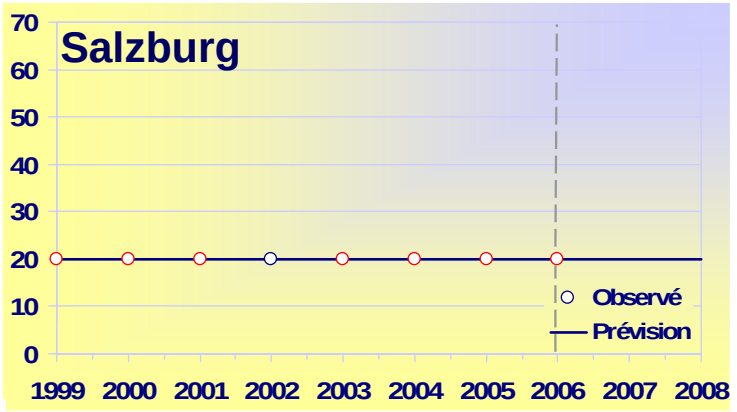


Modèle: Damped-Trend Linear Exponential Smoothing

Imputation: moyenne



Modèle: Simple Exponential Smoothing



Modèle: SMLIN

Source: Elaborations ESeC sur données Eurostat (Labour Force Survey).
Note: Le MAPE (≈10) est naturellement plus haut pour l'imputation stochastique.

App. 2 - Spécification des modèles sur bases de données traitées avec différentes méthodes d'imputation

Tout d'abord, on peut noter que **l'imputation stochastique permet de spécifier des modèles de réalité qui ne produisent pas des estimations indéfiniment constante**, 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 (évidemment) égale à zéro (comme dans le cas du chômage 15-24 ans dans les régions autrichiennes du Vorarlberg et de Salzbourg);

- **les bandes de confiance sont toujours représentables avec l'imputation stochastique**;

- **les modèles automatiquement sélectionnés sont non banal seulement avec l'imputation stochastique**. Emblématique, le cas du chômage (15-24 ans) de la région autrichienne de Kärnten, où la première et la dernière donnée 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.

App. 3 - Application sur un groupe homogène en termes de structure et de performances

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%**

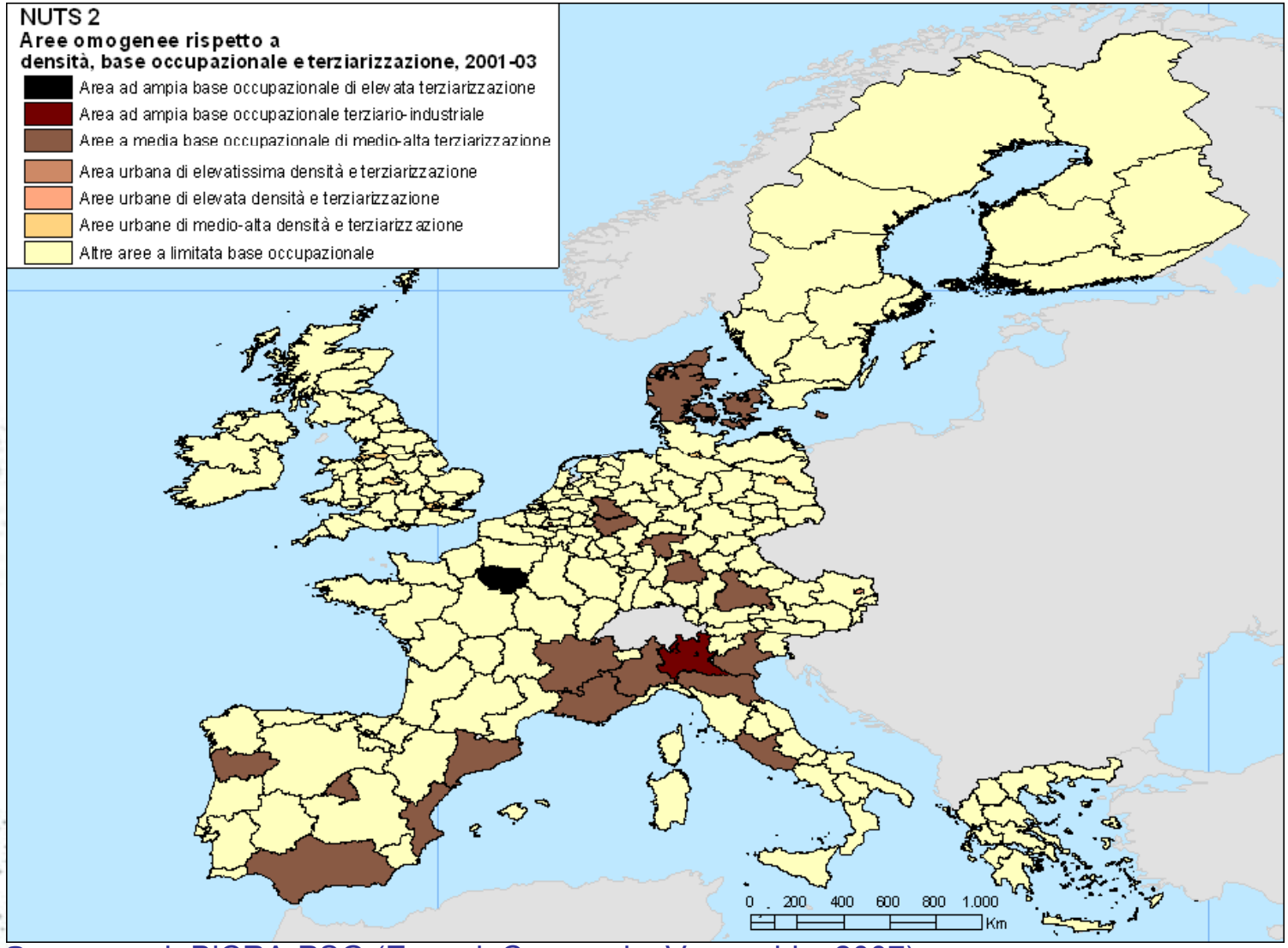
Source: voir BiCRA-PSG (Eusepi, Cepparulo, Verrecchia; 2007)

Régions à base occupational limité (classification basé 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%

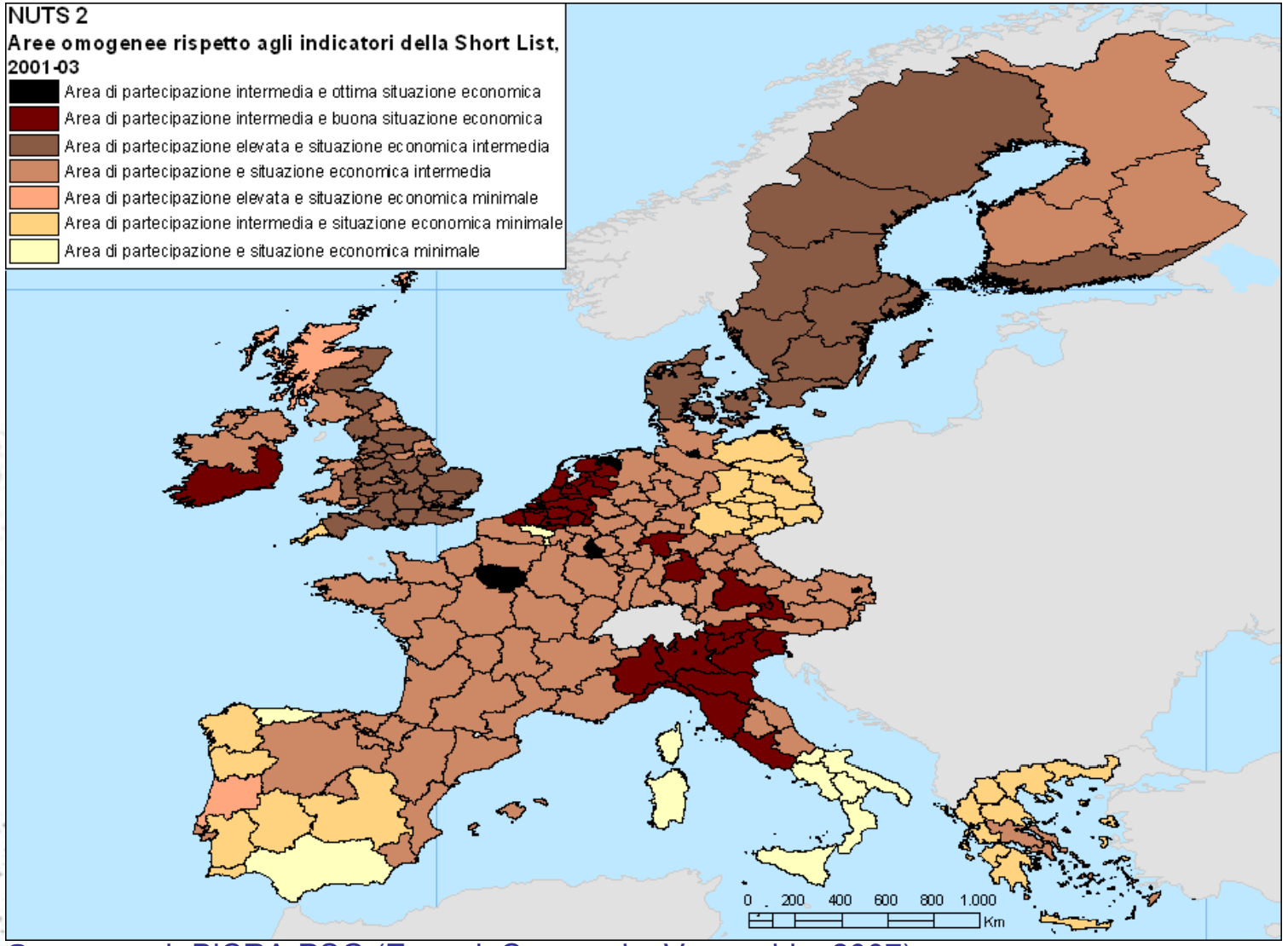
Source: voir BiCRA-PSG (Eusepi, Cepparulo, Verrecchia; 2007)

App. 3 - Application sur un groupe homogène en termes de structure et de performances: Structure



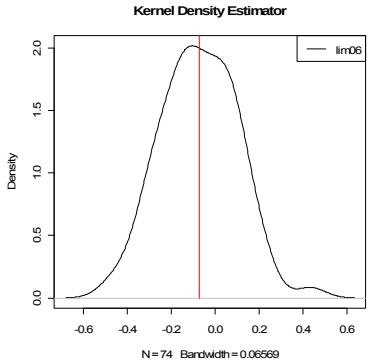
Source: voir BiCRA-PSG (Eusepi, Cepparulo, Verrecchia; 2007)

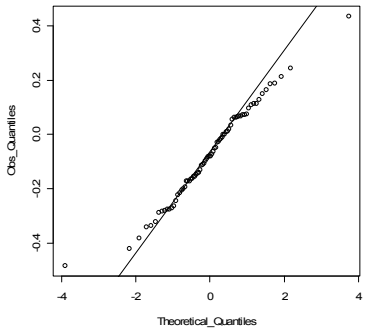
App. 3 - Application sur un groupe homogène en termes de structure et de performances: Performances



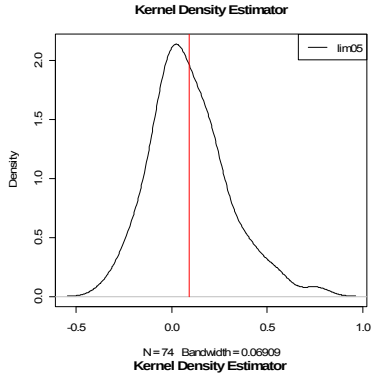
Source: voir BiCRA-PSG (Eusepi, Cepparulo, Verrecchia; 2007)

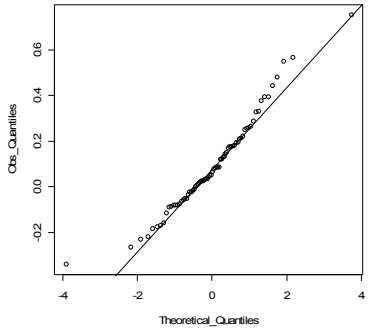
App. 3 - Distribution des logarithme du Chômage 15-24 pour les régions à participation et situation économique intermédiaire (KDE, QQplot, Skewness, Kurtosis), 2000-06



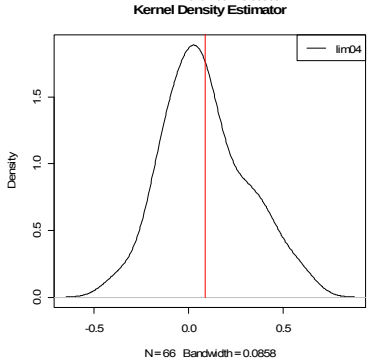


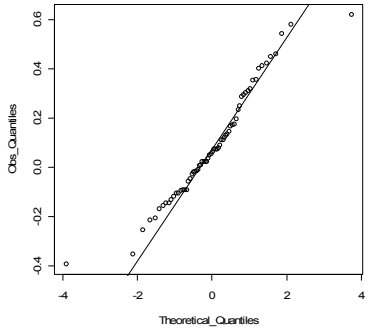
2006	
N:	74
Skewness:	0.101325
Kurtosis:	2.971209





2005	
N:	74
Skewness:	0.668547
Kurtosis:	3.747535





2004	
N:	66
Skewness:	0.374558
Kurtosis:	2.789564

App. 3 - Régions à participation et situation économique intermédiaire: Imputation stochastique - Chômage 15-24 (en centaines), 1999-08

Statistiques sur les indices, 2000-06

	1999	2000	2001	2002	2003	2004	2005	2006
Obs disponibles		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 quad. moyen		15%	17%	24%	21%	26%	25%	17%
Variance		0.0239	0.0300	0.0596	0.0425	0.0667	0.0614	0.0274

Nombres casuels - distributions normales (moyenne et écart quadratique moyen des indices)

		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%

Interpretation des missing value: Imputation stocastique

		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

Conclusions

L'application de l'imputation stochastique et de la ESeC-Rubin et la comparaisons avec les "vieilles méthodes" (considère le cas de la moyenne conditionnelle), montre les améliorations possibles - dans le cadre des séries économiques - résultant d'une **approche multidisciplinaire qui développe les propositions méthodologiques de la théorie:**

- **des échantillons**
- **des nombres indice**

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] Anselmi A., Chiodini P.M., Verrecchia F., «ESeC-Rubin Missing Value Interpretation for a Regional Bottom-Up Hierarchical Forecasting», ESeC Working Paper

[2] 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.

[3] 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

[4] 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.

[5] Eusepi G., Cepparulo A., Verrecchia F., «Bilevel Comparative Regional Analysis - Performances in Structural Grid.», ESeC. Working paper ESeC_WP001

[6] Herzog T.N. and Rubin D.B., «Using multiple imputation to handle nonresponse in sample surveys.» In Incomplete Data in Sample Surveys, Vol. 2

[7] Little R.J.A., Rubin D.B., «Statistical analysis with missing data»,

[8] Martini M., «Numeri indice per il confronto nel tempo e nello spazio»,

[9] Rubin D.B., «Multiple imputation for Nonresponse in Surveys»

[10] Rubin D.B., «Multiple imputation after 18+ year». *Journal of the American Statistical Association*, June 1996, Vol. 91, n°. 434

[11] 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

[12] SAS Institute Inc., «SAS Forecast Server 1.4: Administrator's Guide»

[13] Schafer J.L. and Graham J.W., «Missing Data: Our View of the State of the Art», *Psychological Methods*, 2002, Vol. 7, n° 2

[14] Verrecchia F.,«The Generalised Index Numbers», *Journal of ESeC Short Papers*, 2008, Vol.1, n°1

[15] 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 (26-29.8.2008).

Données manquantes et prévisions: Méthodes à imputation variable

Antonio ANSELMI
SAS Institute, Customer Support

Paola M. Chiodini
Université of Milano-Bicocca

Flavio Verrecchia
ESeC

JMS - Journées de Méthodologie Statistique
Paris, le 24 mars 2009