UNIVERSITÀ DELLA CALABRIA

Dipartimento di Economia e Statistica
Ponte Pietro Bucci, Cubo 0/C
87036 Arcavacata di Rende (Cosenza)
Italy
http://www.ecostat.unical.it/

Working Paper n. 12 - 2011

CRIME AND ECONOMIC PERFORMANCE. A CLUSTER ANALYSIS OF PANEL DATA ON ITALY'S NUTS 3 REGIONS

Rosetta Lombardo
Dipartimento di Economia e Statistica
Università della Calabria
Ponte Pietro Bucci, Cubo 0/C
Tel.: +39 0984 492466
Fax: +39 0984 492421
e-mail: r.lombardo@unical.it

Marianna Falcone

e-mail: maryfalcon@libero.it

Ottobre 2011

Crime and Economic Performance. A cluster analysis of panel data on Italy's NUTS 3 regions

Rosetta Lombardo[a,*], Marianna Falcone

[a] *Dipartimento di Economia e Statistica - Università della Calabria*
Via Pietro Bucci, cubo 0C, 87036 Rende (CS) - Italy

Abstract

Crime is a complex phenomenon which needs to be investigated at appropriate disaggregate temporal and territorial levels of analysis. The specific issue addressed in this paper concerns the possibility of classifying Italian provinces in groups by using a new methodology that combines cluster analysis with panel time series data regarding a wide range of crime indicators, economic performance indicators and other socio-demographic variables. Our main contribution is to show that crime is not inextricably linked to geographic location as is usually believed. From this point of view, the position of the Italian Mezzogiorno has two facets; one is found in those relatively not affluent provinces which are resistant to random street and organized crime while the other facet is encountered in those provinces which are caught in a vicious circle where increasing criminal activity and weak economic performance feed on each other to undermine the security of the population. The pattern for the Center-North of Italy is much more varied and composed of a series of four clusters. In particular, there is a relatively large cluster of provinces which should be considered at risk because of an increasing value added per capita, high rates of population replacement and female employment; all conditions that might attract and encourage criminal activity.

JEL classification codes: C32, E27, D73, K42, O1, R11.

Keywords: Law-breaking behavior, Local economic development, Socio-economic factors, Cluster analysis, Time-series-cross-section data.

*Corresponding author.
Email addresses: r.lombardo@unical.it (R. Lombardo); maryfalcon@libero.it (M. Falcone).

1. Introduction

Interest in an economic analysis of crime was inspired by Becker's, (1968) seminal contribution. Becker's model, extended by Ehrlich, (1973), postulates that individuals rationally decide whether to engage in criminal activities by comparing the expected returns to crime with the returns to legal work. Since then, a large amount of empirical literature (among others, Fox, (1978); Craig, (1987); Avio, (1988)) has developed around the estimation and testing of economic model of crime with aggregate data which rely heavily on cross-section time series (TSCS or panel) techniques (see Cornwell & Trumbull, (1994)). While economic literature has paid great attention to territorial disparities of economic development, it has not focused enough on local differences in crime rates in relationship to differences in socio-economic conditions which point to the existence of clusters in the data (see, *e.g.* Hsieh & Pugh, (1993); Anselin *et al.*, (2000)). A number of empirical studies have focused on the link between crime and economic performance using regression methods to analyze administrative units within a pooled data base and for summarizing relationships between indicators. As Wilson & Butler, (2007) pointed out, regression results with panel or TSCS data can be exceedingly frail and many data sets simply have too many limitations to use in a reliable fashion regardless of the estimation method employed.

The aim of our study is not to build a regression model which relates a response to a set of covariates to examine, for example, the explanatory power of one or more crime offenses on some indicators of economic performance or the impact of deterrence policies or labor-market outcomes on criminal behavior. Rather, our main objective is to find out an effective method for utilizing the combined strength of spatial and temporal dimension of a set of selected indicators to classify the administrative units into two or more groups. More specifically, this paper addresses two issues. The first concerns the classification of a set of administrative units into clusters on the basis of various indicators that are only available for a limited period of time which can be different for different indicators and different units (unbalanced TSCS data). In particular, we apply the cluster analysis of multivariate short time series to contribute to a better understanding of the interactions between criminality, economic performance and geographical location. The second issue this paper addresses, by modeling several related time series together, is the detection of hidden sequence features and development trends also useful for the setting up of potential control schemes of crime. This is important

because certain patterns of crime in time and space may generate inefficiency and slow down local economic development.

The main results of our analysis can be summarized as follows. The Italian provinces are classified in seven distinct clusters which show that crime in Italy is not inextricably linked to geographic location as is usually believed; in fact, one of the most interesting clusters, named "provinces at risk" is distributed all over the Center-North of Italy. From this point of view, the position of the Italian Mezzogiorno has two facets; one is found in those relatively not affluent provinces which are resistant to random street and organized crime while the other facet is encountered in those provinces which are caught in a vicious circle where increasing criminal activity and weak economic performance feed on each other to undermine the security of the population. The pattern for the Center-North of Italy is much more varied and composed of a series of four clusters. In particular, there is a relatively large cluster of provinces which should be considered at risk because of an increasing value added per capita, high rates of population replacement and female employment; all conditions that might encourage criminal activity.

TSCS data calls for modeling heterogeneity due to serial dependence between the observations on each unit over time, and spatial dependence between the observations on the units at each point in time. Our methodological approach has the potential to detect the fact that multivariate short time series arise from very different groups of units and that an econometric model based on the same parameters should be estimated within each group. In practice, we add a completely new technique to the "methodological tool bag" that researchers have at their disposal to analyze TSCS data.

The rest of the paper is structured as follows. In section 2, we discuss data and indicators and propose definitions to ascertain entrenched patterns of economic performance and crime by employing Italian TSCS data at a NUTS 3 (or province) level of disaggregation. Our statistical methodology is described in section 3 where we attempt to gather advantages and cancel out disadvantages of both cluster analysis and cross-sectional time series econometrics. The bridge between the two approaches is a new metric which has proved to be very effective for multivariate short time series. Section 4 presents the basic investigation. Here, we carry out statistical searches for trends and groups by using 24 indicators observed in the 103 Italian provinces. In particular, we analyze an unbalanced TSCS data set made up of yearly observations. Conclusions and directions for future research are then presented in section 5.

2. Conceptual framework for the choice of units and indicators

Criminal activity is a phenomenon associated with money and the acceptance of law-breaking behavior. Dutta & Husain, (2009) point out that crime has serious consequences for the states ability to promote development. In fact, an increase in government expenditure on security crowds out some key investments in infrastructure. Crime discourages entrepreneurial behavior, hinders opportunities for employment and education and has negative consequences for social and psychological aspects of life. It hinders the accumulation of assets. Most signifiant increases in crime rates, experienced in many regions of Italy, regard property-related crimes (Travaglini, (2003)), and this has to do with growth. The empirical evidence shows that crime depresses investments (Pellegrini & Gerlagh, (2004)); growth and tourism (Peri, (2004); Crdenas & Rozo, (2008); Gaibulloev & Sandler, (2008)) while increasing inflation (Al-Marhub, (2000)). Illegal activity and, in particular, corruption diverts resources in the graft-rich environment of public works projects, at a cost to other more valuable social intervention and prevention programs. Law-breaking behavior can have the effect of lowering tax revenues and of pushing poor into the shadow economy. Offenses against the state threaten to erode peoples confidence in governmental institutions. Moreover, crime damages the country's image abroad since investors see it as a sign of instability and apparent chaos (Daniele & Marani, (2011)). These impacts are key to explaining the problems experienced in Italy in terms of economic growth and national well-being.

Organized crime is even more damaging. Perhaps, it is the single greatest obstacle to development not only in the South, but throughout the whole of Italy. Wherever it occurs, organized crime distorts markets. In some cases, for instance, industries bear only part of the cost of legally disposing of their toxic waste by hiring organized crime to manage its disposal clandestinely. Criminal syndicates can lower agricultural and food prices by helping certain business owners to reduce production costs trough the employment of illegal foreign workers, the use of doctored scales or the adulteration of food. Organized crime decreases real estate values, by forcing property owners to sell at ridiculously low prices through intimidation or violence. In general, its activities lead to higher costs for the government, honest business owners and consumers. In addition to extracting money from others, organized crime engages in its own enterprise when it comes to public contracts (especially in

the construction industry). Through a system of coordinated bid withdrawals and a programmed rotation of winners, all of the companies controlled by the syndicates are guaranteed contracts while offering only a minimal discount. This type of contract becomes even more lucrative if the base price of tender is overestimated. The additional profits allow the contract winners to deliver larger bribes to both the corrupt politicians and public officials who organize the bidding.

The empirical research regarding the relationship between economic performance and criminal activity in Italy has so far yielded conflicting results. In fact, although the affirmation that crime negatively influences economic performance finds a wide consensus in public opinion, quantitative research into the Italian case is relatively limited. Crime in Italy is described by some stylized facts: high spatial and time variability of crime activities and the persuasion that organized crime is localized in specific areas, although the localization radius is gradually increasing. Peri, (2004) analyses the effect of crime on long-term economic growth by examining the roles of several variables, among which there is a proxy of social capital. The results demonstrate how crime has had a notable influence on regional development. Buonanno, (2006), using regional data, investigates the relationship between labor market conditions and crime in Italy, accounting for both age and gender in the unemployment measure. His results suggest that unemployment has a large and positive effect on crime rates in southern regions. Mauro & Carmeci, (2007) explore the link between crime, unemployment and economic growth using Italian regional data. The empirical results suggest that crime and unemployment have long-run income level effects. Buonanno & Leonida, (2008) use a panel dataset for the Italian regions over the period 19801995 in order to study whether education exerts a non-market effect on crime rate. Their empirical results indicate that education exerts a negative significant effect on crime rates after controlling for socioeconomic and deterrence factors. Both the share of population with a high school diploma and the average years of schooling of the population are negatively and significantly correlated to every classification of crime rate. Moreover, they find that crime rate depend more on high school graduation rate than college graduation rate. Cracolici & Uberti, (2009) try to capture the differences among types of crimes and across space over a period of two years, at the provincial level. The use of exploratory spatial data analysis allows the authors to detect some important geographical dimensions and to distinguish micro- and macro-territorial aspects of offenses. Detotto & Otranto, (2010) propose a state space model

to analyze the effects of crime on economic growth. They apply the model to the Italian case, using a large data set with monthly frequency to measure the crowding out effect of crime. Bianchi *et al.*, (2010) investigate the relationship between immigration and crime across Italian provinces over the period $1990 - 2003$. They disaggregate reported crimes by type of criminal offense (violent crime, property crime and drug-related crime) and use instrumental variables based on immigration to other countries than Italy to identify the causal impact of exogenous changes in Italy's immigrant population. According to the estimates, immigration only increases the incidence of robberies, while leaving all other types of crime unaffected. Calderoni, (2011), proposes an index measuring the presence of organized crime at the provincial level over the period $1983 - 2009$. The index highlights the concentration of organized crime in its original territories, but also a significant presence in northern and central provinces.

2.1. Data structure

Crime is a very complex phenomenon that needs to be investigated at a suitable disaggregate "territorial" and "sectorial" levels of analysis (*i.e.* administrative units and different types of offenses) to catch the most important local disparities. Hence, suitable tools of analysis should be used to control for the two dimensions of the problem. Our approach involves multivariate time series on socio-economic indicators which are typically collected at a national or sub national level, observed at different points in time, arranged by chronological order and analyzed in tandem (TSCS data). See, on this point, Baltagi, (2006b).

Data can be arrayed with the m territorial units as rows; each unit is described by a fixed set of p indicators concerning various aspects of the economic performance/crime problem. Similar data will exist for the time points $t, t = b_{r,j}, b_{r,j} + 1, b_{r,j} + 2, \ldots, c_{r,j}$. The time intervals $[b_{r,j}, c_{r,j}], r = 1, 2, \ldots, m; j = 1, 2, \ldots, p$ of the indicators do not need to be of equal length and the time spans over which the territorial units are observed may differ for starting and/or ending points, The data are organized into the following three-way array

$$\mathbf{X}_j = \begin{bmatrix} x_{1,j,b_{1,j}} & x_{1,j,b_{1,j}+1} & x_{1,j,b_{1,j}+2} & \cdots & x_{1,j,c_{1,j}} \\ x_{2,j,b_{2,j}} & x_{2,j,b_{2,j}+1} & x_{2,j,b_{2,j}+2} & \cdots & x_{2,j,c_{2,j}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{r,j,b_{r,j}} & x_{r,j,b_{r,j}+1} & x_{r,j,b_{r,j}+2} & \cdots & x_{r,j,c_{r,j}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m,j,b_{m,j}} & x_{m,j,b_{m,j}+1} & x_{m,j,b_{m,j}+2} & \cdots & x_{m,j,c_{m,j}} \end{bmatrix} \quad j = 1, 2, \ldots, p. \qquad (1)$$

In practice, each slide $\mathbf{X}_j$ of (1) is a ragged array, which can be thought of as a rectangular array, some positions of which have been omitted. The unbalanced TSCS structure (1) allows us to have different numbers of observations for different units thus incorporating more information than if we had to restrict ourselves to a balanced TSCS in which all the units had the same number of time points and the same starts and ends. The basic difference, however, is that the unbalanced TSCS case has indicators that are crucially dependent on the length of time-series available for each territorial unit.

2.2. Selection of the units

The units considered in this study are the 103 NUTS 3 (European Nomenclature of Territorial Units for Statistics) regions in Italy. This administrative unit is an ideal unit of analysis for a number of reasons. Mid-level administrative areas, such as provinces, are important because the boundaries of politically defined sub-national regions are often derived from geography and industrial history. In Italy, a number of institutions, from the Prefecture to the Chamber of Commerce and the Province, reflect this level of subdivision and these institutions play important coordinating and supervisory roles which may influence both economic performance and the presence of crime, be it organized or not. In fact, in Italy, law enforcement agencies are frequently organized on a provincial basis. The provincial level offers greater detail than the region and permits identification of different patterns within the Italian regions, particularly within those without a traditional presence of organized crime (Calderoni, (2011)). Moreover, provinces allow us to capture the nature of crime better given that criminal activities are related to a specific area where illegal cartels exert great pressure and/or to a period of time in which adverse economic conditions may increase the probability of crime being committed. Cornwell & Trumbull, (1994) observe that the data on crime and related variables should be acquired at the finest possible

resolution. The finer the disaggregation of the data, the more precisely local government interventions and national economic policies can be studied, implemented and evaluated in relation to the diversity among the various administrative units. On the other hand, it should not be ignored that if the units are small, a substantial amount of crime is often committed by non-resident offenders (Büttner & Spengler, (2003); Edmark, (2005)).

2.3. Selection of the indicators

The set of indicators has been chosen so as to portray the clusters of provinces which are as compact and separate as possible, hoping to find that the unknown but existing groups all cluster around the same set of attributes (Friedman & Meulman, (2004)). The set of pertinent indicators has to be rich enough to encompass all important phenomena while also varied enough to avoid bias. This decreases the risk that neglected factors will distort the results and permits (at least in principle) a detailed explanation of several relationships connecting socio-economic conditions and criminal activity. On the other hand, the "curse of dimensionality" is lurking in the background: too many indicators are not always a good thing because adding more features may increase the noise.

The challenge we have to face is to strike a balance between generality and parsimony to provide a manageable amount of information without overly simplifying the issues. Entorf & Spengler [2000, 2002] considered low economic status, family disruption and urbanization, as important factors of crime. They study the effect of some educational, demographic and economic variables to explain the crime differences between European regions and conclude that socio-economic indicators are important for the understanding of crime in Europe. Akomak & ter Wee, (2008) investigate the link between social capital and crime and attempt to explain why crime is so heterogeneous across space and to what extent levels of social capital can be associated with crime rates. Scorcu & Cellini, (1998) try to figure out which economic variable, from among consumption, wealth, and unemployment, is most closely related to crimes in the long run. According to Edmark, (2005), the proportion of divorced people, of the population with higher education, of foreign citizens, of young men and sales of alcohol, can be considered as factors explaining crime patterns.

In summary, we derive a set of 24 indicators from a comprehensive review of recently published literature. They are organized into four macrofactors:

crime related variables, demographic indicators, economic performance indicators and educational indicators.

2.3.1. Crime related indicators

We consider the following seven different types of crime: drug related crimes; extortion; burglary and larceny; prostitution related crimes; robbery; rape, sexual assault and kidnapping; homicide. We also include the reported crime rate, the foreigner crime rate and the juvenile crime rate.

2.3.2. Demography

Previous research indicates that certain variables have effects on crime that should be taken into account. Demographic indicators are central to many theoretical perspectives and empirical models of criminal behavior. South & Messner, (2000) reviewed studies that explore the multiple linkages and reciprocal relationships between criminal and demographic behavior, focusing on the intersection of criminal and demographic events over the course of a life.

To test the relationship between crime patterns and age structure in the population, we include the following indicators in our analysis: young-age dependency ratio (actual population under the age of 15 divided by the number of individuals of working age $15-64$); population replacement rate (resident population aged $0-14$/resident population aged 65 and over); population replacement rate in active age (resident population aged $15-19$/resident population aged $60-64$).

Among the demographic factors considered, in a number of works, as potential influences on crime which have a significant impact on the growth rate of per-capita output, and hence, on economic performance, we use the average male age at first marriage weighted with specific marriage rates; the average age of women at childbearing calculated on live births; and the number of divorces for 100 marriages. Population density is also taken into account. Its role in the generation or suppression of crime has been the subject of debate for decades. The classic argument is that high density offers opportunities for property crime given that it is a surrogate for the distribution of private property, much of which presents attractive targets for thieves. On the other hand, densely populated areas offer natural surveillance that has the effect of inhibiting violent crime in so far as witnesses are more

abundant and events are more likely to be reported to the police (see Harries, (2008)).

2.3.3. Economic performance

The assessment of economic performance is a demanding and multifaceted endeavour. Besides statistics, it involves a variety of fields belonging to the realm of social sciences. Virtually all studies rely on the unemployment rate as an indicator of economic conditions. Among others, Cantor & Land, (1985) argue that the unemployment rate is meant to proxy the more general health of the economy. Gould *et al*, (2002) state that changes in crime rates can be explained by changes in labor market opportunities for those most likely to commit crime, i.e. young, unskilled men. Raphael & Winter-Ebmer, (2001) point out that, all else being equal, the decrease in potential earnings associated with involuntary unemployment, increases the relative returns on illegal activity. Moreover, workers that experience chronic unemployment have less to lose in the event of arrest and incarceration.

The overwhelming majority of theoretical arguments suggest a strong positive relationship between unemployment and crime. It must be noted, however, that empirical research has not always been able to document a causal effect of unemployment on crime. For instance, Witte & Witt, (2001) observe that young people are more likely to participate in crime long before they participate in the labor market. Existing research suggests that higher unemployment is associated with a greater occurrence of property crime, but this relationship turns out to be insignificant for violent crime (Entorf & Spengler, (2000)). In addition, it is hard to know whether a rise in unemployment rates is causing a rise in crime rates, or is merely a symptom of something else that is the true cause. We can be relatively certain that higher unemployment rates are associated with a greater occurrence of crime, but cannot exclude the reverse causality.

A number of studies (see for example Kapuscinsk *et al.*, (1998) or Hansen, (2006)), find that rising female employment, which is generally thought of as a labor market improvement, is actually positively associated with crime. In essence, the theory proposes that increased entry of women into the labor market increases the overall supply of workers and, thus, subsequently lowers wages. As low wages and crime are generally found to be related, by lowering wages, a rising female employment may increase crime. Moreover, because women tend to have less labor market experience than men, or because they

are discriminated against, their entry into the labor market at a lower point in the earnings distribution puts downward pressure on the wages of males in lower skilled jobs, *i.e.* of those who are more likely to be on the margins of crime.

Saridakis, (2004) observes that female employment can also be considered as a proxy of the changes that have occurred in the families of Western Countries. An increase in female employment means women spend less time at home which, in turn, contributes to lower parental supervision of children and thus, could be associated with an increase in the crime rate. We consider three indicators of economic performance: the total unemployment rate, female employment rate and male youth unemployment rate.

Following Ehrlich, (1973) , we could have considered gross-domestic product (GDP) as a proxy for the general level of prosperity and, thus, as an indicator of illegal income opportunities. GDP is reliable when measured at a national or regional level. Hence, as an indicator of economic performance, we use the value added per capita which is a more reliable aggregate of national account referred at a provincial level.

2.3.4. Education

It is a statistical fact that criminal activity is inversely related to perpetrator education level. Machin *et al.*, (2010) find that criminal activity is negatively associated with higher levels of education and show that improving education may yield significant social benefits and may be a key policy tool in the drive to reduce crime. There are a number of theoretical reasons why education might have an effect on crime. Higher levels of educational attainment raise skills and abilities, imply greater productivity of labor and, therefore, are associated with higher expected returns from legal work (see Forni & Paba, (2000)). Education raises the opportunity cost of time spent engaged in criminal activities and increases the cost associated with incarceration (see Lochner, (2004); Lochner, (2010)).

It is a potentially large influence on individual propensities to offend and, possibly, an important cause of area-level variation in crime rates. Crime statistics indicate that crime rates tend to be lower in areas with higher levels of education, and that these are also areas of higher per capita income and contain a higher proportion of families belonging to the highest socio-economic status. Lochner & Moretti, (2004) make a strong case for increasing high school graduation rates as an alternative to increasing the size of police

forces. Despite promising evidence that education-based policies and early childhood intervention may play an important role in helping reduce crime, evidence is still limited and sometimes mixed. In fact, whether the link between education and crime is causal, or whether it masks a number of possible effects that may not be due to education, is less clear. For example, education may also develop criminal skills; although, this is only likely to be important for certain white collar crimes. Santas, (2007)) shows that there is a connection between growing high school dropout rates and increasing juvenile crime rates. The chance of an uneducated person who is involved in criminal activity being successful is very low. The chances are greater that these dropouts will end up on public assistance or in prison, so costing society money instead of contributing. In fact, a high school diploma is the bare minimum credential necessary to have a reasonable chance in the workforce.

We consider two indicators of education: the share of students enrolled in university courses and the share of early school leavers.

The final selection of indicators used in this study is described in the Table 1.

The source of data used in this paper is the Italian National Institute of Statistics (Istat, (2011)).

3. Statistical Methodology

We intend to build a classification of provinces useful for researchers and policy makers attempt to answer questions concerning efforts at crime control. Given a TSCS, the following scheme is adopted: computation of the distance between two univariate time series. In doing so, an adequate metric is required. The choice of this metric should capture the discrepancies between time series when considering both the observed values and the temporal nature of the indicators. In this paper, we develop a simple, but very effective computational technique to compare relatively short and multivariate time series or sequences. Based on the quantification of the distance between two time series, the same computation is performed for all possible pairs of provinces and for all indicators used in the analysis. Subsequently, the indicator-specific distance matrices are merged into a global distance matrix, which will be used for the cluster analysis of the provinces.

Table 1: Indicators and time intervals.

Class	Code	Variable	Years
Crime			
	IC_1	Reported crime rate	$1999-2008$
	IC_2	Juvenile crime rate	$1999-2007$
	IC_3	Foreigner crime rate	$1999-2005$
	IC_4	Unknown perpetrators crime rate	$1999-2008$
	IC_5	Drug related crimes	$1999-2008$
	IC_6	Extorsion	$1999-2008$
	IC_7	Burglary and larceny	$1999-2008$
	IC_8	Prostitution related crimes	$1999-2008$
	IC_9	Robbery	$1999-2008$
	IC_{10}	Rape, sexual assault and kidnapping	$1999-2008$
	IC_{11}	Homicide	$1999-2008$
Demography			
	ID_1	Mean age of men at the time of first marriage	$1999-2008$
	ID_2	Mean age of women at childbirth	$1999-2008$
	ID_3	Population density	$1999-2009$
	ID_4	Divorces per 100 marriages	$1999-2007$
	ID_5	Population replacement rate	$1999-2008$
	ID_6	Population replacement rate in active age	$1999-2008$
	ID_7	Young-age dependency ratio	$1999-2008$
Econ. Performance			
	IL_1	Unemployment rate	$1995-2009$
	IL_2	Male unemployment rate, $15-24$	$1995-2009$
	IL_3	Female employment rate	$2000-2009$
	IL_4	Value added per capita	$1995-2008$
Education			
	IE_1	Share of students enrolled in university courses	$1999-2009$
	IE_3	Early school leavers	$2004-2009$

3.1. Distance between multivariate short time series

The distance function that quantifies the proximity between objects (in our case, multiple short-run time series) is crucial in determining the outcome of any cluster analysis algorithm. When dealing with many time series of a short duration, it is more important to assign a value to the distance between two observed sequences rather than to measure the strength of relationships between the stochastic processes that generate observations. Consequently, techniques such as autoregressive and spectral representations, dimensionality reduction, compression, decomposition or subsequence matching methods are not useful. On the contrary, in case of short length sequences, it is essential to make the most efficient use of the few observations that are available.

Consider two univariate time series concerning the j-th indicator observed in two different territorial units $X_{r,j}, X_{s,j}$ at the time points, respectively, $b_{r,j}, b_{r,j}+1, b_{r,j}+2, \ldots, c_{r,j}$ and $b_{s,j}, b_{s,j}+1, b_{s,j}+2, \ldots, c_{s,j}$. Let $e_{r,s}^{(j)}, \ldots, f_{r,s}^{(j)}$ be the time interval in which the two time series overlap with $e_{r,s}^{(j)} \geq \max\{b_{r,j}, b_{s,j}\}, f_{r,s}^{(j)} \leq \min\{c_{r,j}, c_{s,j}\}$ The distance between two time series that we wish to evaluate has the form:

$$A_{r,s}^{(j)} = \min_{e_{r,s}^{(j)} \leq L_{r,s}^{(j)} \leq f_{r,s}^{(j)}-1} \left\{ \left[\frac{\sum_{t=e_{r,s}^{(j)}}^{f_{r,s}^{(j)}} |x_{r,j,t} - x_{s,j,t'}|^q}{f_{r,s}^{(j)} - e_{r,s}^{(j)} + 1} \right]^{1/q} \right\} \quad \text{with } q \geq 1 \qquad (2)$$

where $t' = \left(t + L_{r,s}^{(j)}\right) \bmod f_{r,s}^{(j)}$. We assume that $f_{r,s}^{(j)} - e_{r,s}^{(j)} > 2$ for all r, s and j which implies that there are at least three time points at which one time series overlaps with the other. If the starting point of one of the two time series, for example, $X_{s,j}$, were allowed to change, while keeping the same sequence of values of both time series, then (2) would determine the alignment that makes $X_{r,j}, X_{s,j}$ nearer in terms of the Minkowski norm of order q and $A_{r,s}^{(j)}$ is the total amount of distance that, in this case, a point travels between leaving $X_{r,j}$ and reaching $X_{s,j}$. Raveh, (1981)) interpreted (2) as the minimum absolute mean differences between the time series r and s where these differences are computed for the various lags. For each $q \geq 1$, expression in (2) is a metric. We choose $q = 1$ becauses it does not imply a heavier computational burden.

For greater generality, we have admitted the possibility that the time spans can have differing starting and ending points for both different indicators and different provinces. The ability of any measurement of distance to distinguish between objects decreases with the reduction of the number of comparisons. Hence, equation (2) should be modified to give less weight to the time of non-overlap in order to distinguish between identical time series and partially identical time series (*e.g.* identical after start of overlap). This modified version of the distance function is defined in (3) below

$$\delta_{r,s}^{(j)} = \left[1 + \left(\nu_j - n_{r,s}^{(j)}\right)\right] A_{r,s}^{(j)} \qquad (3)$$

where ν_j is the maximum duration of time observed for the j-indicator. The rationale behind (3) is that $A_{r,s}^{(j)}$ is an average over the overlapping time period, so that we can impute the "missing" values in the computation of the

distance between $X_{r,j}$ and $X_{s,j}$ by repeating $A_{r,s}^{(j)}$ once for each time period in which either time series existed but did not overlap with the other. Although there is no specific reason to assume that the unobserved distance between the time series r and s is of a linear type (see, Sinha & Mark, (2005)) for an alternative approach), the choice in (3) has the desirable effect of limiting the formation of clusters which include units that share too few time points. It should be noted that, if $A_{r,s}^{(j)}$ is a metric then $\alpha A_{r,s}^{(j)}, \alpha > 0$ is also a metric (see, for example, Batagelj & Bren, (1995)).

For reason of comparison, it is desirable to normalize distances so that they vary between zero and one:

$$d_{r,s}^{(j)} = \frac{\delta_{r,s}^{(j)}}{\max_{r,s=1,2,\cdots m} \delta_{r,s}^{(j)}}, \quad j = 1, 2, \cdots, p. \tag{4}$$

The $(m \times m)$ matrix of normalized distances $\mathbf{D}_j$ is obtained by computing the coefficient $d_{r,s}^{(j)}$ for all the $m\,(m-1)\,/2$ pairs of units. The matrix $\mathbf{D}_j$ is an indicator-specific distance matrix. Since different indicators could potentially generate a different clustering, a key issue is to decide how the indicator-specific distance matrices should be combined to produce a global measure of distance. We use the linear combination:

$$\mathbf{D} = (d_{r,s}) = \sum_{j=1}^{p} w_j d_{r,s}^{(j)} \quad \text{with} \quad w_j \geq 0, \quad \text{and} \quad \sum_{j=1}^{m} w_j = 1. \tag{5}$$

where $w_1, w_2, \ldots, w_p$ is a system of weights that assists the congruence between clusterings of the same regions based on different sources of information. In the light of (5), condition (4) appears to be necessary otherwise the distances associated with the various indicators would copy the structure of the indicator with the largest distances.

An optimal system of weights for (5) can be obtained by using an expression for how much a certain indicator affects the global distance. This can be derived from the total sum of the squares of the elements in the indicator-specific distance matrices $\mathbf{D}_j, j = 1, 2, \ldots, p$ choosing the weights so as to maximize the variance of the elements in the global distance matrix $\mathbf{D}$ which, in turn, leads to the Distatis procedure, a generalization of classical multidimensional scaling, developed by Abdi *et al.*, (2005). The weights $\mathbf{w} = (w_1, w_2, \ldots, w_p)$ are such that the patterns generated by variables which are most similar to the others get higher weights than those which disagree with most of the rest of the variables.

Distatis requires that $\mathbf{D}_j$s are Euclidean distance matrices. The construction process for the metric (3) gives no guarantee for the euclidity of the indicator-specific matrices, even though each single entry of these matrices is the value of a metric. But, just in case $\mathbf{D}_j$ is not Euclidean, it is possible to determine a constant $\phi_{1,j}$ such that a matrix with elements $d_{r,s}^{(j)} + \phi_{1,j}$ $r \neq s$ is Euclidean. (See Gower & Legendre, (1986)[theorem 7]). Pavoine *et al.*, (2009) show that if the $\mathbf{D}_j$s are Euclidean, then $\mathbf{D}$ is also Euclidean.

3.2. Clustering

Most techniques for cluster analysis are designed for items described by a vector of features whereas applications to time-varying data require clustering methods which are determined by distances. In this paper, we use the partitioning around medoids method (PAM) proposed by Kaufman & Rousseeuw, (1990) because the computed matrix of global distances, whatever the outcome, can be fed into the algorithm in order to form clusters of similar units.

The PAM algorithm requires the number of clusters, k, be known a priori. Initially randomly k medoids are chosen from a set of m provinces. Each remaining province is clustered with the medoid to which it is most similar on the basis of the distance between province and medoid. The medoids are recomputed as the most centrally located province in the cluster; then, in each step, an exchange of a selected province and a non-selected province is made, as long as such an exchange results in an improvement of the quality of the solution. This quality is measured by using a loss function that measures the average dissimilarity in all the clusters. The process is iterated until the re-computation of medoids gives no further improvement. The algorithm terminates after a finite number of iterations.

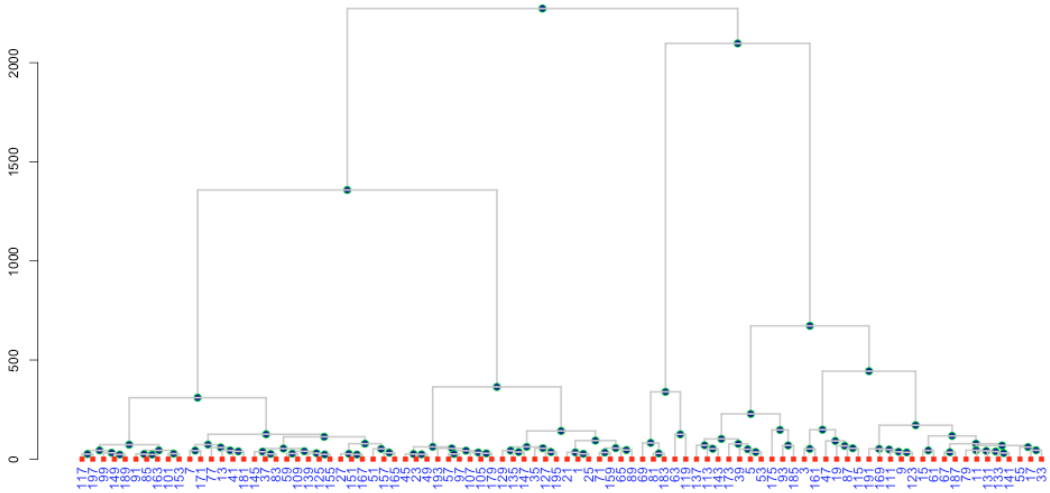
3.2.1. An illustrative application

Clustering of TSCS data can also be seen simply as a means rather than an end in itself. Despite the vast literature on cross sectional data analysis, relatively little attention has been paid to the similarities of territorial units and to their spatial contiguity (see Patacchini, (2008) and reference therein). When dealing with short-run dynamics, it is usual to pool all time series data and estimate a common regression model or a global time series model instead of performing the usual identification, estimation and validation steps

for each one (Fruhwirth-Schnetter & Kaufmann, (2008)). This approach, however, introduces a severe bias if the data-generating mechanism differs substantially between the sequences (Beck & Katz, (1995)). This can be avoided, or at least reduced, by clustering the multivariate time series into homogeneous groups of cross-sectional units so that pooling can only take place within each group. Cornwell & Trumbull, (1994) estimate an economic model of crime by using panel data for $n = 90$ counties in North Carolina over the period $1981 - 1987$. The data are reported as Crime data set in the package plm in R (R Development Core Team, (2009)). Their results suggest that both labor market and criminal justice strategies are important in deterring crime, as well as that the effectiveness of law enforcement incentives has been greatly overstated. The estimation is repeated by Baltagi, (2006a) who confirms the conclusion that county effects cannot be ignored in estimating an economic model of crime. We run the PAM algorithm on the package cluster in R using $p = 20$ variables (region and urban dummies are ignored because they do not vary over time). The partial distance matrices are weighted using the Distatis method . Among the most prominent indicators (*i.e.* those associated with large weights) are the weekly wages in various sectors (whole sales and retail trade; service industry; transportations, utilities, communications and local government). The average sentence and the probability of prison sentence occupy, respectively, fourth and seventh place in the ranking of twenty indicators. The percentage of young males receives the lowest weight.

For simplicity, the counties are divided into two groups as indicated by the dendrogram obtained with the Ward link shown in Figure (1). The cluster on the left is formed by 57 counties and has county 167 as the representative unit whereas the group on the right includes 33 counties with county 187 as the typical unit. By comparing the two leaders, we find that the latter group is characterized by a lower crime rate, higher police per capita, higher "probability" of arrest, lower percentage minority, higher population density and higher weekly salaries with respect to the other cluster. In extreme synthesis, we can conclude that the 90 multivariate short time series arise from two very different groups of counties and that an econometric model based on the same parameters should be estimated within each group.

Figure 1: Hierarchical aggregation of CrimeNC data using the Ward link



3.2.2. A simulation experiment

To study the performance of the distance function presented in section (3.1) the previous section, we apply our technique to a benchmark data set. For this experiment we choose the "Synthetic Control Chart Time Series" (SCCTS) data set collected by Alcock & Manolopoulos, (1999) available from Frank & Asuncion, (2010). The data set contains 600 examples of control charts, each with 60 time series values and there are six different classes with 100 representative examples from each class. We select random samples without replacement of equal sizes from the "normal" group with code $1-100$; the "cyclic": $101-200$; the "increasing trend": $201-300$; the "decreasing trend": $301-400$; the "upward shift": $401-500$; the "downward shift": $501-600$.

We divide the 60 observations of each series into h consecutive equal subsequences of length $n_i = 60/h, i = 1, \ldots, h; h = 3, \ldots, 6$. For a number of cluster $k \in \{2, 3, 4, 5, 6\}$ we select $\{31, 21, 16, 13, 11)\}$ cases for each cluster from SCCTS of the type $1, 2, \ldots, k$. When $k \neq 6$, the classes are chosen at random (without replacement) from $1, 2, \ldots, 6$ in each repetition of the experiment. To simulate time spans of different length, we randomly omit $0, 1, \ldots, t_1$ and $0, 1, \ldots, t_2$ contiguous entries at the beginning and end

respectively of each subsequence, where $t_1 = t_2 \in \{0, 1, \ldots, \lfloor (n_i + 1)/4 \rfloor\}$. The kn_k h-variate time series are classified using the global distance matrix (5) obtained by using the Distatis weights together with the PAM algorithm which is available in the R-package *cluster*.

Table 2: Mean and std.dev. of r from 250 simulations of SCCTS

Method	p		$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Proposed	3	$\mu(r)$	0.855	0.812	0.773	0.718	0.664
		$\sigma(r)$	0.086	0.037	0.018	0.010	0.004
	4		0.875	0.847	0.820	0.760	0.717
			0.079	0.027	0.016	0.007	0.007
	5		0.891	0.815	0.758	0.691	0.642
			0.059	0.029	0.015	0.009	0.006
	6		0.822	0.805	0.743	0.696	0.635
			0.101	0.029	0.013	0.008	0.006
Euclidean	3	$\mu(r)$	0.628	0.614	0.568	0.553	0.517
		$\sigma(r)$	0.130	0.027	0.010	0.006	0.003
	4		0.666	0.609	0.583	0.553	0.544
			0.127	0.027	0.010	0.009	0.003
	5		0.753	0.633	0.592	0.558	0.555
			0.103	0.027	0.010	0.005	0.004
	6		0.694	0.634	0.597	0.581	0.553
			0.124	0.025	0.011	0.004	0.003

To compare the stability of the results of our metric, the data generation is repeated for 250 times for each k and for each h. In all the experiments, we set k to the number of classes in the data set. Since we know the cluster to which each sequence belongs, we can use the known clustering membership and assess the quality of the agreement between the known and the computed cluster membership. The adjusted Rand index r is chosen as the cluster validation measure (see Hubert & Arabie, (1985)). We also report the results obtained using the simple Euclidean metric (Keogh & Kasetty, (2003)) because, at the very least, one should ask wether it makes sense to introduce a new product if it is only slightly better than the one that it replaces. The numerical summaries are given in Table 2.

Our metric is manifestly superior to the Euclidean metric and this reassures us that the additional computations are compensated for better clustering results. The best segmentation of the original time series is at $p = 4$ for which the index r for our metric is generally greater than it is for other alternative subdivisions. The mean of r increases until it reaches its peak at $p = 4$ at which point, precipitation occurs. This "evolution" does not occur for $k = 2$, probably because of the important role played by the pairs of classes that are most often confused with each other (normal/cyclic and decreasing trend/downward shift). The impact of omitted values at the start and/or the end of the sequences, apparently, does not compromise the ability to capture the dynamics of the time series, at least not for the amount of missingness adopted in our experiment. In fact, the average value of r is always around 0.5 for $k = 6$.

4. Toward a taxonomy of Italian provinces

In this section we focus on the co-movements across time and joint extremal behavior exhibited by variables referring to different aspects of education, socio-demographic, economic, and criminal factors that are observed at a NUTS 3 level. In this respect, we have developed a typological procedure to group provinces so that similar units are put into the same group and dissimilar units into different groups. With this in mind, we apply the statistical methodology described in section 3 to the data set discussed in Section 2 which consists of $p = 24$ indicators in the $m = 103$ Italian provinces. Table 3 reports the weights assigned to each indicator under the Distatis weighting procedure.

Table 3: Distatis weights of the indicators

Crime	IC_1	IC_2	IC_3	IC_4	IC_5	IC_6	IC_7	IC_8	IC_9	IC_{10}	IC_{11}
	3.83	2.10	4.60	2.34	3.80	4.15	3.74	3.95	3.39	3.71	4.12
Demography	ID_1	ID_2	ID_3	ID_4	ID_5	ID_6	ID_7				
	4.07	5.01	2.57	5.10	4.94	6.08	5.25				
Econ. Performance	IL_1	IL_2	IL_3	IL_4							
	5.75	5.46	5.95	5.37							
Education	IE_1	IE_2									
	1.98	2.75									

The Distatis procedure is very helpful in the establishing of a hierarchy of influence among indicators which can be measured by considering the weight

each has in the global distance matrix. If all the indicators were weighted equally, then each indicator would be weighted 4.17 (based on a total of 100). In this sense, the most influential indicators are those included in the macrofactor "Economic performance" ($IL_i, i = 1, 2, 3, 4$) because their weight is greater than the average rating plus the standard deviation of the ratings. The indicators contained in the macrofactor demography are also relevant. With the exception of population density (ID_3), all the other demographic indicators have a weight above the average rating. This particularly, referes to the divorce rate (ID_4), the population replacement rate in active age (ID_6) and the young-age dependency ratio (ID_7). The first crime indicator in order of importance is the foreigner crime rate (IC_3) followed by extortion (IC_6) and homicide (IC_{11}). The indicators which contribute the least to global distances are the share of students enrolled in university courses (IE_1), the unknown perpetrators crime rate (IC_4) and, surprisingly enough, the juvenile crime rate (IC_2).

One element which emerges from our study is the presence of convergence groups in the TSCS data set, *i.e.*, a tendency of Italian provinces to cluster around a reduced number of poles of attraction. Table 4 presents the findings of the PAM algorithm applied to the global distance matrix. Figure 2 shows the thematic map resulting from the clustering. The number of clusters is determined with the help of the *R* package *clValid* which contains functions for validating the results of a clustering analysis. In particular, we consider internal measures which use information that is intrinsic to the data in order to assess the quality of the clustering (see Handl *et al*, (2005)). The specific internal validation measures included in the package are connectivity, silhouette width, and Dunn index.

In each graph we find a positive or negative jump in the curve, or a local peak in correspondence of $k = 7$ clusters. Based on this choice, the PAM technique proposes:

1. three large clusters *i.e.* fragile provinces, core provinces and provinces at risk;
2. two medium sized groups: robust and smart provinces;
3. a small group of metropolitan areas;
4. a singleton cluster formed by the province of Naples.

The cardinalities of the clusters are reported in the "units" column; the other column on the right of the number of units in the cluster, shows the membership of the provinces. The "medoids" column shows the representative

Table 4: Clustering of Italian provinces.

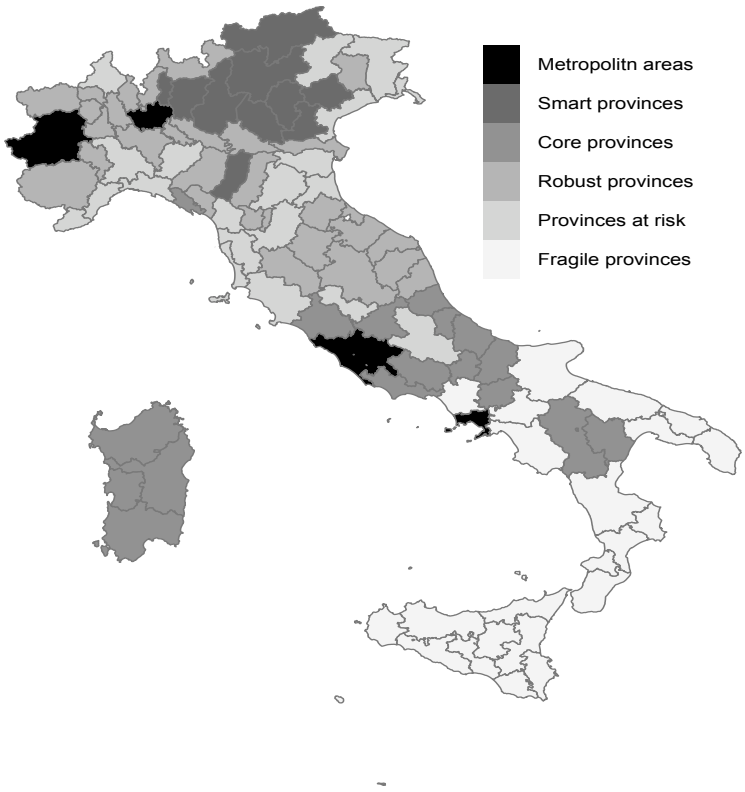
	Name	Medoid	Units									
C_1	Fragile provinces	Taranto	22	Ag	Av	Ba	Br	Cl	Ce	Ct	Cz	Cs
				En	Fg	Le	Me	Pa	Rg	Rc	Sa	Sr
				Kr	Ta	Tp	Vv					
C_2	Robust provinces	Campobasso	17	Aq	Bn	Ca	Cb	Ch	Fr	Is	Lt	Mt
				Nu	Or	Ps	Pz	Ri	Ss	Te	Vt	
C_3	Provinces at risk	Udine	21	Al	Bl	Bo	Fi	Fe	Ge	Go	Gr	Im
				Sp	Li	Lu	Pc	Pi	Ra	Sv	Tr	Ts
				Ud	Ve	Vb						
C_4	Core provinces	Novara	29	An	Ao	Ar	Ap	At	Bi	Co	Cr	Cn
				Fc	Lo	Mc	Mn	Ms	Mo	No	Pr	Pv
				Pg	Pu	Pt	Pn	Po	Rn	Ro	Si	So
				Va	Vc							
C_5	Smart provinces	Vicenza	10	Bg	Bz	Bs	Lc	Pd	Re	Tn	Tv	Vr
				Vi								
C_6	Metropolitn areas	Milan	3	Mi	Rm	To						
C_7	Naples	Naples	1	Na								

or typical province of the cluster whose characteristics are decisive in the identification process of the group. In fact, the interpretation of clusters is possible on the basis of the values of indicators for the typical units of the clusters.

A cursory review of the cluster compositions in Table (4) and the cluster mapping in Figure (2) suggests that clusters can be ordered according to the degree of economic performance and crime rates as reflected by the resulting configuration of the given set of indicators, particularly in the cluster medoids. The first two clusters (fragile and robust provinces) form the so-called Mezzogiorno of Italy, but constitute two opposite sides of the same coin.

Cluster C_1 - Fragile Provinces - is made-up of 22 provinces, many of which are geographically contiguous and historically characterized by the presence of crime cartels. It contains a number of large provinces, some medium and a few small provinces. Fragile provinces lack the functional authority to provide basic security within their borders and the institutional capacity to be responsive to basic social needs for their populations. In fragile provinces, investment in human capital is low in the period considered. Economic performance is weak; in fact, the cluster is characterized by a very high unemployment rate and by a low value added per capita. Poor labor market performance is a distinctive feature of this group's profile. Furthermore, given that the hidden economy constitutes an important phenomenon in this

Figure 2: A taxonomy of Italian provinces



area, it is also characterized by weak legality. The Southern provinces in the cluster are closely together, suggesting that spatial proximity may affect their labor market performance. Provinces with very high initial unemployment rates could be penalized by the proximity of provinces with similar initial conditions. In the period under consideration, the average total number of crimes in general and that of crimes committed by young people in particular is high.

Cluster C_2 - robust provinces - is a cross-regional cluster of 17 provinces localized in the Center-South of Italy. The provinces are small, with low population density, and tend to be resistant to law-breaking behavior. C_2 is typified by high investment in human capital, the highest of those registered

in the seven medoids. This cluster is characterized by a high unemployment rate, which is higher among the young, and by a moderate average value added per capita. However, agglomerations of firms integrated through a complex web of economic and social relationships, known as industrial districts (e.g. Pescara, Isernia, Sassari, Matera, Frosinone), fall into the group. Usually the small towns, where there is active social control, are the quietest. In fact, the average number of total crimes committed in robust provinces is rather low, the lowest among those registered in the seven leader provinces. The average numbers of incidents of extortion, robbery, drug-related crime and murder are the lowest in absolute. One might conclude that it is a cluster which, though clearly not particularly successful in terms of economic performance, does not create great concerns in terms of safety. Probably, the investment in human capital and the strong socio-cultural anchor of the industrial districts in the cluster discourages criminal activity.

Cluster C_3 - Provinces at risk - consists of 21 provinces localized in the Center-North. The cluster comprises regional capitals, a number of provinces with access to the sea and several seaport provinces. Port regions always seem to be at an advantage when compared to those provinces which are not situated by the sea or on rivers. C_3 is characterized by significant investment in human capital. The economic performance is quite good. The average unemployment rate is, in fact, very low, the youth unemployment rate is not high, and the value added per capita is high. On average, a conspicuous number of crimes is registered. The number of incidents of extortion and drug-related crime is not negligible. The number of crimes committed by young people is surprisingly high. It is surprising because a number of young people are engaged in studies, with, therefore, little time to devote to illegal activities, and, above all, because there is a lower youth unemployment rate than in other provinces where the total number of crimes committed by young people is lower.

Cluster C_4 - core provinces - is spatially the most dispersed cluster. In fact, it is made up of 29 provinces from nine different Regions. In general, these provinces include only one city and this constitutes the quasi-exclusive center for activities in the province. Some crime attracting provinces belong to this cluster. Novara, for example, is a crime attracting province because of its high per capita value added and because of its geographic position; it is located on the national border and is an important crossroads for commercial traffic between Milan, Turin, Genoa and Switzerland. Rimini, the capital of entertainment, is invaded by over a million of tourists each year and has

the record for crimes per inhabitant. A number of industrial districts (*e.g.*, Parma, Prato, Pesaro, Biella, Pavia) belong to this cluster. The average unemployment rate in the reference period is low, the youth unemployment rate is not too high and the value added per capita is high; therefore economic performance is good. Investment in human capital is low. The number of total crimes is not high on average; the number of crimes committed here by young people probably reflects the fact that few young people pursue an education and the percentage of the young who are unemployed cannot be ignored. Drug-related crimes, extortion, robberies and murders are not negligible.

Cluster C_5 - smart provinces - is a small cross-regional cluster of 10 provinces located in the Center-North. The cluster has low investment in human capital. The distinctive feature of this cluster is an excellent economic performance. The average unemployment rate during the reference period and the youth unemployment rate are the lowest among those observed in the seven medoids while the value added per capita is the highest among all the medoids, except for Milan. The industrial districts can be found in Brescia, Bergamo, Reggio Emilia, Vicenza, Verona and Treviso. A high number of total crimes is registered, on average, while the number of crimes committed by young people is low, probably thanks to the fact that few young people are unemployed. The level of drug-related crime, extortion, robbery and murder is moderate or high.

Cluster C_6 includes three metropolitan areas, each comprised of a major city and its related communities: Milan, Rome and Turin. The elements in this cluster are crime importers because of their high per capita value added. Rome, for instance, is the capital city and seat of parliament and the ministries. Turin is the fourth Italian city in terms of population after Rome, Milan and Naples and the country's third economic pole; it is the capital of the Italian car industry. Milan, Rome and Turin are all big cities with a population that does not coincide with the number of residents. A number of tourists visit Milan and Rome and, thanks to the presence of large Universities and research centers, many young people from all over Italy and abroad live there for study reasons.

Cluster C_7 - Naples - is a singleton cluster formed by a single province, Naples. It is a metropolitan area in the South of the Country which is often associated with the organized crime but it is not a part either of C_6 or of C_1. Some characteristic features of Naples are a high unemployment rate, an extremely high youth unemployment rate and low value added per capita.

Investment in human capital is low. Naples has the highest number of incidents of extortion, the highest number of robberies and the highest number of homicides for the period considered. Compared to the other three metropolitan areas, it has the lowest number of drug-related crimes, the lowest number of total crimes and, unexpectedly, the lowest number of crimes committed by young people. This is unexpected because of the extremely high youth unemployment rate. Perhaps young people are attracted by criminal organizations and they commit different crimes from those usually committed by young people elsewhere. On the other hand, the propensity to report crimes in the province could simply be low or lower than that registered elsewhere.

R code and the data used in this paper are available from the authors upon request.

5. Conclusions and suggestions for future research

The purpose of this paper is to analyze spatio-temporal links between economic performance and various criminal activities in the 103 Italian provinces for the period $1995 - 2009$. More specifically, we make an attempt to determine compact and separate clusters of provinces by using an unbalanced time series of cross-section data set concerning 24 indicators of crime, economic performance and other socio-demographic aspects. In practice, we add a new technique to the "methodological tool bag" that researchers have at their disposal to analyze TSCS data.

The evidence indicates that there are significant disparities in the spatio-temporal evolution of crime across Italian provinces. These disparities stem from many factors, including structural differences in the geographical distribution of value-added, unemployment, education, and other sociodemographic indicators. For instance, provinces with higher crime rates are often linked with high rates of family disruption (divorces), youth unemployment and female employment. Our main contribution is to show that crime is not as inextricably linked to geographic location as is usually believed. From this point of view, the position of the Mezzogiorno, a long discussed controversial subject, has two facets; one is found in relatively non-affluent provinces which are not highly subject to random street crime or organized crime and the other is to be found in provinces caught in a vicious circle in which increasing criminal activity and weak economic performance feed on each other to undermine the security of the population. The pattern for the Center-North of Italy is much more varied and composed of a series of four

clusters. In particular, there is a relatively large cluster of provinces which should be considered at risk because of an increasing value added per capita, a high level of population replacement rate, and a good trend in female employment; all conditions that might encourage criminal activity. A special mention must be given to the cluster composed of the three economically most important Italian cities (Milan, Turin, Rome) and Naples. There is a strong relationship between the spatial characteristics of these metropolitan areas and crime rates. The evidence points to cluster C_6 and Naples as provinces sharing most of the characteristics of other clusters. They are simultaneously core and smart provinces, as well as fragile provinces (Naples) or provinces at risk (Turin).

Diversification within the Italian economy and the geography of crime evidentiate that a crucial role should be attributed to locally-implemented policies that are able to work with territorial differences, even though national level policy also plays an important role. The identification of a mix of national and local policy measures is preferable. Results suggest the need for economic policies which are able to increase the level of employment, especially in some Southern areas, encourage greater investment in education (especially for Core and for Smart provinces) and increase vigilance, particularly in Core and in Robust provinces, in order to avoid the risk that they may become attractive to crime.

Starting from our findings, it would be interesting to further clarify the role played by distinct aspects of labor market quantity (joblessness) and quality (secondary sector work and low-wage jobs) in increasing illegal activities. This could be carried out, for instance, by estimating a regression model for each cluster, after a suitable allocation of metropolitan areas to the clusters.

It should be mentioned that there is a limitation to our analysis resulting from the use of the NUTS 3 classification. This choice is not optimal for all the situations that matter. Many Italian regions (NUTS 2 level), are entirely included in specific clusters. Provinces do not necessarily correspond to homogenous and self-contained regions. Perhaps, the most appropriate administrative unit for the type of analysis presented in this paper is a mixture of administrative areas at NUTS 2 and NUTS 3 levels. Admittedly, we computed the metric distance between provinces neglecting the spatial interdependence between them. A possible improvement in the effectiveness of our clustering algorithm can be a distance function which captures spatial contiguity effects.

From the point of view of the statistical methodology, future studies should address the clustering building process by carrying out a different PAM study for each indicator. In this sense, consensus clustering is a promising field for interdisciplinary research because it combines multiple individual clustering results into a single consensus solution so as to improve the accuracy and stability of clustering. Finally, another fruitful line of investigation that might be addressed is the reduction of the number of indicators, by using aggregate variables (*e.g.* dynamic principal component analysis). In this sense, we note that the weights of the partial distance matrices provide a reliable means for ranking different indicators with respect to their ability to distinguish between the clusters.

References

Abdi, H. and O'Toole, A. J. and Valentin, D. and Edelman, B. (2005).DISTATIS: the analysis of multiple distance matrices. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, (USA), 42–47.

Akomak, I. S. and ter Weel, B. (2008). The impact of social capital on crime: evidence from the Netherlands. *IZA Discussion Paper*, N. 3603.

Alcock, R. J. and Manolopoulos, Y. (1999). Time series similarity queries employing a feature-based approach. In 7th Hellenic conference on informatics, Ioannina, Greece.

Al-Marhub, F. (2000). Income inequality and inflation: the cross-country evidence. *Contemporary Economic Policy*, **18**, 428–439.

Anselin, L. and Cohen, J. and Cook, D. and Gorr, W. and Tita, G. (2000). Spatial analyses of crime. Pp. 213–262 In Duffee D. (Ed.) *Criminal Justice 2000, Volume 4, Measurement and Analysis of Crime and Justice*, National Institute of Justice, Washington, DC.

Avio, K. L. (1988). Measurement errors and capital punishment. *Applied Economics*, **20**,1253–1262.

Baltagi, B.H. (2006). Estimating an economic model of crime using panel data from North Carolina. *Journal of Applied Econometrics*, **21**, 543–547.

Baltagi, B.H. (2006). Unbalanced panel data: A survey. *Statistical Papers* **47**, 493–523.

Batagelj, V. and Bren, M. (1995). Comparing resemblance measures. *Journal of Classification*, **12**, 73–90.

Beck, N. and. Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Journal Review*, **89**, 634–647.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, **76**, 169–217.

Bianchi, M. and Buonanno, P. and Pinotti, P. (2010). Do immigrants cause crime? *Journal of the European Economic Association*, Forthcoming. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1785005

Buonanno, P. (2006). Crime and labour market opportunities in Italy (1993-2002), *Labour. Review of Labour Economics and Industrial Relations*, **20**, 601–624.

Büttner, T. and Spengler, H. (2003). Local determinants of crime: distinguishing between resident and non-resident offenders. Discussion Paper N. 03-13, ZEW Centre for European Economic Research.

Buonanno, P. and Leonida, L. (2009). Non-market effects of education on crime: evidence from Italian regions.*Economics of Education Review*, **28**, 11–17.

Calderoni, F. (2011). Where is the mafia in Italy? Measuring the presence of the mafia across Italian provinces. *Global Crime*,**12**, 41–69.

Cantor, D. and Land, K. C. (1985). Unemployment and crime rates in the post-World War II U.S.: A theoretical and empirical analysis. *American Sociological Review*, **50**, 317–332.

Crdenas, M. and Rozo, S. (2008). Does crime lower growth? Evidence from Colombia. Working paper, No. 30. Commission on Growth and Development.

Cornwell, C. and Trumbull, W. N. (1994). Estimating the economic model of crime with panel data. *Review of Economics and Statistics*, **76**, 360–366.

Cracolici, M.F. and Uberti, T. E. (2009). Geographical distribution of crime in Italian provinces: a spatial econometric analysis. *Jahrbuch für Regionalwissenschaft*, **29**, 1–28.

Craig, S.G. (1987). The deterrent impact of police: An examination of a locally provided public service. *Journal of Urban Economics*, **21**, 298–311.

Daniele, V. and Marani, U. (2011). Organized crime, the quality of local institutions and FDI in Italy: A panel data analysis. *European Journal of Political Economy* **27**,132–142.

Detotto, C. and Otranto, E. (2010). Does crime affect economic growth? *Kyklos*, **63**, 330–345.

Dutta, M. and Husain, Z. (2009). Determinants of crime rates: Crime deterrence and growth in post-liberalized India. MPRA Paper No. 14478

Edmark, K. (2005). Unemployment and crime: is there a connection? *Scandinavian Journal of Economics*, **107**, 353–373.

Ehrlich, I. (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *The Journal of Political Economy*, **81**, 521–565.

Entorf, H. and Spengler, H. (2000). Socioeconomic and demographic factors of crime in Germany: evidence from panel data of the German states. *International Review of Law and Economics*, **20**, 75–106.

Entorf, H. and Spengler, H. (2002). *Crime in Europe: causes and consequences*, Springer-Verlag, Berlin.

Forni, M. and Paba, S. (2000). The sources of local growth: evidence from Italy. *Giornale degli Economisti e Annali di Economia*, **59**, 1–49.

Fox, J.A. (1978). *Forecasting Crime Data - An Econometric Analysis*. Lexington Books, Lexington, Mass.

Frank, A. and Asuncion, A. (2010). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.
http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes *Journal of the Royal Statistical Society. Series B*, **66**, 815–849.

Fruhwirth-Schnetter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, **26**, 78–89.

Gaibulloev, K. and Sandler, T. (2008). Growth consequences of terrorism in Western Europe. *Kyklos*, **61**, 411–424.

Gould, E. D. and Weinberg, B. A. and Mustard, D. (2002). Crime rates and local labor opportunities in the United States: 1979 – 1995. *Review of Economics and Statistics* **84**, 45–61.

Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5–48.

Handl, J. and Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post- genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Hansen, K. (2006). Male crime and rising female employment. Centre for Longitudinal Studies, Institute of Education, University of London.
http://www.cls.ioe.ac.uk/

Harries, K. (2008). Property crimes and violence in United States: An analysis of the influence of population density. *International Journal of Criminal Justice Sciences*, **1**.

Hsieh, C. C. and Pugh, M. D. (1993). Poverty, income inequality, and violent crime: A meta-analysis of recent aggregate data studies. *Criminal Justicie Review*, **18**, 182–202.

Hubert, L. and Arabie, P. (2005). Comparing partitions. *Journal of Classification*, **2**, 193–218.

ISTAT (2011). http://en.istat.it/dati/db_siti/

Kapuscinski, C. A. and Braithwaite, J. and Chapman, B. (1998). Unemployment and crime: Toward resolving the paradox. *Journal of Quantitative Criminology*, **14**, 215–243.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Keogh, E. J. and Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, **7**, 102–111.

Lochner, L. A. (2004). Education, work and crime: a human capital approach. *International Economic Review*, **45**, 811–843.

Lochner, L. (2010). Education Policy and Crime. In P. Cook, J. Ludwig and J. McCrary (eds.), *Controlling Crime: Strategies and Tradeoffs, Chapter 11*, University of Chicago Press, Chicago. Forthcoming.

Lochner, L. and Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *The American Economic Review*, **94**, 155–189.

Machin, S. and Olivier, M. and Vuji, S. (2010). The crime reducing effect of education. IZA Discussion paper No. 5000.

Mauro, L. and Carmeci, G. (2007). A poverty trap of crime and unemployment. *Review of Development Economics*, **11**, 450–462.

Patacchini, E. (2008). Local analysis of economic disparities in Italy: a spatial statistics approach, *Statistical Methods & Applications*, **17**, 85–112.

Pavoine, S. and Vallet, J. and Dufour, A-B and Gachet, S. and Herve, D. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, **18**, 391–402.

Pellegrini, L. and Gerlagh, R. (2004). Corruption's effect on growth and its transmission channels. *Kyklos*, **57**, 429–456.

Peri, G. (2004). Socio-cultural variables and economic success: evidence from Italian provinces 19511991. *The Berkeley Electronic Journal of Macroeconomics. Topics in Macroeconomics*, **4**, article 12.

R Development Core Team, (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Raphael, S. and Winter-Ebmer, R. (2001). Identifying the effect of unemployment on crime. *Journal of Law and Economics*, **44**, 259–283.

Raveh, A. (1981). A graphic display for characterization of seasonal pattern similarities of time series. *The Statistician*, **30**, 179–182.

Santas, M. (2007). A Risk analysis of crime and high school dropout rates. *Research Paper, AYSPS*, Georgia State University.

Saridakis, G. (2004). Violent crime in the United States of America: A time-series analysis between 1960-2000. *European Journal of Law and Economics*, **18**, 203–221.

Scorcu, E. A. and Cellini, R. (1998). Economic activity and crime in the long run: An empirical investigation on aggregate data from Italy, 1951–1994. *International Review of Law and Economics*, **18**, 279–292.

Sinha, G. and Mark, D. M. (2005). Measuring similarity between geospatial lifelines in studies of environmental health. *Journal of Geographical Systems*, **7**, 115–136.

South, S. J., and Messner, S. M. (2000). Crime and demography: Multiple linkages, reciprocal relations. *Annual Review of Sociology*, **26**, 83–106.

Travaglini, G. (2003). Property Crime and Law Enforcement in Italy. A Regional Panel Analysis $1980 - 95$. *Giornale degli Economisti*, **62**, 211–240.

Wilson, S. E. and Butler, D. M. (2007). A lot more to do: the sensitivity of time-series cross-section analyses to simple alternative specifications. *Political Analysis*, **15**, 101–123.

Witte, A. D. and Witt, R. (2001). Crime causation: Economic theories. *Encyclopedia of Crime and Justice*.