Chapter 5

Application of Non Parametric Maximum Likelihood Estimation to Greek data

5.1 Using official crime reports instead of real crime counts

Since one can argue that analyzing reported crime counts will not lead in valid inferences, as it is obvious that official crime statistics do not necessarily reflect the “true” nature of crime in our society, in this paragraph we give a short discussion on using official crime data.

Black (1951), states that criminal statistics are a regular feature of the administration of justice in all civilized countries. He also states that they give a fair indication of the present crime situation.

The first problem of official crime counts is under-reporting. There are many reasons why an individual may not report an incident to the police, having as a result under-reported crime. In the sequence we will present a brief summary of these factors (for a more detailed discussion of these factors see Skogan (1994) or MacDonald (2001)).

Demographic factors are likely to influence reporting behavior. In addition to age, gender and ethnicity, victimization surveys have shown that certain groups within inner cities are far more likely to experience burglary and this repeated experience tends to reduce their reporting inclinations as they develop a lower expectation of the reporting outcome (Maguire, 1997). It has also been shown that individuals in higher income groups, particularly older people, tend to be more insured than lower income people or those who are unemployed, and consequently are far more likely to report an incident (Lewis, 1989).

In addition to these socio-economic factors, individual attitudes are also likely to influence reporting behavior. For example, if a victim perceives the police to be ineffectual, or has had a negative experience of the police, then that person is probably more reticent to contact the police about an incident than otherwise would be the case. Similarly, if the victim has a sense of culpability about the incident (for example, they left the window open), then this will probably reduce the likelihood of reporting the incident (Skogan, 1994). In addition, just as individual attitudes to the police may affect reporting inclinations, then so might individual criminality. For example, individuals who are involved in criminal activities such as drug taking are probably unlikely to want to involve the police should they become victims of property crime.

Finally, there are numerous incident-specific factors that probably have a very strong influence on whether or not to report an incident. Undoubtedly if an incident results in financial loss to the victim (either due to stolen property or damage) then this is an incentive to report the crime, not least because this may be a requirement for an insurance claim. However, it should be noted that when asked, victims typically state that these types of incidents would have still been reported even if an insurance claim were not being made (Budd, 1999). These incident-specific factors are typically thought of to influence the individual's assessment of the costs and benefits of reporting (Goldberg and Nold, 1980), and include factors such as the seriousness of the crime, the perceived threat from the incident, and when the incident occurred.

Moreover, as mentioned before about the way that the police and the Government are gathering the crime statistics, there is also the problem of under-recording crime. Both under-reporting and under-recording have become subjects of interest for several researchers.

Black (1970), states that the major uses of official crime statistics have taken two forms and each involves a different social epistemology, a different way of structuring knowledge about crime. One employs official statistics as an index of the “actual” or “real” volume and morphology of criminal deviance in the population. According to Black, those who follow this approach typically consider the lack of fit between official and actual rates of crime to be a methodological misfortune. Historically, measurement of crime has been the dominant function of crime rates in social science. A second major use of official statistics abandons the search for "actual" deviance. This is managed either by defining deviance with the official reactions themselves, or by incorporating the official rates not as an index of deviant behavior but as an index of social control operations. In effect this second range of work investigates "actual" social control rather than "actual" deviance.

Black states that when official statistics are used as a means of measurement and analysis they usually function imperfectly. For this, in his paper he follows an alternative strategy, which makes official records of crime an end rather than a means of study. His strategy treats the crime rate as itself a social fact, an empirical phenomenon with its own existential integrity. From this standpoint crime statistics are not evaluated as inaccurate or unreliable. Black states that they are an aspect of social organization and cannot, sociologically, be wrong.

MacDonald (2002), examines the differential between recorded crime statistics and actual experiences of crime in the England and Wales. MacDonald explores what data are available to researchers in Europe and the United States, and how recorded crime levels can be affected by different factors that vary across countries.

Criminologists argue that these official crime statistics are socially constructed (Maguire, 1997). This is a consequence of how society defines crimes, of the discovery of crimes, and of the reporting and recording of crimes. Perhaps the biggest problem that researchers face is with respect to what criminologists call the "dark figure" of crime, that is, those hidden crimes which are experienced by the public but are not included in official recorded crime statistics. The importance of the "dark figure" for modeling crime is that any change in the public's willingness to report crimes, or a change in police reporting practices, can give rise to a change in the recorded crime rate that may not truly reflect the changing experience of crime by victims. Excluding social definitions of crime and unnoticed (victimless) crime, the "dark figure" of crime is a primarily a consequence of two sources of error: under-reporting (by victims) and under-recording (by the police).

MacDonald (2002), uses data from the British Crime Survey to estimate a probability of reporting equation in order to evaluate the influences on the "dark figure" of crime. It then becomes clear that not all crimes are reported to the police. MacDonald states that, although it might be argued that non-reporting is related to how serious the crimes are considered by the victims, given the very low reporting rates for domestic violence, it must be the case that factors other than the seriousness of the crime have a substantial influence on reporting behavior. MacDonald also notices that the proportion of crimes reported to the police is not constant over time. In order to model crime, it is important to determine whether or not this variation in reporting over time is random, or whether it varies systematically. If the latter is true, then MacDonald it should be somehow possible to take account of the "dark figure" when using official crime statistics. MacDonald also notices that just in terms of under-reporting, there are some categories of crime for which official statistics are going to be particularly plagued by the "dark figure" (vandalism, common assault, and domestic violence).

As far as the under-recording is concerned, MacDonald states that although a considerable proportion of the "dark figure" is due to underreporting, police recording practices also have an impact on the discrepancy between official crime statistics and the experience of victims. MacDonald argues that this is partly because some incidents that are believed to be crimes by victims are not notifiable. However, in an extensive review of police recording practices for the London's Home Office, Burrows et al. (2000) found that there are many reasons why the police may not record crimes that are not due to how incidents are defined in law. The authors report that the primary reason for under-recording is police discretion, and in particular the application of an evidential standard to allegations of crime that is not applied uniformly across police forces (or across years). They can not be certain, though, why different forces vary in their use of discretion (for example it may be related to police numbers or it may be politically motivated).

Therefore, it becomes clear that the two sources of measurement error combine to yield a series of official statistics that do not reflect the full extent of victimization. However, MacDonald (2002) argues that they are not a problem for economic models of criminal activity if the British Crime Survey measure of crime and the official recorded crime rate follow each other over time, and maintain the same order of magnitude of difference. MacDonald notes, however, that the British Crime Survey does not reveal a complete picture of victimization, and is itself limited due to the nature of surveys (for example, the British Crime Survey is likely to be subject to sampling errors, respondent recall and evasion bias, and general problems due to how representative the sample is. Nonetheless, MacDonald points out that the comparison shows that British Crime Survey does track the recorded figures reasonably well. However, the differences between the two series are not constant over time.

These facts were known even decades ago. Black (1951), recognizes the "dark figure" in official crime statistics and he also admits that it can not be ignored nor neutralized, because it is by no means constant. He claims that the "dark figure" is relative to the type of crime and criminals, the strength and efficiency of the police, the changing attitude of the public in their inclination to report suspected crimes and to prosecute alleged offenders, and to changing methods of recording complaints made to the police. He also notes that unless the strength of the police force is constantly adapted to the volume of crime, the proportion of undisclosed crime is bound to increase with the high tide of criminality.

However, Black states that the number of crimes known to the police, as the nearest possible approach to the source, has been generally recognized as the best available index of the volume of crime.

Yannaros (1993), also states that the problem of estimating unobserved events, such as unreported crimes, remains unsolved. As far as the methods used to estimate the true number of crimes, such as interviewing possible victims and prisoners, he comments that since both victims and prisoners may find the questions asked 'sensitive' it is impossible to get reliable estimates of the fraction of crimes which remain unreported.

Yannaros also comments that the approach based on the properties of success counts in the binomial situation, which has been studied for a long time, is based in an assumption that is unrealistic in practice (see Olkin et al., 1981, for the long history of this problem). He even notes the non-existence of a better model, which reflects the difficulty of the problem.

Since there is no better approximation but the official statistics of the crime activity we use the official crime counts of that period to obtain a picture of crime activity in Greece. Moreover, we know at least that the method used by Greek Police and the Government to collect that data has not changed for the period 1987-1997. Thus, although the problem of under-reporting still exists, the factor of under-recording by the Authorities can be assumed to be constant for that time period.

5.2 Greek Data considerations

Borowick (1998), as mentioned before, has shown that reporting crime counts follow some kind of mixed Poisson distribution, semiparametric maximum likelihood (SML) methods, as those presented in the previous chapter, can be applied to the available annual Greek crime data.

In our application we use in fact non-parametric maximum likelihood estimation (NPMLE). The algorithm used is proposed by Karlis (2001). In order to derive the NPMLE, Karlis used a hybrid algorithm that is based on the typical EM algorithm. The basic ingredients of his algorithm were the following:

1. He started with a mixing distribution having as many number of components as the results of Lindsay (1995) advocate. This sets an upper bound on the support points which for the Poisson case is just the half of the different values of the data at hand.
2. Having decided on the number of initial points he selected them randomly, the λ's are selected by choosing uniform random number s in the interval $(0, \max X_i)$ while the mixing proportions were set equal to uniform (0,1) random numbers rescaled so as to sum to 1.
3. After the selection of initial points the classical EM algorithm for finite Poisson mixtures started running. After 100 iterations he checked for redundant points that can be dismissed without make the loglikelihood to decrease. When redundant components were identified he merged those components. There are two kind of redundant points

- mixing probability close to 0, namely smaller than 10^{-6} implying a component with very small probability which eventually, if the algorithm keeps running, will end to an empty component, and
- components parameters too close together, i.e. λ's that are closer than 10^{-6} in absolute value, implying two components with the same mean; this implies that if one merges these two components the loglikelihood does not change.

  However, we took care to ensure that the merged solution has not worst likelihood than the one before merging.

4. He stopped the algorithm when a certain criterion indicated that the algorithm cannot progress any further. The criterion used was that the relative difference of the loglikelihood in two successive EM iterations was smaller than 10^{-12}.
5. Finally, he used the gradient function to verify that the solution found is truly the NPMLE. If not, the algorithm is started again from different initial values.

Some interesting points related to the algorithm are the following:

- For each sample we started from different initial sets of values until the gradient function based criteria to be satisfied.
- This algorithm is not optimal in any sense but it is easy to be programmed. Alternatively more sophisticated algorithms can be used (see, Böhning, 2000).
- The algorithm is in fact a backward algorithm and we get rid of redundant components.

Karlis' algorithm was written in Pascal in order to be applied in Greek data.

For the application of the algorithm, the Greek counties populations for the corresponding years were also needed. For the years 1981 and 1991, the populations can be obtained by the Greek Census of 1981 and 1991, respectively. However, for the analysis to be more accurate, these populations could not be used as an approximation for all the other years from 1987 to 1990 and from 1992 to 1997.

In order to obtain more accurate results we used in our analysis approximations of Greek counties populations in cooperation with Virras. Virras (2001), in his MSc thesis has applied Survival Analysis techniques to Greek counties collecting primary data for them for the period 1971-1993. He states that the progress of a region (growth or not) depends on its attractiveness, meaning the ability of a region to draw business units and the right blend of people to run them. He expresses this ability as a variable called Basic Image of the region. Basic Image is a function of a multitude of economic, social and environmental factors, whose values lie in the interval [-1,1]. The Basic image of a given area measures the degree to which this area satisfies a set of basic criteria common for all movers (employers, professionals, unskilled workers, skilled workers, etc.). An area satisfying those criteria is considered, by all potential movers, as worth a closer examination and as potential final choice. Positive Basic Image indicates an attractive region, while negative Basic image a repulsive one. Virras states that if the Basic Image is quantified properly, one would expect that changes in its value and measurable changes of the region's population due to migration would generally agree in sign (i.e. both positive or both negative).

Virras has calculated the Basic Image values of every region for the period 1971-1993. Using the Greek census data of 1971, 1981 and 1991 and the results of his application Virras computed approximations of the annual Greek counties populations for the period 1971-1993.

Hence, we use these population approximations for the period 1987-1990 and 1992-1993. For the years 1994-1997 we use the populations of 1993. In Appendix A the Greek counties populations used are given.

In order to apply the NPMLE algorithm, we created 246 separate files for each of the offenses under consideration, for each of the years 1987 to 1997, in ASCII format, with two columns. The first column contains the count of the reported offenses for the corresponding year. The second column contains the population of each county for the same year. Using Pascal, we run Karlis' algorithm, for each of the files separately, obtaining one ASCII file containing the results of each run. In Appendix A we give one input file and the corresponding output file as examples.

5.3 Obtained distribution parameters using the NPMLE algorithm

Since we assume that the reported crime counts come from a mixed Poisson, the number of its subpopulations is the first parameter we look for. Using this number we can cluster the Greek counties. As mentioned before, the algorithm is run for each offense for each year separately. Thus, the clustering of the counties in this stage can only be done for the corresponding offense for each year separately.

For example, we obtained that offenses concerning antiquities for the year 1997 constellate the Greek counties in four clusters, since the estimated number of subpopulations found by the algorithm to be four (4).

For each obtained subpopulation, its members are also obtained. In Appendix B all the obtained memberships of the Greek counties are given.

Moreover, for each obtained subpopulation the mean λ is given. This is of great importance, since it is the mean offense count for the obtained cluster.

For example, for the offenses concerning antiquities, for the year 1997, the four obtained clusters have means:

$$\lambda_1 = 1.12464,\ \lambda_2 = 8.00206,\ \lambda_3 = 33.24285 \text{ and } \lambda_4 = 51.61833.$$

Thus, counties of the first cluster have low crime activity concerning antiquities, whereas those of the fourth cluster have high crime activity.

As we can see the clusters are given in ascending order, corresponding to higher values of mean λ, as the number of cluster increases.

5.4 Mapping the results of the NPMLE application

The obtained clusters can be graphically represented in a map. Coloring the county members of each cluster with a different color, a simple look at the map gives a full picture of crime activity for each offense for each year.

However, there is not yet software that can perform that coloring for Greek counties automatically, simply by giving the corresponding membership for each county. In order to present, though, an example of the mapping of the clusters, we found a Greek map and we colored it using Photoshop.

The same map will be used in the following sections, colored accordingly, to give a graphical presentation of the clustering results. The editing of the maps is performed using Photoshop.

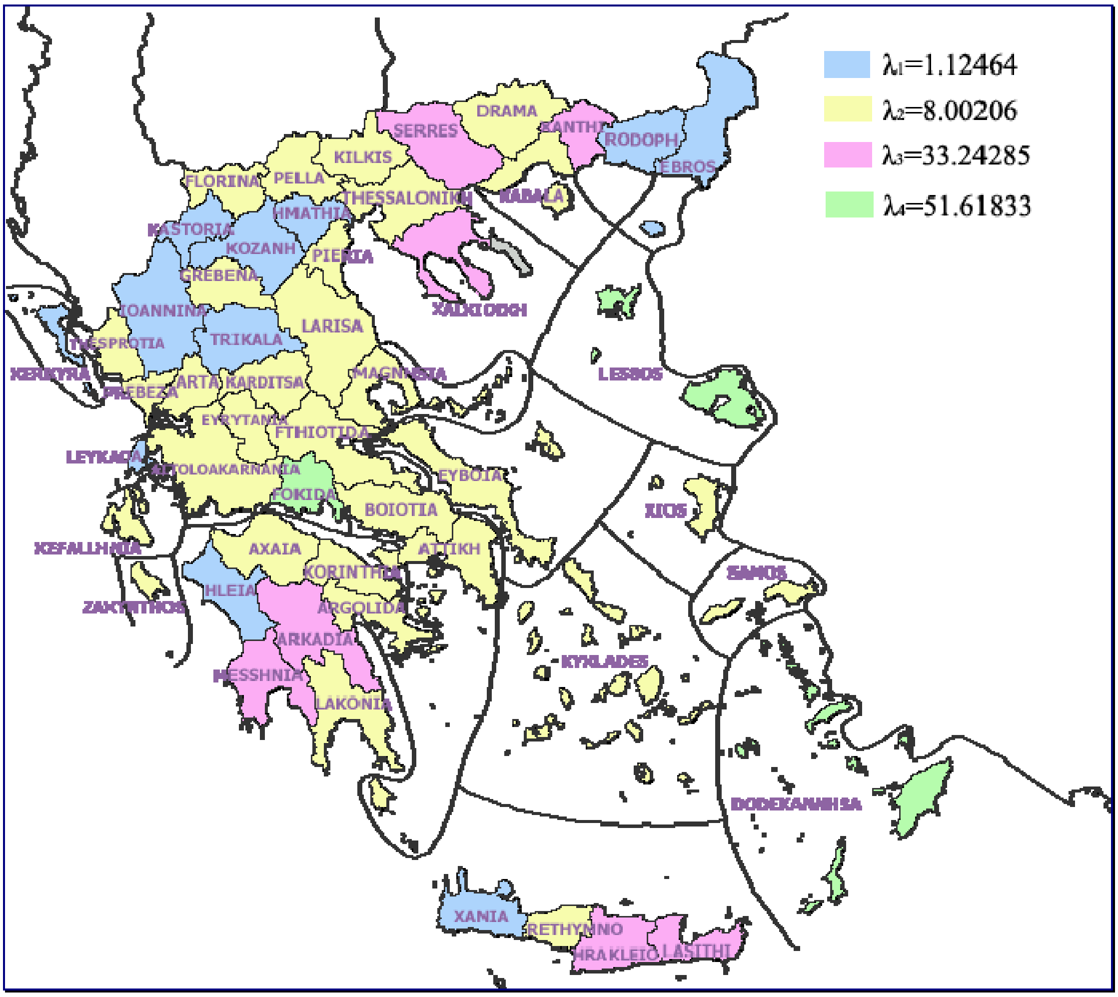

Figure 57. Mapping of Greek counties clustering according to offenses concerning antiquities for the year 1997.