Electronic Journal of Applied Statistical Analysis
EJASA (2013), Electron. J. App. Stat. Anal., Vol. 6, Issue 1, 118 – 129
e-ISSN 2070-5948, DOI 10.1285/i20705948v6n1p118

OUTLIER DETECTION WITH KERNEL DENSITY FUNCTIONS IN MONITORING THE ISTAT LFS DATA PRODUCTION PROCESSES

Rita Lima*

Italian National Institute of Statistics, Rome, Italy

Received 30 April 2012; Accepted 31 October 2012
Available online 26 April 2013

***Abstract**: To calculate the thresholds that determine the normal operating region for a data production process, it is proposed a method based on a time-varying tolerance bounds. The proposal is applied on the Italian Labour Force Survey to report how this new decision rules impact on survey quality.*

***Keywords**: Monitoring data production, AAPOR, time series analysis*

1. Introduction

The aim of the Labour Force Survey (LFS) in the National Statistical Institute of Italy (ISTAT) is a highly reliable labour market data at the national produced and disseminated on monthly and quarterly bases and provincial levels on annual bases (Council Regulation EC, No 577/98).
The sample frame is a stratification of the primary units (municipalities) in each NUTS-2 (Region) based on the population of the municipality. The ultimate sampling units are the households and are drawn randomly from municipal registers. Stratification of the municipalities in each NUTS-2 region is based on the population of the communes. The households are rotated according to a 2-(2)-2 rotation plan. Households are interviewed during two consecutive quarters. After a two-quarter interval, they are again interviewed twice in the corresponding two quarters of the following year. As a result, each household is included in four waves of the survey.
In ISTAT LFS data are collected through computer assisted personal (CAPI) or telephone (CATI) interviews carried out by professional interviewers to Italian private households.
In general, data quality is based on several international performance criteria about how reduce the potential bias incurred by the deviation of a normal production process [6], [7], [8], [9] e [10].

* Email: lima@istat.it

Moreover, operational definitions and formulas to develop a harmonized approach for the calculation of these quantitative indicators are suggested by international dispositions [21], [22]. According to these dispositions, some key indicators are produced and monitored on a regular basis for the assessments of non-sampling error impact on survey estimates allowing a continuous quality check at interviewer's level [15]. In fact, a crucial step is detecting LFS data that do not conform to a well defined notion of normal behavior, or conform to a well defined notion of outlying behavior, though it is typically easier to define the normal behavior. Considering the possible detection of production process anomalies, also referred to as outliers, the problem is that defining a representative normal region is challenging and the boundary between normal and outlying behavior is often not precise. Moreover no standard indication is given about how the cut-off threshold should be chosen to the definitions of the tolerance bounds that determines the normal operating region.

In the absence of a prior threshold, a common practice in ISTAT household sample surveys is to identify the key indicators that are usually calculated in the other ISTAT household surveys as benchmarks in LFS production process monitoring. Such benchmark data sets allow a standardized comparative evaluation of outlier detection techniques in ISTAT and hence are very useful. But often the lack of such benchmark data sets has forced researchers to perform some statistical techniques.

Most outlier detection rules for multivariate data are based on the assumption of symmetry of the underlying data distribution although this assumption is frequently violated in practice.

Recently, more attention has been given to the application of control charts.

Their use is suggested to measure the stability of a statistical process of data production by means of process variables. ISTAT have used the control charts to identify variation for each interviewer in telephone (CATI) interviews [12].

This paper recovers an attempt to define the normal operating situation for evaluate the LFS interviewer performance as well as their impact on survey quality based on calculation of percentiles [11]. But its novelty is a definition of tolerance intervals based on average weighs all past observations equally to model information on the production process in time series like those of the ISTAT LFS continuous survey, carried out during every week of a year from 2006: The proposal can be classified both into non parametric and statistical model-based approach, since it is based on local density estimation using kernel functions. It is demonstrated the effectiveness of this time-based algorithm by using a number of synthetic and real data sets throughout the LFS production process.

The remainder of this paper is organized as follows. Section 2 presents a brief review on tolerance interval for outlier detection. Section 3 contains a short description of LFS data used in ISTAT monitoring production process. Section 4 illustrates the proposed method and section 5 concludes.

2. A Brief Review on Outlier Localization using tolerance interval

Several techniques have been formulated for outlier detection in literature[1]. The underlying principle of any statistical approach for outlier detection arises from the following definition - *An outlier is an observation that lies very far from the bulk of the values of data in either direction* [3], [19]. In other words, measurements not falling within the "normal" ranges or *tolerance intervals* may indicate the presence of an abnormality. David (1981) stated that a tolerance interval is similar to a conference interval in that "*a tolerance interval has random terminals, say L and V* "(p. 16) [4]. But David also pointed out that the two differ in that "*whereas a confidence interval is designed to cover, with prescribed probability, a population parameter such as mean, variance, or a quantile, the requirement of a tolerance interval (L, V) is that it contains at least a proportion* γ *of the population with probability* β *(with* $0 \le \beta$ *and* $\gamma \le 1$*)*" (p. 16) [4].

In practice, a tolerance interval is essentially a confidence interval on a specified proportion of a population distribution. An upper tolerance limit is a number such that there is a specified level of confidence that a specified proportion of the population has values at or below that number. Thus, if *f(x)* is a continuous probability density function (*pdf*), the tolerance interval $L(\gamma,\beta)=(L,V)$ is such that:

$$\Pr\left\{\int_{L}^{V} f(x)dx \ge \beta\right\} = \gamma \qquad (1)$$

Suppose now that $x_1, x_2, ..., x_n$ is a random sample from a distribution with continuous pdf *f(x)*. Let *L=X(r)* and *V= X(s)*, where *X(r)* and *X(s)* denote the *rth* and *sth* order statistics, respectively, from this sample and $s > r$. Then $L(\gamma,\beta)$ reduces to:

$$\Pr\left\{F(X_s) - F(X_r) \ge \beta\right\} = \gamma \qquad (2)$$

Further if *f(x)* is a $N(\mu,\sigma^2)$distribution with both μ and σ^2 being unkown, Wald and Wolfowitz [24] proposed a means of constructing a tolerance interval with, that is:

$$\begin{cases} L_1(\gamma,\beta) = m - zs \\ L_2(\gamma,\beta) = m + zs \end{cases} \qquad (3)$$

[1] See Ben-Gal I., Outlier detection, in: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

where $z = \sqrt{\frac{n-1}{\chi^2_{n-1,\gamma}}}$, m is the sample mean, s^2 is the sample estimate of the variance and $\chi^2_{n-1,\gamma}$ is the $(1-\gamma)$ quantile of chi-quadrato distribution with (n-1) df with the property that $\Pr(\chi^2 > \chi^2_{n-1,\gamma}) = \gamma$.

Although traditionally the normal distribution has been used as the target distribution, there are several non-parametric methods available for estimating tolerance intervals [2].

Within the class of non-parametric methods, tolerance interval can perform very well if a proper bandwidth is chosen [1], [2] e [4].

3. LFS Indicators System

The proposed method is based on the quantitative indicators which are used for monitoring the performance of about 300 interviewers of the CAPI network in each Italian region for the 4 quarters of the years from 2006 to 2010.

In each quarter of the year, it is possible to have details about these indicators for each one of the 13 weeks (*reference weeks*) when the interviews of the sampling households are carried out.

Some of the most important indicators, according the AAPOOR standards [21], are:

The Coverage Rate (CR)

It is the number of complete interviews divided by the number of interviews planned to achieve the precision of the designed survey estimates. The higher the coverage rate, the lower the sampling error for the survey.

The Proxy Rate (PR)

It is the number of households with at least one member for whom a proxy interview (any adult in the household to provide information for other household members) had to be taken in relating to the number of interviewed household. Proxy responses do not appear to substantially impact on key labour force data produced from the survey. However, some LFS data items may be affected such as working hours and wages, detailed classifications of industry and occupation.

The Refusal Rate (RR)

It is the proportion of all cases in which a household unit is not interviewed for refusal or break-offs of all potentially eligible cases. Non-respondents may have different characteristics on average compared to those of respondents to the survey. Depending on the level of non-response and the extent that differences exist, the survey estimates could be biased toward the characteristics of survey respondents.

The Substitution Rate (SR)

It is the number of replacements of an originally selected household by a substitute household relating to the number of interviewed households. Implications for data quality are a reduction of sample size and a higher non response rate. The literature is not in favour of substitution option, at least not with probability samples [15], [23]. Substitution is even recommended when the non response rate exceeded 35% [23].

In 2010, the average CR was around 95% while SR was around 27% (table 1).

Table 1. Statistical characteristics for CAPI fieldwork rates. Year 2010.

Field Rate	Mean	Dev Std	Min	Max
CR	94,8	10,8	22,9	100
SR	27,3	26,3	0	100
PR	29,1	22	0	100
RR	5,8	9,9	0	54,2

4. A Kernel Density Approach

The proposed procedure operates in two phases:
1) building the probabilistic model which comprises of estimating the kernel density distribution of LFS quality indicators;
2) determining outliers where a test instance is compared to the model to determine if a value is an outlier or not.
Here the problem of detection of outliers for a classification of fieldwork is solved producing the simple rule IF – THEN which considers an unit to be "O.K." if the value of its field rates is within the tolerance interval and "not O.K." otherwise [12].
Its details are summarized below:

4.1 Phase I: building the probabilistic model

The first step performs density estimate. Since it does not make any assumption about the type of the density and the underlying data distribution of ISTAT LFS field rates are generally skewed and multimodal [11], a non parametric kernel estimate that allows the identification of the function density for observed distributions of field rates looks more promising.
Thinking of the large body of published literature on non-parametric density estimation approaches[2], the field rates' weighted kernel density estimate (KDEs) is calculated weighting observations with the quota of the interviews assigned to each interviewer to the total of interviews planned. So observations with higher weights have more influence in the computations. Let $(x_1, x_2, \ldots, x_n)$ be an *iid* sample drawn from some distribution with an unknown density f. The weighted kernel density estimate (KDE) of f is as follows:

$$f(x) = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} W_i \varphi_h (x - X_i) \qquad (4)$$

where $h > 0$ is a smoothing parameter (the *bandwidth parameter*); W is the quota of the interviews assigned to each interviewer to the total of interviews planned and:

[2] G. R. Terrell and D. W. Scott. Variable kernel density estimation. The Annals of Statistics, 20(3):1236-1265, 1992.

$$\varphi_h(x) = \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{x^2}{2h^2}\right) \tag{5}$$

is the standard normal density rescaled by the bandwidth. More crucial than the exact kernel choice is the choice of *h*: for a fixed kernel function, increasing *h* implies more smoothing, less variability and more bias in the density estimation. As the optimal value of *h* is unknown, approximations methods are required [17], [20].
The normal function is selected for the kernel function property that's the number of modes in the density estimate decreases monotonically as the value of the bandwidth parameter increases. This property is not, in general, shared with other kernels [17]. This means that the number of features in the estimated density is a decreasing function of the amount of smoothing.
A basic problem with this bandwidth selector is its sensitivity to outliers.
If $h \to 0$ and $nh \to \infty$, then the optimal bandwidth is:

$$h_{AMISE} = \left[\frac{1}{2\sqrt{\pi}n\int(f'')^2}\right]^{1/5} \tag{6}$$

in the sense that it minimizes the so-called mean integrated squared error index (MISE, Wand and Jones; 1993). Its expressions contain the second derivatives of the unknown density *f* being estimated, and so approximations are required (AMISE).
In this paper the estimated optimal *h* is in the sense that minimizes the AMISE and it is obtained using the Silverman's rule of thumb [17], [20]:

$$h = 0{,}9\min\left[\hat{\sigma}, (Q_3 - Q_1)/1{,}34\right]n^{-1/5} \tag{7}$$

where Q_3 and Q_1 are the third and first sample weighted quartiles, respectively and the weighted variance takes the form

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} W_i (X_i - \overline{X})^2}{\sum_{i=1}^{n} W_i} \tag{8}$$

based on the weighted sample mean.

4.2 Phase II: determining outliers

The next step is to determine if a given data instance is an outlier with respect to the model or not. The threshold is achieved by direct estimation of the percentiles of the field rates' weighted kernel density estimate and doesn't refer to any specific benchmark.
It is obtained only by the use of information on the LFS statistical production process in time series. A general analysis of time series of quarterly field rates plot shows a cyclic factor within the production planning process. The first week of the quarter is usually characterized by a

strong motivation of the interviewers which results in good performance in terms of response rate. The following weeks (time 2, time 3, time 4), the motivation is still high and the interviewer has learnt how to react to objection by the respondents, thus limiting the refusals.
Hence to adjust the tolerance interval to the cyclic factor, in a reference week of each reference quarter the cut-off is the value between: a) the percentiles of the fitted KDEs in the week of the quarter in the previous years and b) the simple average percentiles of the fitted KDEs of past data in time series[3].
In this is simple way, fieldwork data are smoothed, random variation are removed and trends and cyclic are shown. The adjustment relies on obtaining good results reducing the field rates' cyclic factor in the quarter due to the interviewer's performance refer to a reference week. In the forth week the best results are obtained.
From the 5th up to the 10th week tiredness starts, thus producing a decrease in response rate and an increase in non-response for end period. The last weeks of the quarter (time 11, time 12, time 13) the interviewers receive a motivation kick, as they want to close the quarter with good results, so the substitution mainly due to refusals or excess in attempts increases.

5. A Kernel Density Approach

In order to analyse the interviewers' behaviours, the weighted unvaried kernel density are estimated in each week of each quarter of 2010 and the density estimate for CR, SR, RR and PR are plotted (figures 1,2, 3 and 4).
Datasets used contained different percentage of outliers thus providing a diverse test and illustrating wide capabilities of the proposed framework. A dynamic analysis can also be carried out by taking into account data from different quarters.
In the 2th quarter the existence of two modes in figure 3, for example, suggests that there are two distinct groups: one composed of the "best" interviewers which results in good performance in terms of SR, and another consisting of the "worst". The first mode is much less pronounced than the second, which indicates that the two groups are not of the same size: there are relatively few "worst" interviewers, and distinctly more "best" interviewers. Further, the first mode is located just to the left of the value 10 on the X-axis, while the second is found at around 60.
Tolerance intervals are based on confidence limits, calculated using KDEs. Here, for synthesis sake, results are reported only for some weeks of the 4th quarter of 2010 and are presented in tables 2 through 5.
To give an overall summary of the performance of the method, it is evaluated: (a) the dimension of the data set, (b) the proportion of outliers in the data, including artificial outliers and known outliers (c) the degree of separation of the outliers from the rest of the data , i.e. 'outlyingness'.
One might question the performance of this procedure. In fact there is no ground-truth reference to validate the performance of kernel density estimation.
Here the validity of the method has been tested comparing its results to the benchmarks which are already used in ISTAT household's surveys.
The results show that the method performs efficaciously so it may be considered as an appropriate and simple method to use for the detection of outliers.

[3] The "simple" mean weighs all past observations equally.

Moreover, to give an overall summary of the performance of the method, it is evaluated: (a) the dimension of the data set, (b) the proportion of outliers in the data, including artificial outliers and known outlier (c) the degree of separation of the outliers from the rest of the data , i.e. 'outlyingness'.
However, further studies are needed to actually evaluate the performance of different approaches for computation of an optimal bandwidth.
This is a concern, however, because the behavior of the estimated tolerance intervals depends on this parameter. The claims made and issues raised above are presently being investigated and the results will be reported at a later stage.

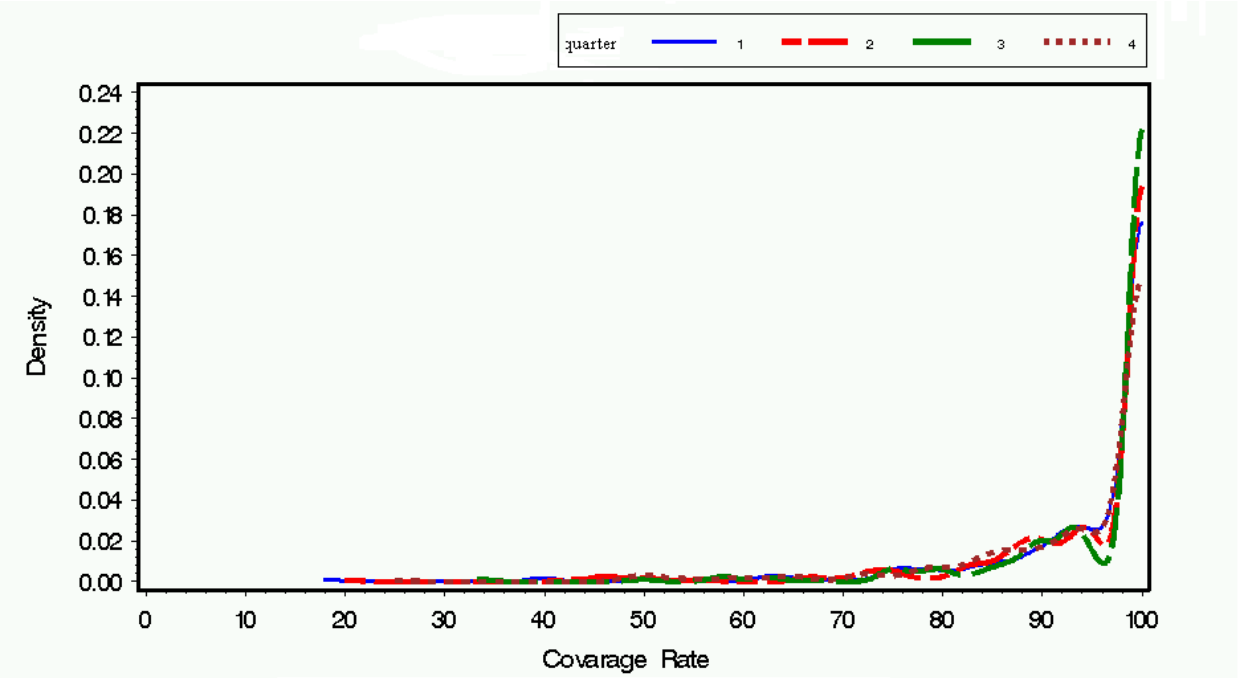

Figure 1. Density estimates of Coverage Rate by quarters and bandwidth.

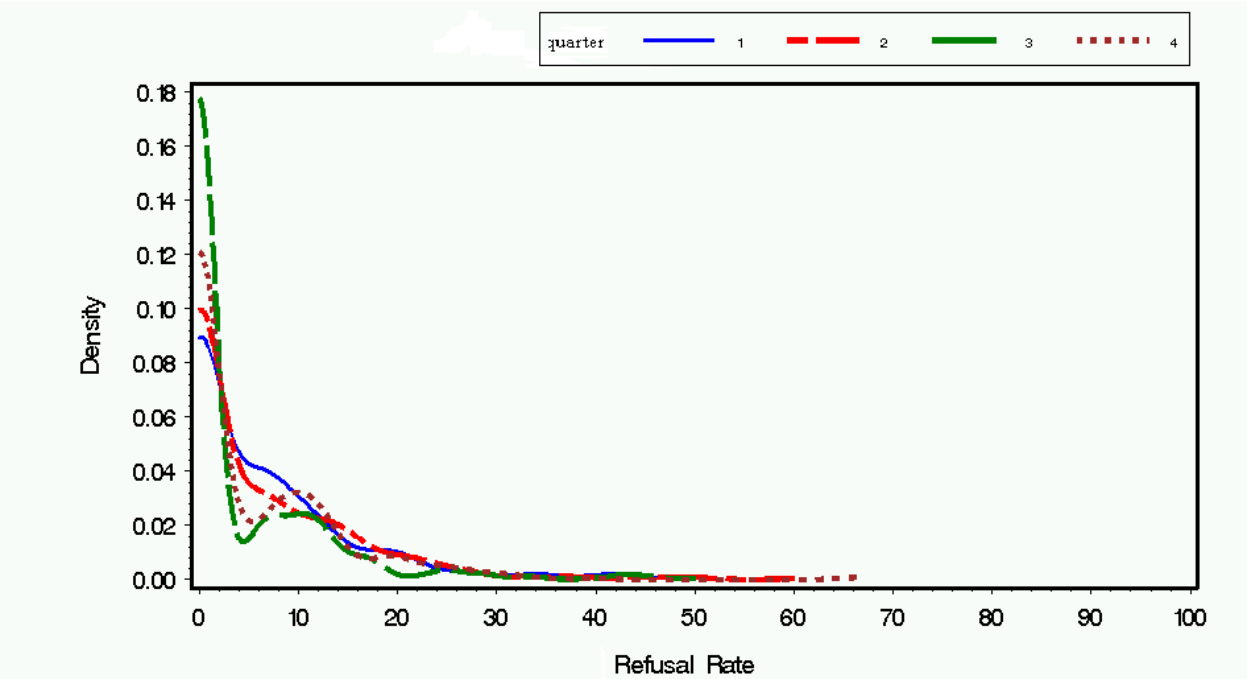

Figure 2. Density estimates of Refusal Rate by quarters and bandwidth.

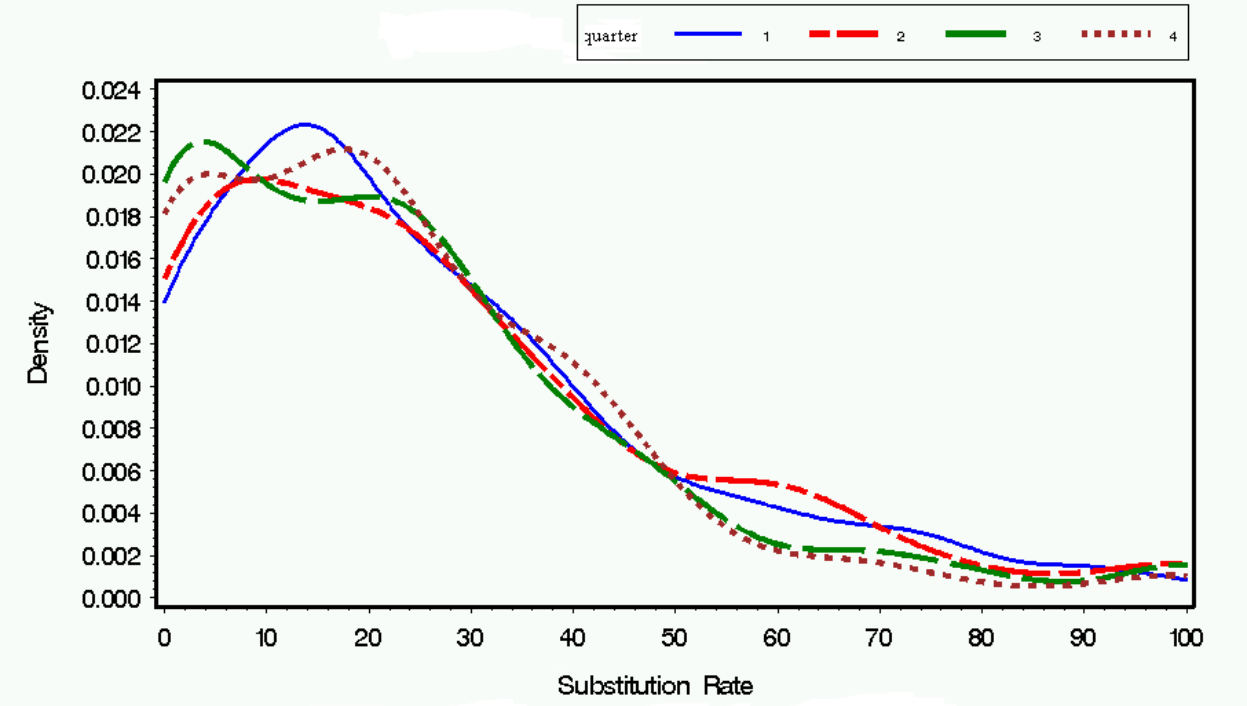

Figure 3. Density estimates of Substitution Rate by quarters and bandwidth.

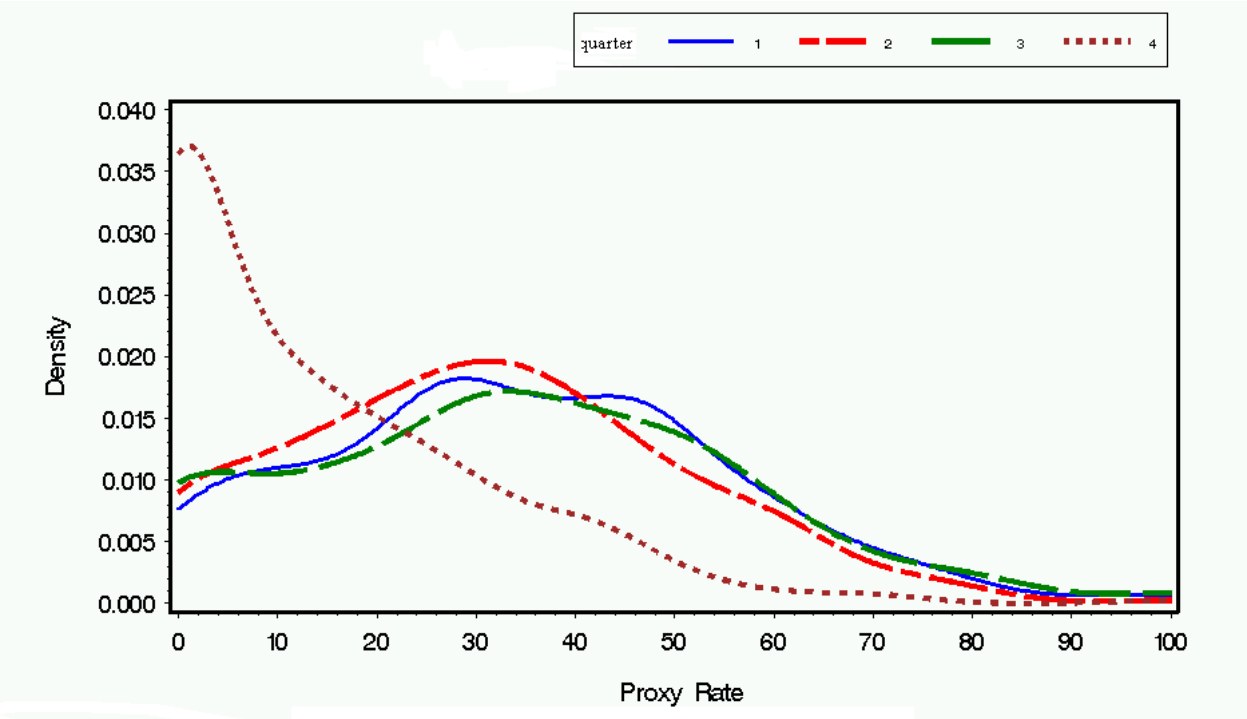

Figure 4. Density estimates of Proxy Rate by quarters and bandwidth.

Table 2. Tolerance limits for the 2th week in the 4th quarter of 2010.

Percentile Level	Percentile Estimate				Tolerance Interval							
					CR		RR		SR		PR	
	CR	RR	SR	PR	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	80	0	0	0	68	80	0	0	0	0	0	0
10	86	0	0	5	79	83	0	0	0	0	5	6
25	92	0	0	19	90	93	0	0	9	9	16	15
75	100	8	35	50	100	100	11	11	39	36	48	47
90	100	17	60	67	100	100	19	20	62	53	63	63
95	100	23	72	79	100	100	26	25	76	64	71	67

Table 3. Tolerance limits for the 6th week in the 4th quarter of 2010.

Percentile Level	Percentile Estimate				Tolerance Interval							
					CR		RR		SR		PR	
	CR	RR	SR	PR	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	77	0	0	0	64	75	0	0	0	0	0	0
10	86	0	0	5	75	85	0	0	0	0	3	6
25	94	0	8	15	90	94	0	0	9	9	18	17
75	100	9	33	46	100	100	10	9	37	36	53	46
90	100	17	50	65	100	100	19	18	57	54	68	63
95	100	22	67	75	100	100	26	23	76	68	76	73

Table 4. Tolerance limits for the 9th week in the 4th quarter of 2010.

Percentile Level	Percentile Estimate				Tolerance Interval							
					CR		RR		SR		PR	
	CR	RR	SR	PR	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	72	0	0	0	21	63	0	0	0	0	0	0
10	83	0	0	0	46	79	0	0	0	0	2	9
25	92	0	7	0	76	90	0	0	10	9	15	19
75	100	9	33	23	100	100	8	9	47	40	51	50
90	100	14	46	40	100	100	19	20	73	60	70	64
95	100	20	63	46	100	100	25	25	89	70	79	67

Table 5. Tolerance limits for the 12th week in the 4th quarter of 2010.

Percentile Level	Percentile Estimate				Tolerance Interval							
					CR		RR		SR		PR	
	CR	RR	SR	PR	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	76	0	0	0	38	69	0	0	0	0	2	6
10	82	0	0	0	56	75	0	0	0	0	3	12
25	92	0	5	12	82	90	0	0	10	10	17	24
75	100	8	38	44	100	100	10	12	46	40	54	50
90	100	20	55	63	100	100	22	20	71	58	72	65
95	100	27	67	70	100	100	28	27	85	69	81	75

References

[1]. Azzalini, A. 1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68, 326–328.

[2]. Baíllo A. and Cuevas A. (2006). Parametric versus nonparametric tolerance regions in detection problems. *Computational Statistics*, 21, 523-536.

[3]. Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*. J. Wiley & Sons, New York.

[4]. David, H. A. (1981). *Order Statistics* (2nd ed.), New York: John Wiley.

[5]. Di Bucchianico A., Einmahl J.H. and Mushkudiani N.A. (2001). Smallest Non Parametric Tollerance Regions. *The Annals of Statistics*, 29, 1320–1343.

[6]. Eurostat (2003). *Standard Quality Indicators-producer-oriented*, Document n° Eurostat/A4/Quality/03/General/Standard Indicators.

[7]. Eurostat (2002a). *Definition of Quality in Statistics*, Document n° Eurostat/A4/ Quality/02/General/ Definition.

[8]. Eurostat (2002b). *Standard Quality Report*, Document n° Eurostat/A4/Quality/ 02/Standard/Report.

[9]. Eurostat (2002c). *Glossary on Quality in Statistics*, Document n° Eurostat/A4/Quality/ 02/Standard/Glossary.

[10]. Eurostat (2002d). *How to make a quality report - version 1*, Document n° Eurostat/A4/Quality/02/General/How To Make a Quality Report.

[11]. Giuliani G., Lima R., Ranaldi R. (2011). Classification using Kernel Density Estimation: an application to the ISTAT's Labour Force Survey, in *The 8th Scientific Meeting of the Classification and data Analysis Group of the Italian Statistical Society*, Pavia, 7-9 September 2011.
[12]. Grassia M.G., Simeoni G. (2005). Strumenti per il monitoraggio del processo: l'uso delle carte di controllo - in Muratore M. G., Sabbadini, L.L. Signore M., *Il monitoraggio del processo e la stima dell'errore nelle indagini telefoniche. Applicazioni all'indagine sulla Sicurezza dei cittadini*, ISTAT Metodi e norme (cap. 4).
[13]. Hidiroglou, M.A. and Srinath, K.P. (1981). Some estimators of population totals from a simple random sample containing large units. *Journal of the American Statistical Association*, 76,.690-695.
[14]. Keats J. B. and Hubele N.F. (1989). *Statistical Process Control in Automated Manufacturing*. New York, M. Dekker.
[15]. Kish L. (1965). *Survey Sampling*. New York, Wiley Classics Library.
[16]. ISTAT (2006). http://www.istat.it/dati/catalogo/20060830_00
[17]. Jones, M. C., J. S. Marron, and S. J. Sheather. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*,91, 401–407.
[18]. OECD (2011). www.oecd.org/statistics/qualityframework
[19]. Rousseeuw P. J. and Leroy A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons, New York.
[20]. Silverman, B. W. (1986). *Density Estimation*. London: Chapman and Hall.
[21]. The American Association for Public Opinion Research (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition*. AAPOR.
[22]. The International Monetary Fund (2001). Toward a Framework for Assessing Data Quality. *Working Paper*, P/01/25.
[23]. Verma, V. and Gabilondo, L. (1993). *Family Budget Surveys in the EC: Methodology and Recommendations for Harmonization*. Eurostat, Statistical Document 3 E. Luxembourg:Eurostat.
[24]. Wald, A., and Wolfowitz, J. (1946). Tolerance Limits for a Normal Distribution. *Annals of Mathematical Statistics*, 17, 208-215.