ROBUST ESTIMATES OF INVESTMENTS FROM THE BANK OF ITALY'S BUSINESS SURVEY

Paola Battipaglia

Research Department, Banca d'Italia

Via Nazionale 91, 00184 Rome, Italy*

April 2000

* E-mail address for correspondence: battipaglia.paola@insedia.interbusiness.it

Abstract

Genuine (representative) outliers are common with many variables in business surveys because of the natural variability of the underlying population distributions. Traditional type estimators (like the sample mean) are not robust to extreme values in the sample and may therefore grossly misrepresent the true population parameter.

Most theoretical work on robust estimates of location focuses on small departures from symmetric distributions. Some recent applied research on asymmetric distributions has introduced winsorisation as an effective and flexible tool for reducing the mean square error of the sample estimators of location in the presence of a few representative outliers.

We explore the impact of winsorisation on the estimates of the annual rate of change in industrial investments, derived from the Bank of Italy's annual business survey. Results are checked by a simulation experiment, testing two versions of winsorised estimators and ten different cut-off thresholds for each one.

KEYWORDS: Outliers; Survey estimates; Robust statistics; Winsorised estimators.

Introduction

In all surveys, outliers in the collected data represent a serious problem, as they affect the reliability of the inference which is made from the sample to the population. While some of the extreme data in a sample are identified at the editing stage as measurement or coding errors, others are genuine occurrences from a skewed variable. This type of "representative" outliers, which are common in business surveys, convey relevant information about the non-sampled units of the population and therefore require attention.

Traditional type estimates (like the sample mean) are not robust to extreme values in the sample. That is, they may be mislead by a few, or even just one, unusual observations and grossly overestimate or underestimate the true population parameter. To reduce the impact of these observations, winsorisation is a method based on the intuitive approach of adjusting down (or up) any sampled value greater (or smaller) than a cut-off. By allowing for different choices of the cut-off value and the outliers treatment, this method makes it possible to carry out different strategies and amounts of intervention. Some recent experience at the United Kingdom's Office for National Statistics and at the Australian Bureau of Statistics show that this method can be quite effective in producing robust estimates and deserves further research.

This paper explores, with the help of a simulation experiment, the impact of winsorisation on the estimates of the annual rate of change in industrial investments, derived from the Bank of Italy's annual business survey. Section 1 introduces the survey, the analysis variable and the classic estimator with its standard errors; section 2 is a brief overview of robust statistics developed in the context of survey estimates; section 3 proposes two winsorised estimators. Finally, sections 4 and 5 assess the performance of the two robust methods and their behaviour with different cut-offs. Conclusions are drawn in section 6.

1. The survey and the estimation problem

1.1 The Bank of Italy's business survey

Since the seventies, the Bank of Italy has been carrying out a business sample survey, focused on the medium and large industrial firms. The annual sample is made of a panel of about 1,000 firms, covering approximately 10 per cent of the corresponding population. The questionnaire collects information on more than 200 variables. The bulk of it, which is kept constant across different waves, covers structural features - like firms' ownership and employment structure– as well as short term dynamics for employment, investments, sales, prices and productivity; the financial section investigates the various financing sources and possible bank credit constraints. A further set of questions is changed every year to provide information on up-to date research issues. Interviews are personal, and are carried out by the Bank's personnel working in the local branches.

Quality checks on the collected data are carried out at different levels: at the time of the interview, by comparison with information given by the same firm in the previous wave; at the data entry stage, by checks on the range and other bounds –like totals; during the data processing, by more in-depth checks on longitudinal and cross-time consistency.

1.2 The analysis variable and its sample estimate

One of the statistics of main concern which is routinely estimated from the survey is the annual rate of change in investments, for the total survey population as well as for some strata such as geographical area and employment size bands. The sample estimate is calculated as the ratio between the two estimated annual totals :

$$r = \frac{\sum w_i^s I_i^t}{\sum w_i^s I_i^{t-1}} \qquad (1)$$

where:

I_i^t, I_i^{t-1} are the amounts of investments of the firm i at times t and t-1,

$w_i^s = \frac{N_h}{n_h} \ \forall \, \text{i} \in \text{stratum h}$ is the sample expansion factor, defined as the ratio between the number of firms in the population (N_h) and the corresponding number in the sample (n_h)[1].

The estimate can be alternatively viewed as the weighted average of the individual rates of change in the sample[2]. Here weights become twofold: they represent the product between the expansion factor w_i^s -accounting for the sampling design- and a leverage component, which is equal to the amount of investments at time t-1. That is:

$$r = \frac{1}{\sum w_i} \sum \frac{I_i^t}{I_i^{t-1}} w_i \qquad (2)$$

where:

$w_i = w_i^s w_i^q$ is the global weight,

$w_i^q = I_i^{t-1}$ is the leverage factor,

$w_i^s = \frac{N_h}{n_h}$ is the sample expansion factor.

[1] It will be noticed that weights are kept constant for the two years involved in the ratio. This is due to the lack of updated information on the distribution of firms in the population and relies on the fact that both I^t and $I^{t-1}$are measured by the same interview.

[2] A problem arises as regards the observations with zero investments at time t-1: see further on, section n.3.

Consistently across time, both the distributions of annual amounts and relative changes in investments are positively skewed, with a long right tail along which a few extremely large observations lie. Outliers tend to occur in the right tail because every year there are a few firms which make exceptionally large investments; on the left side, there is a small mass with zero investments and occasionally very few negative values, coming from disinvestments. Hence, the zero value represents in practice the left bound of the two distributions. The distributions of relative changes, as they can be reconstructed from the grossed-up sample, are reproduced for the last six years in figure 5.

Since the population variability is huge, the standard error of the traditional sample mean tends to be very large for the estimates of both totals and percentage rates of change in investments. Particular concern has recently arisen for the estimate of the rate of change at the geographical area level: in Southern Italy, where the investments are usually more erratic (also because of the public incentives occasionally introduced to stimulate the local economy), the standard error of the traditional estimator is around ten. To reduce the magnitude of this error, the sample design has been changed starting from the current wave and it has been planned to almost double the number of sampled businesses located in Southern Italy. This intervention is expected to reduce the standard error by one half[3]. However, the natural variability of the analysis variable is such, as to suggest that traditional sampling practice needs to be integrated with a robust estimation procedure to ensure that each particular sample be protected against the unduly large influence of a few extreme observations.

[3] Under a stratified design, the contribution of each stratum to the total variance of the estimator is proportionate to the stratum variance times the stratum weight in the population. Oversampling from the strata where the analysis variable has the largest variance is therefore an efficient way of reducing the standard error of the final estimator.

2. Robust statistics

2.1 Representative outliers and their impact on traditional sample estimates

Sample outliers, that is observations far away from the bulk of the sampled data, may arise from different sources, like genuine variability in the behaviour of the units in the population, measurement errors or coding errors. A useful categorisation of outliers (Chambers, 1986) distinguishes between:

- representative outliers**,** which are sample elements with values that have been correctly recorded and which can therefore not be assumed to be unique in the population;
- non-representative outliers, which are sample elements with values incorrect or unique in some sense.

In the process of analysing survey data, non-representative outliers are usually detected and corrected during the editing, while representative outliers are used in the estimate as they are assumed to convey relevant information about the underlying population.

In standard design-based methodology, each observation in the sample is grossed-up by the reciprocal of the selection probability (or sampling fraction) to obtain the so called 'expansion estimator'. When a particular sample contains a few –or even just one- outlying observations, this method can produce an estimate which is very far from the true value, although the procedure remains unbiased in the long run. Getting a more stable estimate requires finding some way to bound the influence of these extreme data. This is a task of great concern to both the survey statistician and the analyst. "The argument that the [traditional expansion] procedure is unbiased falls on deaf ears. The client is not interested in what happens in the long run –he wants an estimate as close to the population parameter as possible for that particular case, and may even feel that a

better estimate would be obtained if the offending observations were discarded" (D.T. Searls, 1966).

Reducing the impact of extreme data on sample estimates is usually obtained by either deflating their weights or modifying their values[4]. Determining the appropriate weights can be difficult, especially if there are relatively few outliers (see P. Smith, 1997, Hidiroglou & Srinath, 1981). In practice, the outliers usually receive unit weights, under the implicit assumption that they are unique in the population. Although this approach is handy, it presents some clear drawbacks. It does not deal properly with observations which are extreme with regard to only a subset of the analysis variables. Moreover, it makes no attempt to use the information in the sample to estimate the contribution of the non-sampled outliers to the population parameter (see Chambers, 1986).

2.2 Winsorisation as a tool to deal with representative outliers

The most well known technique for value-adjustment of outliers is Winsorisation. Winsorised estimators reduce the influence of outliers by adjusting down towards a cut-off any sampled value greater than the cut-off. This can be done in two ways: by replacing every outlier with a value that is exactly the cut-off, or by changing them to an intermediate value between their original one and the cut-off. In the literature, these two kinds of estimators are called **type I** and **type II** Winsorised estimators.

In **type I Winsorisation**:

$y^*_i = y_i$ for any $y_i < K$ (where K is the cut-off)

$y^*_i = K$ otherwise.

[4] The well-known trimmed mean, which involves complete rejection of all the observations which exceed the cut-off, can be regarded as an extreme version of the first method, where the values identified as outliers are allocated zero weight.

In **type II Winsorisation** the choice of the weights can vary. For example, it is possible to set the weight to represent the (overall or per stratum) sampled fraction (f = n/N); in this way, the "credit" which is given to each observation depends on its representativeness according to the sample design. The rule then becomes:

$y^*_i = y_i$ for any $y_i < K$ (where K is the cut-off)

$y^*_i = f\, y_i + (1-f)K$ otherwise.

With this choice, observations coming from totally enumerated strata receive full weight.

Under type II framework, Winsorisation can also be extended to ratio, regression and generalised regression estimation (Kokic-Smith, forthcoming).

For skewed populations, where outliers are concentrated in one tail of the distribution, winsorisation is usually one-sided. The amount of bias introduced by treating only one tail of the distribution is the price to be paid to achieve protection against very large standard errors. Some recent experience with winsorisation in the context of business surveys (mostly focused on the estimate of totals) has shown that, for a wide range of cut-offs, the amount of bias introduced is negligible, while the standard error reduction is of considerable magnitude (see, for example, Kokic & Bell, 1994).

To account for the possible bias, winsorised estimators are usually compared with the traditional method by a quadratic loss function incorporating both the bias and the standard error. The most common global error indicator is the mean square error:

$MSE = Bias^2 + (\text{Standard error})^2$.

3. Winsorisation applied to the Bank of Italy's business survey

Like in many business surveys, the Bank of Italy's sample estimates can be crucially affected by some extreme (usually very large) data in the sample. This is especially worrying for particular domains, where the sample is more sensitive to outliers. After a careful screening at the editing stage, the Bank's current practice to deal with "true" outliers is to allocate them unit sample weights. Recently, in the presence of extremely large values concentrated in very small sub-samples and specific variables, trimmed means have also been used in official publications.

The purpose of this piece of research was to find a possible more refined, still simple tool to be regularly used in the analysis and editing of the survey data. In an initial stage, alternative approaches were tried, including transformations in logarithmic, reciprocal and square root scale. This was in order to achieve symmetry and hence use the results of the asympthotic theory based on these assumptions. However, the results were not straightforward (negative and zero values requiring special treatment) and non satisfactory, as symmetry was not fully achieved in any of the new scales. Unhappy with these findings, we decided to build on recent positive experience with winsorisation, which has the appeal of an intuitive approach providing a simple and flexible tool.

This section develops two types of winsorised statistics, which could be implemented on the survey data to produce robust estimates of the rate of change in investments. Section 4 will test their performance by a simulation experiment.

The need to deal with an estimate of relative change posed a preliminary problem as for the choice of the variable to be treated: independent winsorisation of the two estimates of totals, as it was tried in a first approach, would not guarantee any result in terms of the target statistic. It was therefore decided to work on the distribution of changes directly, comparing two different approaches: one-sided winsorisation of the largest values in the skewed distribution of relative

changes $r_i = I_i^t / I_i^{t-1}$ and two-sided winsorisation of the outlying positive and negative values in the distribution of absolute changes $D_i^t = I_i^t - I_i^{t-1}$.

Before starting with any procedure, it was necessary to decide whether to exclude or transform the ratios with zero investments in the base year. It was opted for the replacement of all the zeros with an extremely small value, set equal to 0.1 million liras: this allowed all the observations in the analysis to be retained. Consistency of the main results was then checked over the subset of defined ratios.

The remainder of this section explains the two methods in detail.

3.1 One-sided winsorisation

Many economic variables–such as investments, sales or stocks, are positively skewed. That is, outliers are concentrated in the right tail, which is therefore very long and thin. To produce robust estimates of means or totals for these distributions winsorisation is usually one-sided; that is, only the largest observations are identified and treated as outliers. Such a method obviously introduces a downward bias, whose magnitude must be checked against the reduction gained in the standard error (e.g. by a combined loss function; see section 2.2).

In the context of the Bank of Italy's survey, one-sided winsorisation was applied to the distribution of individual changes in investments: $r_i = I_i^t / I_i^{t-1}$, to obtain a more robust version of the target statistic expressed as under (2).

The presence of outliers in the leverage component of the weights required their preliminary treatment to avoid overweighting some observations. Winsorisation was therefore carried out in two steps, which are described in the following.

Step 1: preliminary treatment of outliers in the weight component

The auxiliary variable "number of employees", which is measured by the survey with little error, was used to identify extreme observations in the distribution of I^{t-1}. Since the distribution of the ratio: $\text{Investments}_{(t-1)}$/ $\text{Employees}_{(t-1)}$ looks positively skewed, (see figure 6) one-sided type II winsorisation was carried out with a fixed cut-off corresponding to the 99th percentile of the distribution of the ratios, weighted by the expansion factor w_i^s only.

Type II winsorisation was implemented according to the formula:

$$I_i^{t-1(wns)} = I_i^{t-1} \qquad \text{for any } I_i^{t-1} < K \text{ (where K is the cut-off)}$$

$$I_i^{t-1(wns)} = f_h I_i^{t-1} + (1 - f_h)K \qquad \text{otherwise, h corresponding to the employment-size strata.}$$

At the end of step one, $I_i^{t-1(wns)}$, i.e. the winsorised version of I_i^{t-1}, replaced for each observation both the weight component w_i^q and the denominator of the relative change r_i. This modified r_i was named r_i^{wns1}.

Step 2: winsorisation of the weighted distribution of relative changes

Step 2 was concerned with the analysis of the distribution of the annual rate of change: $r_i^{wns1} = I_t^t \,/\, I_i^{t-1(wns)}$, with winsorised weights: $w_i^{(wns)} = w_i^s I_i^{t-1(wns)}$. The positive skewness of this distribution suggested that only the extremely large observations should be treated, as before. The final estimate of the ratio became:

$$RW2 = \frac{1}{\sum w_i^{(wns)}} \sum \frac{I_i^t}{I_i^{t-1(wns)}} w_i^{(wns)}$$

3.2 Symmetric winsorisation

Although in some contexts the bias introduced by one-sided winsorisation has been proved to be negligible (see for example Kokic and Bell, 1994), it is obviously an unpleasant feature, which makes the procedure much more sensitive to the choice of the cut-offs. Moreover, even when used on a single variable, where the trade-off between bias and variance is favourable, the induced bias can severely affect the mean square error performance of derived variables (Kokic and Smith, 1998). In this respect two-sided winsorisation, simultaneously dealing with upper and lower outliers, is more appealing.

The second approach applied two-sided winsorisation to the distribution of absolute annual changes $D_i^t = I_i^t - I_i^{t-1}$, weighted by the expansion factor w_i^s, which accounts for the sampling design. Each employment size stratum was treated as a separate distribution, in order to avoid mistaking for outliers the movements of the largest firms. The resulting winsorised difference $\left[D_i^t\right]^{wns}$ was finally used as the numerator of the winsorised ratio:

$$DW1 = \frac{\sum w_i^s \left[D_i^t\right]^{wns}}{\sum w_i^s I_i^{t-1}} + 1 \ .$$

3.3 Sensitivity of the results to the cut-off value

When a bias is introduced by asymmetric treatment of outliers, the overall performance of winsorised estimators can be very sensitive to the choice of the cut-offs. The issue of finding a mathematical expression for optimal cut-offs (i.e. those minimising the total mean square error of

the corresponding winsorised estimator) has been addressed so far for the estimation of totals, see for example Kokic&Bell, 1994, Cruddas&Kokic, 1996.

The cut-off selection can rely on previous survey information or refer to some regression models. While the first strategy requires a preliminary robust endogenous estimate for the stratum mean μ_h (which can be obtained, for example, by pooling previous repeats of the survey), generalised regression estimation replaces the (stratum) mean with the expected value under the estimation model.

As a preliminary exploratory analysis, this paper addressed the general issue of the sensitivity of the results to the choice of the cut-offs. The performance of both one-sided and two-sided winsorisation was checked under a set of values, ranging from the first to the tenth percentile of the weighted empirical distribution. Referring to the mean square error as the indicator of global performance, the shape and slope of the curve defined by the values corresponding to these different cut-offs were examined.

4. Evaluation by a simulation experiment

The main properties of sample estimators -such as their precision and variability, cannot be defined unless their sample distribution is known. Information on it is usually derived either from asymptotic theory, or from simulation experiments, which generate repeated independent samples under the same essential conditions of the real survey. When the estimators do not have a known distribution, the simulation experiment is more appropriate because its flexibility allows the approximate reproduction of any kind of estimate. For these reasons, a simulation experiment seemed most appropriate in this context: by repeatedly sampling from a large data base of balance sheets identified as the parent population, the bias, standard error and mean square error of the winsorised estimator could be computed and compared with those of the traditional mean.

The main features of the simulation experiment are described here; section 5 provides the results.

4.1 Choice of the population

According to the last Census on Industry – 1996 - the target population of the Bank of Italy's Survey on Investment consists of about 11,000 businesses.

With one year lag, the balance sheets of about 80 per cent of this population are collected in the data base 'Centrale dei Bilanci'. Although individual matching between the survey sample and this data base may be poor (mainly because of the different timing at which information is collected), the distributions share similar features. Both are positively skewed with a long right tail; the percentiles are very close, especially the medians. On the contrary, the mean values tend to be very different, particularly by geographical areas. This is not surprising, since the sample estimators are known to have very large standard errors at this level.

On the whole, relying on these findings, the data base 'Centrale dei Bilanci' was judged a suitable reference population for the simulation experiment. In order to make its size as large as the survey target population, the gap was filled by randomly reproducing at the stratum level as many units as missing. In this way, it was possible to generate samples which matched both the actual sample size and the actual sampling rate.

4.2 The repeated sampling process

The simulation experiment consisted in the generation of one hundred independent samples per each of the last three available years 1996-1998, reproducing in size and features the survey design. As in the real survey, the population had been stratified into six employment size bands: the

last stratum –corresponding to the largest firms, was completely enumerated, while from the first five strata units were drawn at random with selection probabilities equal to the achieved sampling rates (i.e., the sampling fractions at the design stage times the average response rate).

Before carrying out the estimation procedure, each unit in each sample was allocated an expansion weight: for the sake of simplicity, these weights were fixed as the reciprocal of the selection probabilities[5].

5. Results

The simulated samples allowed for a comparative assessment of the performance of the simple expansion estimator and of its two winsorised versions. The overall loss function was set equal to the mean square error [MSE=bias^2+variance]: this was calculated for each of the three years and ten cut-off levels.

While winsorisation was carried out only once for each sample – at the total distribution level - the sample means of the winsorised distributions were calculated also over critical sub-samples. Particular attention was paid to the effects of winsorisation at the area level, where the standard error of the traditional estimator is quite large. The following table presents the results obtained by applying winsorisation to the extreme two per cent of the distribution[6]. R is for the traditional estimate, calculated without any intervention on the outliers; RW2 is the result of asymmetric outlier treatment; DW1 is the value under two-sided winsorisation.

[5] The real survey weights are more accurate and at the same time more variable: in fact, they use the firm's industry and geographical location as second-level stratification variables and, for each stratum, the actual number of interviews instead of the ex-ante selection probabilities.

[6] As regards RW2, two per cent refers to step two, while at step one the threshold remained fixed at the 99th percentile.

	Bias *(percentage points)*				Standard error *(percentage points)*				Mean Square Error *(percentage points)*			
	1996	**1997**	**1998**	**Mean(*)**	**1996**	**1997**	**1998**	**Mean**	**1996**	**1997**	**1998**	**Mean**
ITALY												
R	0.06	0.01	0.12	0.06	3.52	4.32	3.81	3.88	0.12	0.19	0.15	0.15
RW2	6.22	5.15	3.34	4.90	3.11	3.16	3.54	3.27	0.48	0.36	0.24	0.36
DW1	0.31	0.24	-0.33	0.29	3.16	3.39	3.31	3.28	0.10	0.12	0.11	0.11
NORTH												
R	0.18	0.21	0.17	0.18	3.55	3.95	3.72	3.74	0.13	0.16	0.14	0.14
RW2	5.72	5.01	3.69	4.81	3.19	3.09	3.44	3.24	0.43	0.35	0.25	0.34
DW1	0.45	0.22	-0.06	0.25	3.15	3.39	3.34	3.29	0.10	0.12	0.11	0.11
SOUTH												
R	-3.65	-6.72	-3.51	4.63	17.87	27.51	23.56	22.98	3.33	8.02	5.67	5.67
RW2	13.63	5.95	-2.40	7.33	11.46	15.01	19.60	15.36	3.17	2.61	3.90	3.23
DW1	-3.31	-1.35	-5.68	3.45	15.55	16.36	16.00	15.97	2.53	2.69	2.88	2.70

(*) Average of the absolute values.

At the total sample level, a modest intervention appears to produce very modest results. Under one-sided winsorisation, the mean square error is indeed even larger than with the traditional mean: this is because the bias introduced by the asymmetric procedure offsets the gains achieved in terms of standard error. Two-sided winsorisation does not raise any problem in terms of bias, while it achieves some gains in terms of variability. A very similar pattern of error is identified for all the

estimators over the subsample of Northern Italy, which is in fact very close to the total sample in many respects.

The last section of the table presents the errors associated with the estimates obtained by calculating the mean rate of change of the Southern firms over the winsorised distribution. Here, where the performance of the traditional estimator is very poor in terms of standard error, winsorisation achieves very good results. On the average of the three years, the mean square error is cut down by 40 and 60 percentage points by the one-sided and two-sided approach respectively; in 1997, where the largest gain was achieved, the mean square error drops from 8 to less than 3 percentage points.

The bias introduced on specific sub-samples by one-sided winsorisation of the total distribution was unpredictable: it is therefore not surprising that both the sign and magnitude of the bias change considerably across years. Two-sided winsorisation was not expected to introduce any systematic bias, and in fact in this regard it performs very similarly to the standard mean[7].

The following two figures allow for a visual assessment of the effect of winsorisation on the estimated rate of change in investments for Southern firms. The black bar to the left represents the 'true' percentage change (i.e., the one worked out by using the whole data set "Centrale dei Bilanci"). Next to it are the estimates obtained by carrying out no outlier treatment ("R"), one-sided winsorisation ("RW2") and two-sided winsorisation ("DW1"); the heights of these three bars correspond to the average value of the estimate computed over the one hundred simulated samples. The width of the vertical bands is twice the standard error. The two charts differ in the amount of data treated by winsorisation: two and ten per cent of the distribution respectively. The effect of winsorisation in terms of standard error reduction appears quite remarkable, the more the greater the percentage of treated outliers.

[7] Neither R or DW1 is expected to be biased. The amount of bias found in this experiment is probably due to the relatively small number of replicated samples.

Figure 1

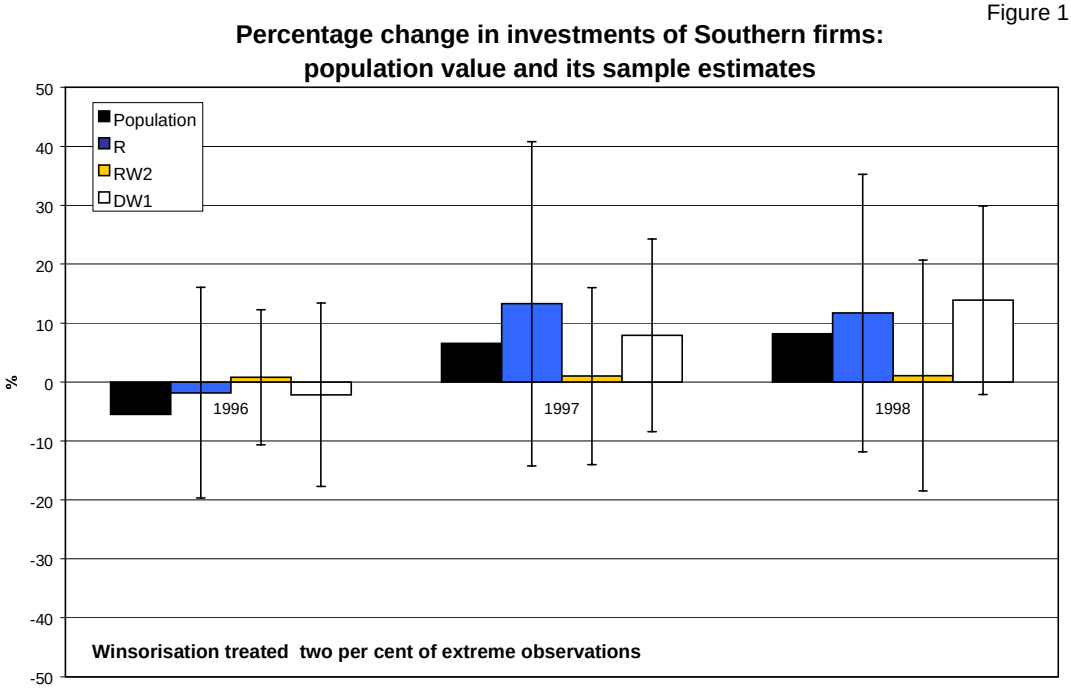
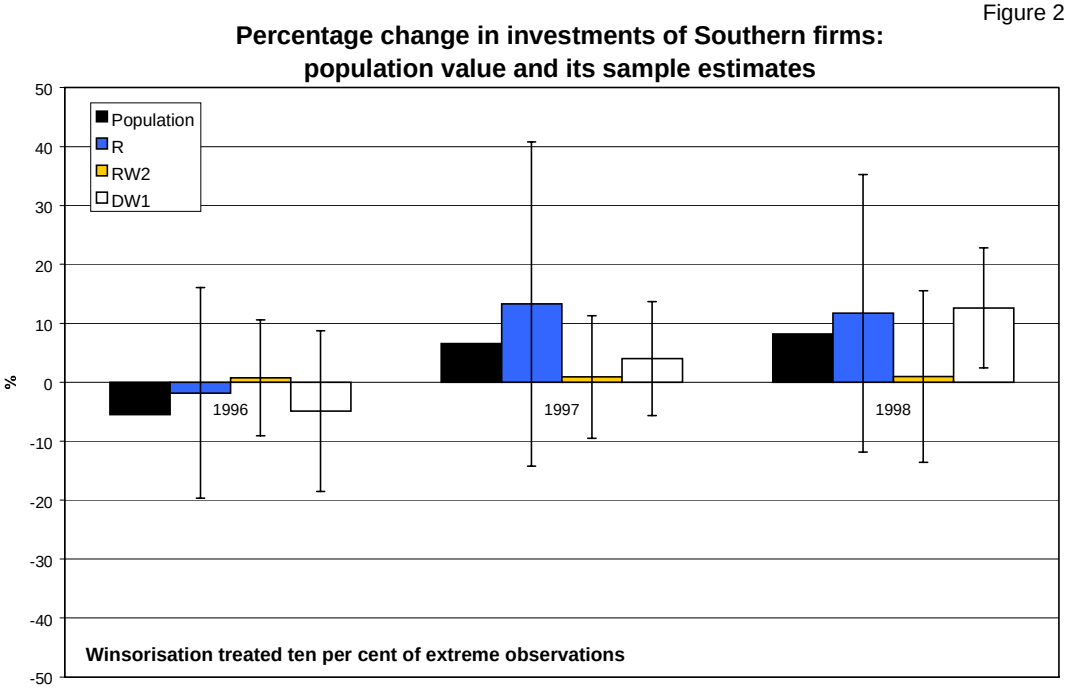


Figure 2



The behaviour of the total error and its components with respect to different cut-off values was examined by the shape and slope of the corresponding curve. The plots for individual years

1996-1998 are presented in figures 7-12. Since time-consistency proved pretty strong, comments will be here focused on average results.

The two pictures which follow give the mean square error at the total sample level and for the sub-sample of Southern firms as the average over the three years. The cut-offs, corresponding to the first ten percentiles of the total distribution, are displayed along the x axis; the traditional estimate R is marked as the origin (cut-off equal to zero).

Figure 3

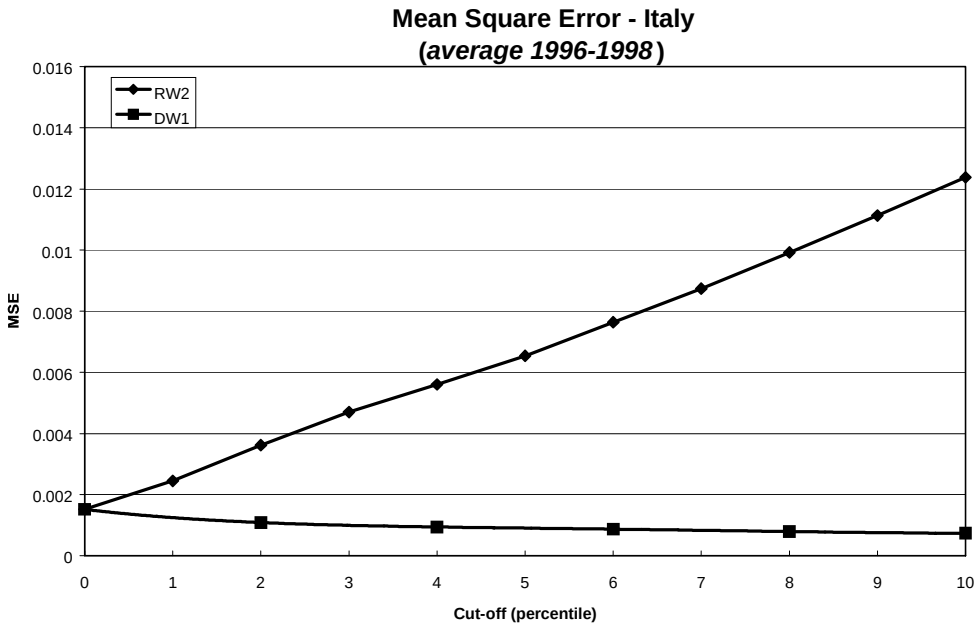
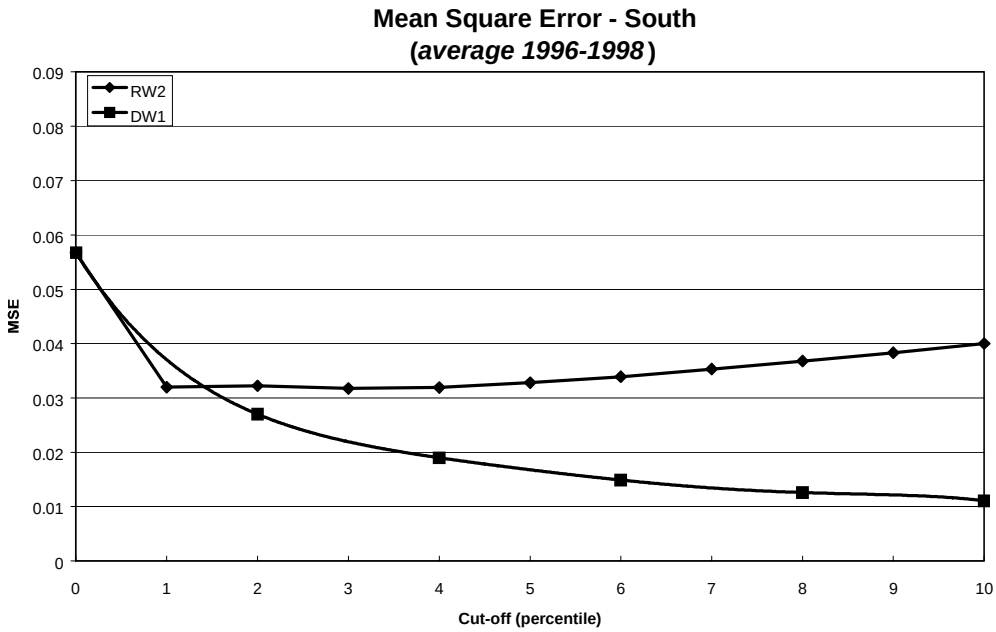


Figure 4



As a preliminary point, it can be noticed that the plotted curves vary approximately monotonically as the cut-off moves; this is a good property, as it means that the picture can be regarded as informative beyond the ten cut-off points selected for this experiment.

At the total sample level, the mean square error under one-sided winsorisation gets quickly larger as the cut-off value gets smaller. This is explained by a steep increase in the bias, which is associated with only a modest reduction in the standard error. The symmetrically winsorised estimator performs much better: the net gain is very small but always positive. This is because, with a bias component very close to zero, the reduction in the standard error is entirely transferred to the total error indicator.

Over the subsample of Southern Italy, both winsorisation methods produce a significant reduction in the mean square error. However, two-sided winsorisation outperforms the asymmetric procedure as soon as the proportion of treated data gets larger than one per cent.

6. Conclusions

Representative outliers (that is, outliers which are believed to represent other units in the population) are a serious problem which is common to many business sample surveys. Recent experience with winsorisation has shown that this is an effective as well as flexible tool to reduce the variance of sample estimators in the presence of extreme observations. Basically, this method consists in changing the values of the outliers towards a cut-off, which can be determined by using information internal and/or external to the empirical distribution of the analysis variable. When the target statistic is a weighted rate of change, a preliminary problem arises, since there is more than one series involved and it is not obvious how to combine the intervention to reduce the variability of the final estimator.

This paper faces the problem of instability which arises in estimating the annual rate of change in industrial investments from the Bank of Italy's annual survey. Since both the amount of investments and their movements are characterised by a large variance due to the presence of a few –usually positive- outliers, the standard errors associated with the traditional estimator (i.e. the mean of all the sampled values, each one weighted by an expansion factor) can be very large. This is especially true for certain sub-samples, where outliers tend to assume very extreme values.

Moving from other approaches –like the traditional scale transformation, which had proved unsatisfactory, this piece of research explores the properties of different kinds of winsorised estimators by setting up a repeated sampling simulation experiment.

From a very large data base ("Centrale dei Bilanci"), which has a good coverage of the survey target population, one hundred independent samples were drawn, reproducing closely the actual survey design. The sample selection procedure was carried out three times, once for each of the last three available years of the data base.

For each sample, the distribution of the rates of change and of the differences between investments at time t and t-1 were treated with one sided and two sided winsorisation respectively; in this way, two alternatives to the traditional expansion estimator were produced.

The performance of the three estimators was compared –across the three selected years- on the whole sample and on the subsample corresponding to Southern Italy, which in recent years has been dramatically affected by outliers.

Finally, the sensitivity of the results to the choice of the cut-offs was explored by working out ten different values for each type of winsorised estimator, each one obtained by treating as outliers a different proportion of data (from one to ten per cent of the empirical distribution).

Results are encouraging. By appropriately linking the final sample estimate to the empirical distribution of individual changes, winsorisation appears to be successful in reducing the standard error of the traditional expansion estimator. Where the problem of outliers is more serious, such as

for the subsample of Southern Italy, the two methods here introduced achieved a standard error reduction of about 30 percentage points over the average of the last three years.

Simulation evidence shows that, when bias and variance are assessed together, two-sided winsorisation outperforms the one-sided approach consistently. This is because, when only upper outliers are treated, the bias introduced in the winsorised estimates tends to be large. As the amount of intervention increases, it rises steeply and dominates the mean square error.

On the whole, the experiment showed that the method is promising, particularly in the two-sided version, which is always as good as, and in many occasions much better than, the traditional estimator. One-sided winsorisation, on the contrary, is very sensitive to the choice of the cut-off: here the trade-off between bias and variance is most often unfavourable, since the price to be paid for a variance reduction is an unacceptably large distorsion in the final estimator.

Further research on this issue could improve the performance of the winsorised estimators. For example, identification could be supported by information derived from previous repeats of the survey, or from some suitable covariates, measured without error in the same survey. In the latter case, outliers treatment would consist in reducing the distance of each observation from the expected value under the model. Furthermore, in a fuzzier implementation of the method, the amount of adjustment could depend not only on the survey weights, but on the extremity of each observation as well. This would allow exploitation of the information conveyed by the extreme data.

There are, of course, even more sophisticated methods of obtaining robust sample estimators. Some of them, which tackle robustness by binding the influence function, deal with the two problems of treatment and estimation simultaneously (see Cowell, Gomulka and Victoria Feser, 1999). However, these methods mostly rely on a parametric approach, which often does not match the need for quick and multivariate estimation that is typical of survey-based inference.

Whatever the method which is finally selected, it is convenient in many respects to embed it in a written procedure, so as to make the criteria more objective. This is especially true for public institutions, which want to make their analyses as transparent and replicable as possible. However, it must not be forgotten that the intellectual curiosity and experience-based wisdom of the analyst may often add an invaluable contribution to the understanding and hence the accountability of any estimate.

Acknowledgements

This work has been started during my 5 months visit at the Department of Statistics of the London School of Economics and Political Science. Thanks are due to Luigi Cannari and Giovanni D'Alessio for their invaluable suggestions, to Jeremy Penzer (L.S.E.) for his helpful comments and encouragement and to P. Smith (U.K. Office for National Statistics) for his kind co-operation in providing references and unpublished material.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust estimates of location*. Princeton, Princeton University Press.

Atkinson, A.C. (1985) *Plots, Transformations and Regression.* Oxford, Clarendon Press.

Bickel, P.J. and E.L. Lehmann (1975). Descriptive statistics for nonparametric models II. Location. *The Annals of Statistics,* 3, 1045-1069.

Caperaa, P. and L.P. Rivest (1995) On the variance of the trimmed mean *Statistics and Probability Letters* 22, 79-85,

Carroll, R.J. (1979). On estimating variance of robust estimators when the errors are asymmetric**.** *Journal of the American Statistical Association, 74,* 674-679.

Collins, J.R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *The Annals of Statistics, 4,* 68-83.

Chambers, R.L. (1986). Outlier Robust Final Population Estimation. *Journal of the American Statistical Association,* Vol. 81, No. 396, 1063-1069.

Cowell, F. Gomulka, J. and Victoria Feser, M-P. (1999) INEQ Manual. London School of Economics, STICERD.

Cruddas, M. and Kokic, P. (1996) The Treatment of Outliers in ONS Business Surveys. *Proceedings of the GSS(M) methodology conference*. ONS, Newport.

Daniell, P.J. (1920). Observations weighted according to order. *American Journal of Mathematics, 42*, 222-236.

Hampel, F.R. et al (1986) *Robust statistics – The approach based on influence functions*. New York: Wiley Series in Probability and Mathematical Statistics.

Hampel, F.R. (1985*)* The breakdown point of the mean combined with some rejection rules. *Technometrics,* 27, 95-107.

Hidiroglou M.A. and K.P. Srinath (1981) Some Estimators of a Population Total from Simple Random Samples Containing Large Units. *Journal of the American Statistical Association*, Vol. 76, No. 37, 690-695.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (eds.) (1983) *Understanding* r*obust and exploratory data*. New York: Wiley Series in Probability and Mathematical Statistics.

Holt, D. and Smith T.M.F. (1979) Post Stratification *Journal of the Royal Statistical Society*, Ser. A, 142, 33-46.

Huber, P.J. (1981). *Robust statistics.* New York: Wiley Series in Probability and Mathematical Statistics.

Huber, P.J. (1972). The 1972 Wald Lecture. Robust statistics: a review. *The Annals of Mathematical Statistics, 43 n.4,* 1041-1067.

Jaeckel, L.A. (1971). Some flexible estimates of location. *The Annals of Mathematical Statistics, 42 n.5,* 1540-1552.

Jenkins, O.C., L.J. Ringer and H.O. Hartley (1973) Root estimators. *Journal of the American Statistical association,* 78, 414-419.

Kokic, P.N. and P.A. Bell (1994) Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. *Journal of Official Statistics,* 10 n.4, 419-435.

Kokic, P. and P. Smith (1998) *Winsorisation of Outliers in Business Surveys* (forthcoming).

Kokic, P. and P. Smith *Outlier-robust estimation in sample surveys using two-sided winsorisation* (forthcoming).

Lehmann, E.L. (1983). *Theory of point estimation*. New York: Wiley Series in Probability and Mathematical Statistics.

Rousseeuw, P. and A. M. Leroy (1987) *Robust regression and outliers detection*. New York: Wiley Series in Probability and Mathematical Statistics.

Searls, D.T. (1966) An Estimator for a population mean which reduces the effects of large true observations. *Journal of the American Statistical Association,* Vol. 61, 1200-1204.

Smith, P. (1997) *Winsorisation: effects in practice*. MQ 040 ONS

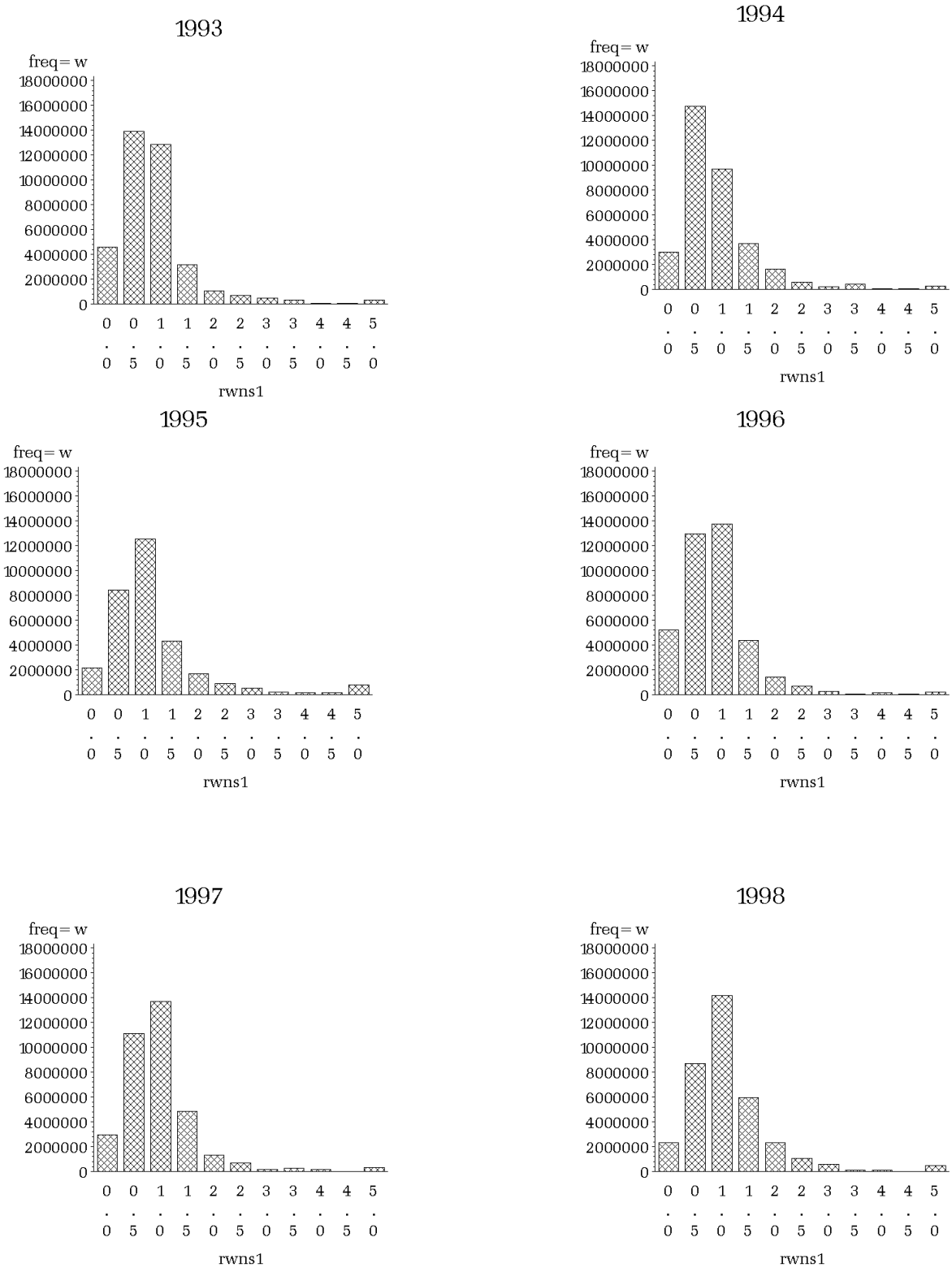
Smith, P., P. Kokic and S. Hibbitt (1997) *Two-sided winsorisation for the ABI*. MQ 045 ONS

Stigler, S.M. (1976). The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. *Journal of the American Statistical Association, 71,* 956-959.

Victoria-Feser, M-P and Ronchetti, E. (1994). Robust methods for personal-income distribution models. *The Canadian Journal of Statistics, 22,* 247-258.

Figure 5

Individual annual changes[8]

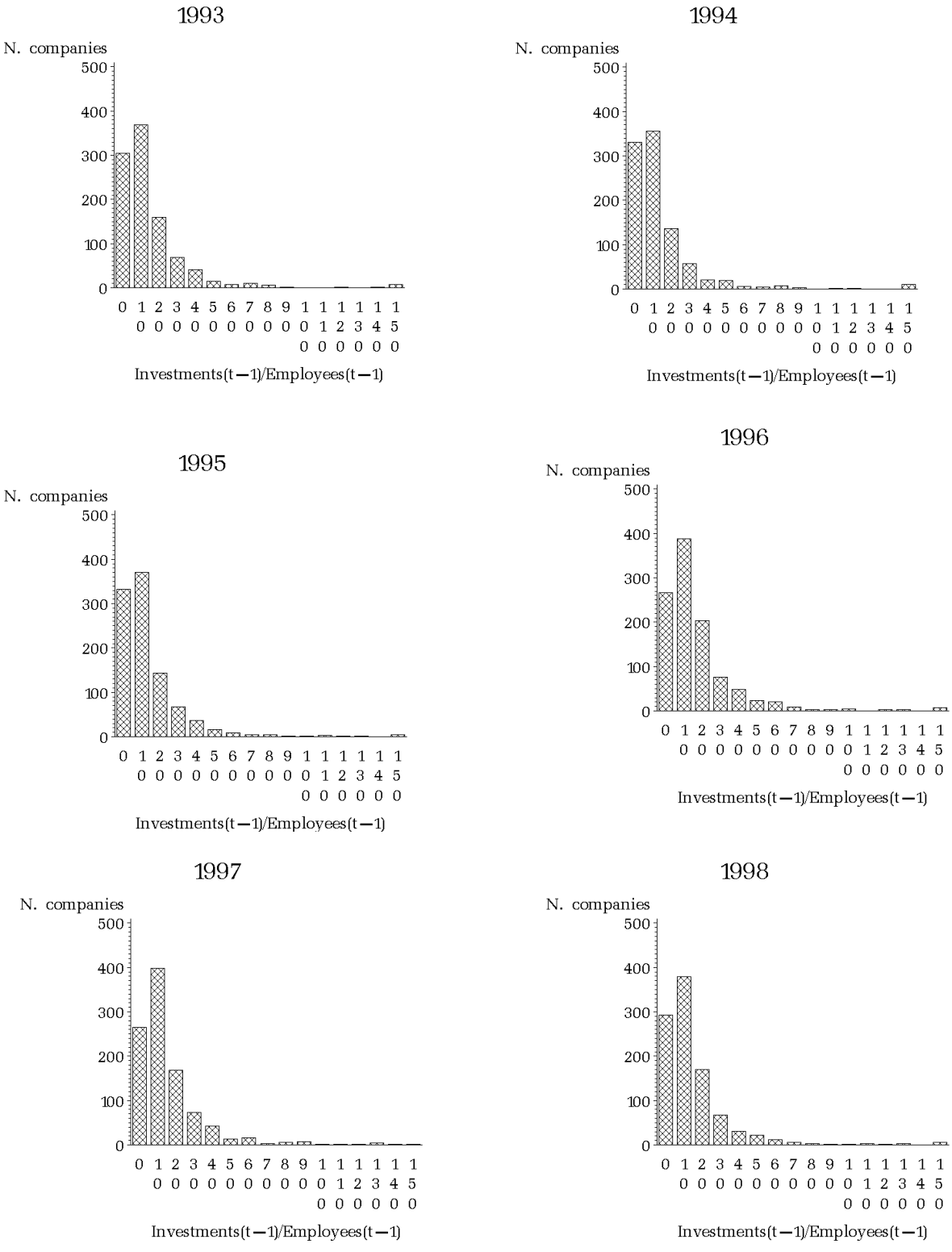


[8] Weighted empirical distributions from the annual surveys (winsorised weights). Each distribution has been truncated at a different cut-off point, ranging from the 95th to the 99th percentile.

Figure 6

Ratio investments/employees[9]

(millions of Italian liras)



[9] Weighted empirical distributions from the annual surveys. Each distribution has been truncated at a different cut-off point, ranging from the 95th to the 99th percentile.

Figure 7

Performance of the winsorised estimators with different cut-offs 1998, Italy

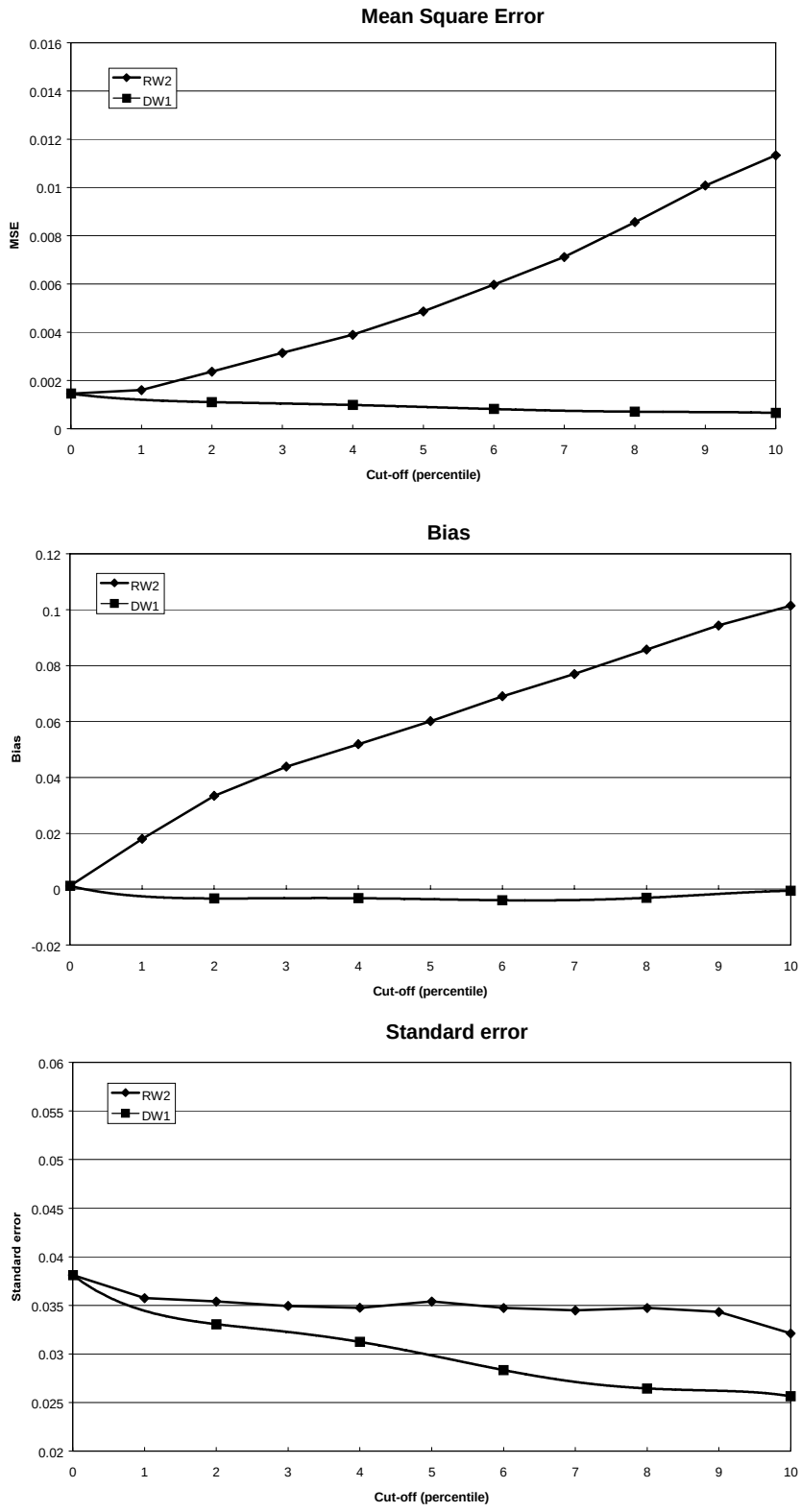


Figure 8

Performance of the winsorised estimators with different cut-offs 1998, South

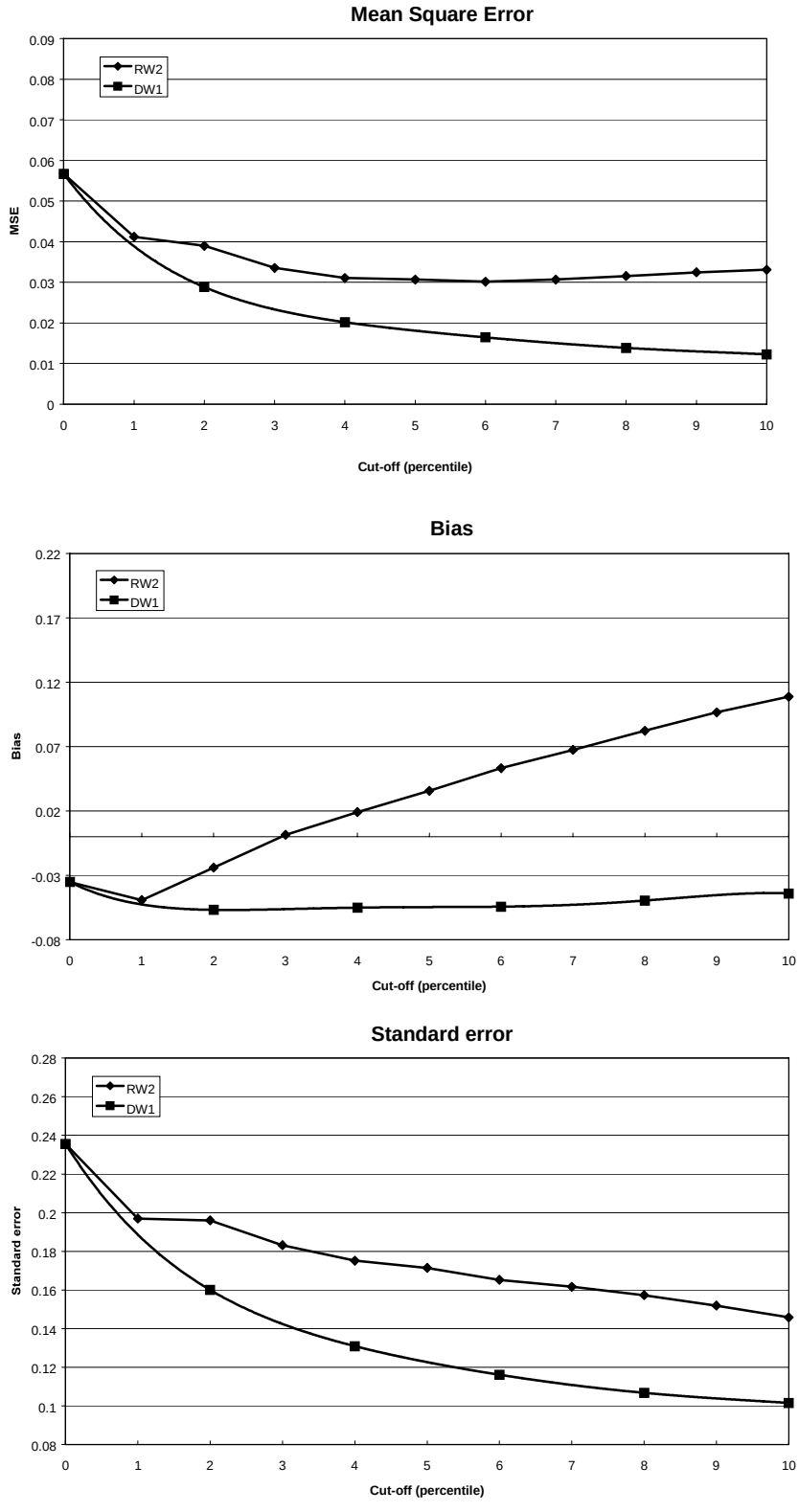


Figure 9

Performance of the winsorised estimators with different cut-offs 1997, Italy

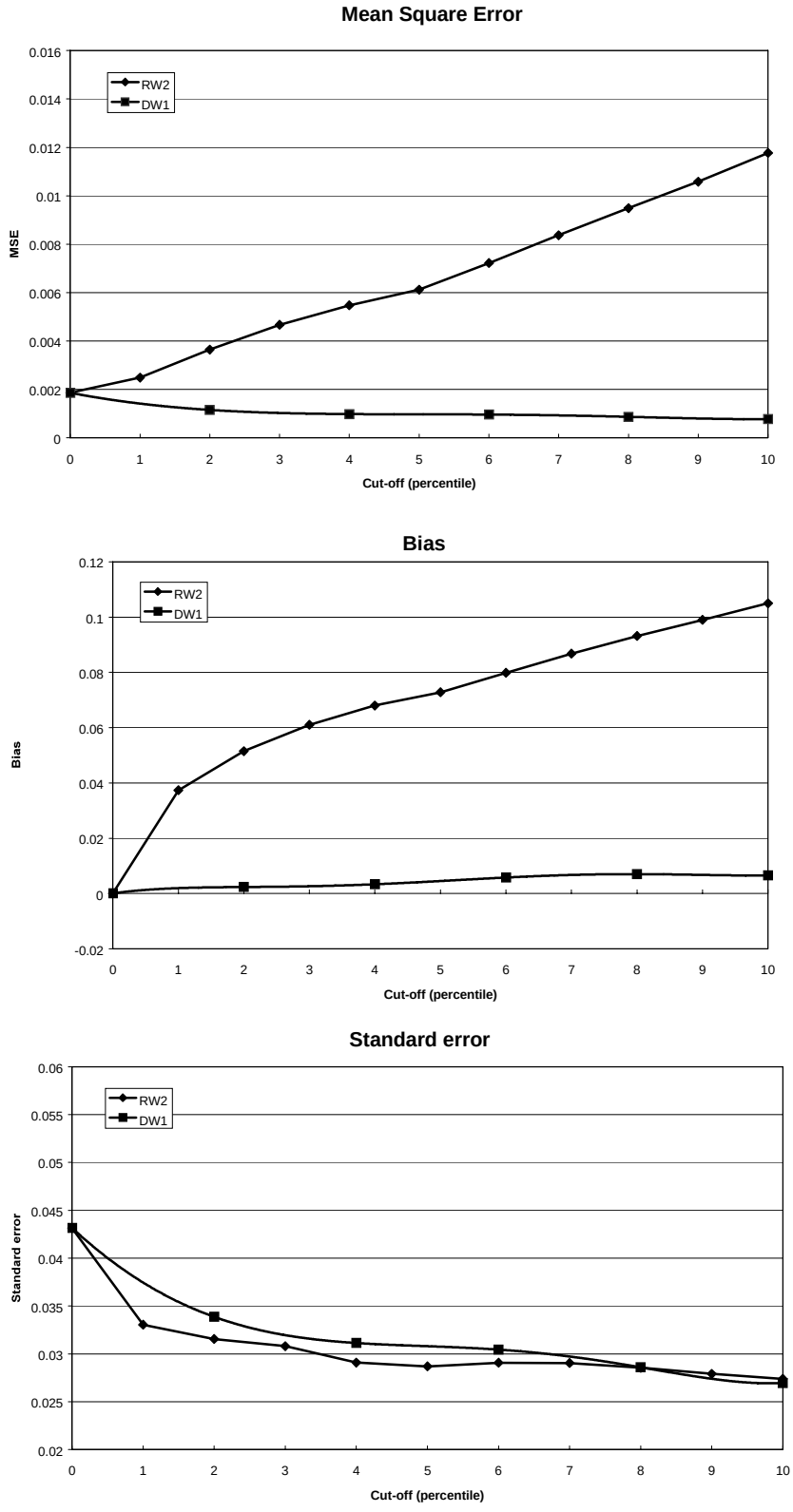


Figure 10

Performance of the winsorised estimators with different cut-offs 1997, South

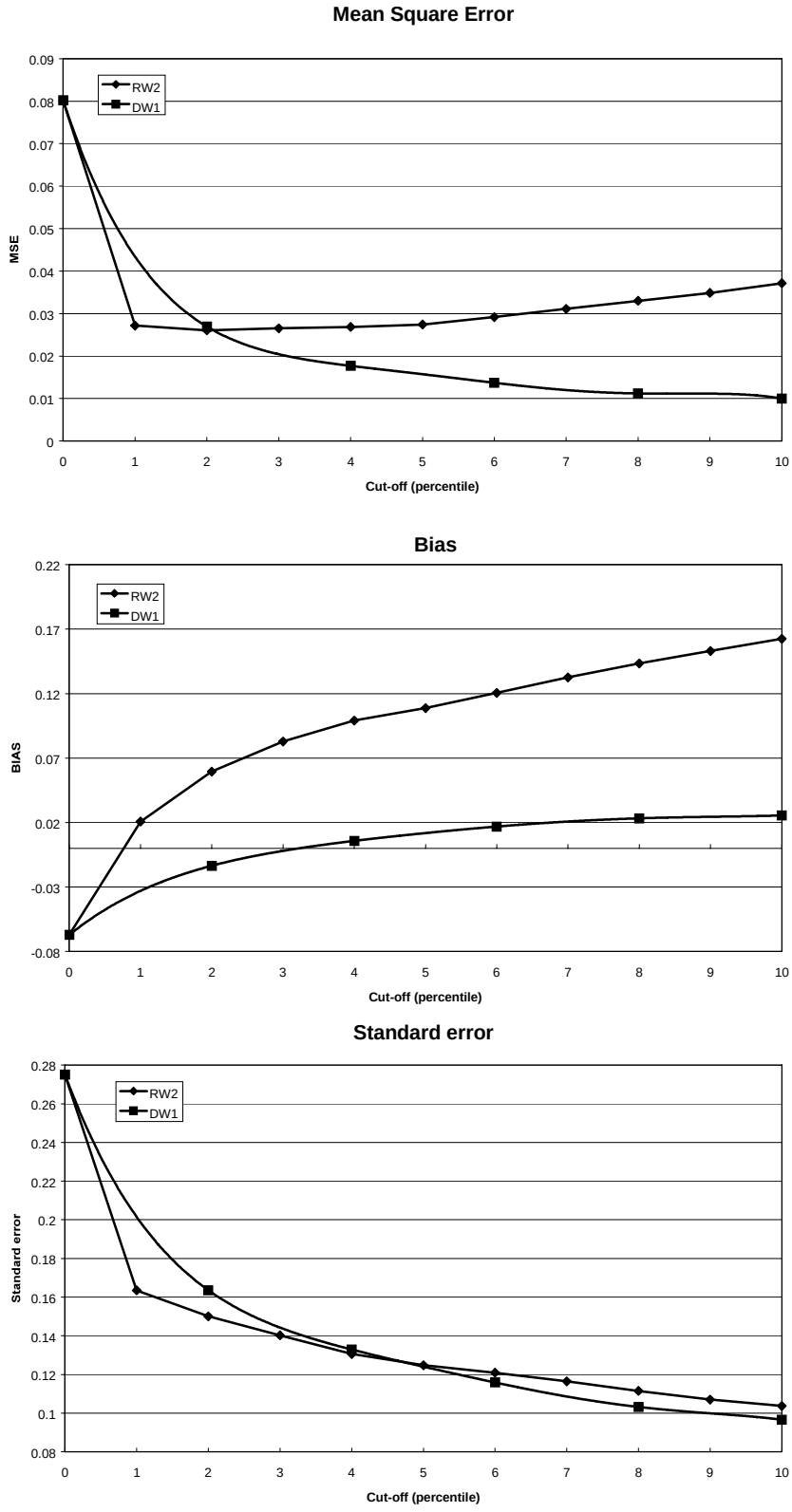


Figure 11

Performance of the winsorised estimators with different cut-offs 1996, Italy

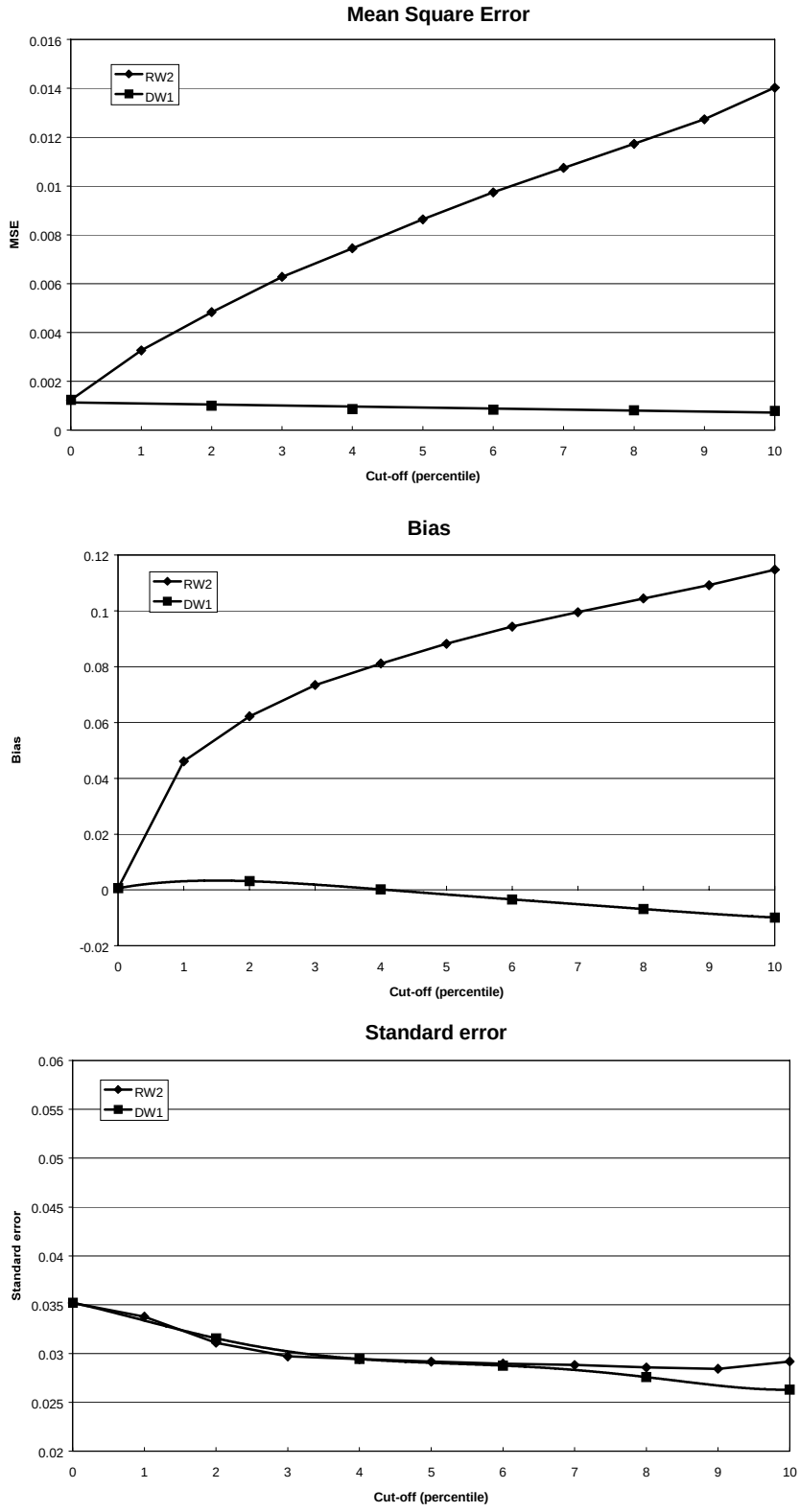


Figure 12

Performance of the winsorised estimators with different cut-offs 1996, South