Outliers identification and handling: an advanced econometric approach for practical data applications

G. Palmegiani | LUISS University of Rome | Rome | Italy

DOI: 10.1481/icasVII.2016.d24c

ABSTRACT

Before implementing any statistical analysis it is crucial to check whether outliers are present in the dataset because their existence imply a distorted analysis that leads to unreliable results and incorrect policy decisions. To overcome that, this paper proposes and advanced econometric approach based on a linear regression framework that is capable of identifying and handling vertical outliers under three alternative scenarios. This methodology, that can be generalized to any other dataset, is applied to a real irrigation dataset on 15 countries sorted out in two groups in order to compute reliable summary statistics on unit irrigation costs that help governments, international organizations and water management in implementing cost efficient water investment decisions as well as knowing how much to pay on average, at maximum, and at minimum, for a given irrigation investment project type located in a given country. Once vertical outliers are handled, two main conclusions are found. First, unit irrigation costs tend on average to be lower for Middle-East countries rather than Afro-Arab countries. Second, is not shown that large scale irrigation investment projects are on average more expansive than small scale irrigation investment projects since it depends on both the country and scenario chosen. Although, the first result is in line with the reality that gives to Afro-Arab countries higher average unit irrigation costs, the second result confirmed at pooled and country level has not been sufficiently addressed by the literature and opens to new different water policy recommendations.

Keywords: Outliers, Unit Irrigation Costs, Method of Moments (MM) Estimator

PAPER

1. Introduction

Outliers are atypical observations, present in virtually every dataset, that can strongly distort any statistical analysis producing unreliable results that lead to incorrect policy decisions. This is especially true, as the pioneer Edgeworth (1887) work has shown, when Ordinary Least Squares (OLS) estimator is used as a tool for outliers identification. This happens for two main reasons. First, being this estimator an equal weighted sum of dependent variable observations, it treats with an equal weight observations that are potential outliers from observations that are not potential outliers while discrimination should be needed instead. Second, being also the OLS estimator based on a minimization of squared residuals that means practically minimizing their variance, when outliers are present the variance of the residuals increases considerably and the OLS estimator is largely influenced as well. It implies distorted estimates as well, unreliable results and incorrect policy decisions. These reasons imply that the OLS estimator does not represent a good tool, in reality is the worst, to identify outliers unless a prior different estimator or analysis concludes that these atypical observations are not present. In fact, something which is not typical cannot be identified using an estimation tool that is typical and constructed under a series of regularity conditions. Then, as pointed out by Huber (1964), Rousseeuw and Yohai (1987), Yohai (1987), Maronna (2006) an other outlier identifying tool is needed.

Furthermore, outliers are mainly of four types according their influence on the OLS estimator. Vertical outliers are those observations that are outlying respect to the space spanned by the dependent variable but not outlying in the space spanned by the regressors. Horizontal outliers are those observations that are outlying respect to the space spanned by the regressors but not outlying in the space spanned by the dependent variable. Good leverage points are observations that are outlying in the space spanned by the regressors but that are located close to the regression line. Bad leverage points are observations that are both outlying in the space of explanatory variables and located far from the true regression line. As mentioned, these types of outliers influence OLS estimates in different ways. Vertical outliers affects OLS estimates and in particular they affects the estimated regression intercept. Horizontal outliers affects OLS estimates and especially they affects the slope coefficients of the regression. Good leverage points does not affect the OLS estimates because they are points located closely to the regression line but it affects statistical inference since they do inflate the estimated standard errors. Bad leverage points affects significantly the OLS estimates of both the intercept and slopes coefficients.

In addition, outliers do not come alone especially in survey data but are almost always accompanied by missing values, large number of variables, sampling weights and measurement errors. Handling each of these data features requires further analysis and study.

This paper outperforms all these challenges proposing an advanced econometric approach for outliers identification based on regression framework estimated using Method of Moments (MM) rather than OLS in a three scenarios context. This methodology, that can be generalized to any other dataset, is applied to a real irrigation dataset in order to compute reliable unit irrigation costs that can help governments and international organization in formulating specific territorial policy recommendation and advices, to enhance the water population conditions as well as suggesting the right price per hectare that an international investor should efficiently pay for an given irrigation investment project in a given region or country.

The rest of the paper is organized as follows. The next section presents the dataset used as well as some basic variables' assumptions, Section (3) describes the methodology used to identify and handling outliers, Section (4) applies the methodology described before computing reliable summary statistics on unit irrigation costs at pooled, region and country level as well as by irrigation type level in a three scenarios context, Section (5) concludes summing up the main paper findings and finally Section (6) gives possible further extensions.

2. The dataset

The dataset is collected by aggregating specific country irrigation portfolios spreadsheets. The aggregated dataset consists of two main variables:

- Total = is the total irrigation investment cost (expressed in unit of United States Dollars (USD));
- Hectares = is the surface (expressed in hectares) covered by each irrigation investment project.

Regarding the total cost variable, since each investment project can be devoted to more than one types of irrigation project, indeed, large scale (LS) projects, small scale (SS) projects and rehabilitation and modernization (RM) projects, it has been calculated exactly the proportion of the total cost devoted to each type or irrigation project. It has been assumed the total cost variable is free from measurement errors. Regarding the hectare variable, it is known just the entire irrigation investment surface while the surface devoted to specific type of irrigation projects remains unknown. Further, it has been assumed both that the hectare variable is properly measured and free from outliers. It implies that, just vertical outliers are present in the real dataset and just statistics to identify vertical outliers are needed. Finally, the total number of countries involved are 15 sorted out in Middle-East countries and Afro-Arab countries while the total number of irrigation investment projects here considered are 225.

Given these assumptions about variables, the main question of interest is how calculate reliable summary statistics on unit irrigation costs at pooled, region and country level and by type of irrigation project. Because having reliable unit irrigation costs it is crucial to implement correct water and agriculture policies at these territorial levels and for these investments types.

3. The cost analysis: the methodology used

Given the variables' assumptions, the summary statistics on unit irrigation costs have been worked out after identifying outliers respect to the cost variable. Once vertical outliers has been identified, the mean, the standard deviation, the maximum value and the minimum value of unit irrigation costs are computed at pooled, region and country level and by type or irrigation project in a three scenarios context.

As mentioned, the outliers identification has been carried out using a linear regression of the log of total cost on the log of hectares covered and on the log of squared hectares estimated by MM rather than OLS. Then, the functional regression form used to identify outliers is:

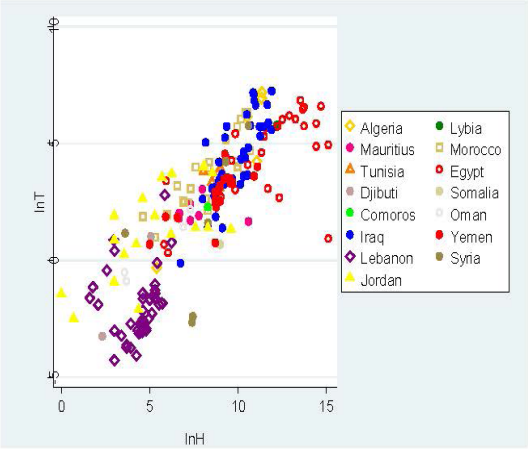
$$ln(TC_{i,h,k}) = b_0 + b_1\, ln(H_{i,h,k}) + b_2[ln(H_{i,h,k})]^2 + \varepsilon_{i,h,k} \quad forall \quad i = \{1,2,3,\ldots,225\}, \; h = \{LS,SS,RM,ALL\}, \; k = \{pooled, region, country\} \qquad (1)$$

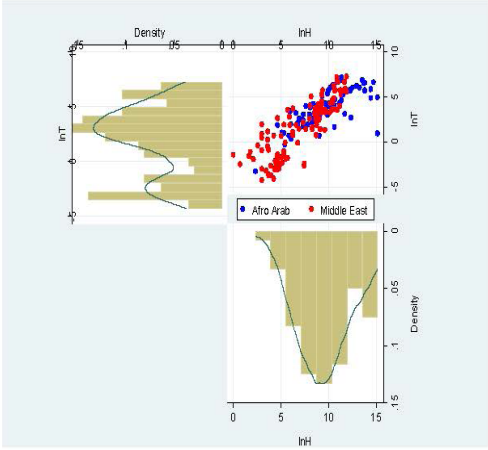
where, $ln(TC_{i,h,k})$ denotes the log of total irrigation investment cost for observation *i*, irrigation type *h* and territorial aggregation *k*, $ln(H_{i,h,k})$ denotes the log of hectares covered by the irrigation investment project for observation *i*, irrigation investment type *h* and territorial aggregation *k*, $[ln(H_{i,h,k})]^2$ denotes the squared of logs along the hectares covered by the irrigation investment for observation i, irrigation investment type h and territorial aggregation k, $E_{i,h,k}$ denotes the regression error term for observation

i, irrigation investment type *h* and territorial aggregation *k* and finally b_0 is a constant term while b_1 and b_2 are the elasticities.

This is a log-log functional regression form with a quadratic regressor term. This functional form has been tested severally and confirmed to be optimal for the starting dataset at left of Figure (1) at different territorial levels for the following reasons. Regarding the log-log form, first, when logs of both dependent and independent variable are taken the distributions of variables are more near to a normal distribution as shown at right of Figure (1) and it is coherent with classic regression assumptions. Second, when logs are taken the R-adjusted increases considerably and it is usually desirable when a statistician have to choose between levels and logs. Third, the Box-Cox test rejects both the functional form in levels and the function form log-lin, then it is confirmed further on that both levels and log-lin forms are not appropriated. While regarding the inclusion of the quadratic term, this choice has been motivated by the necessity to capture no-constant elasticity between total costs and the hectares covered that allows increasing or decreasing marginal costs effects. Further, this term has been included only when, inside the regression estimated MM, it is significant according to the t-test. Furthermore, when it is t-significant this term is also capable of improving R-adjusted considerably.

Figure 1: *The starting log-transformed dataset [at left] and log-variables distributions [at right] for all countries (this figure, at left, points out the starting dataset variables by scattering the log of total irrigation cost versus the log of hectares covered by country, while at right, it shows the related depended and independent variables distributions with the same scatter now sorted out by region)*





As remarked, the regression shown in Equation (1) has been estimated MM rather than OLS to discriminate between potential outliers observation and no potential outliers observations. Then, once estimated, it has been drawn out a measure of vertical outlyingness given by the robust standardized residuals (denoted by $S_{stdresi,h,k}$) and flagged outliers as those observations that have a robust standardized residual larger than 2.25 in absolute value.

Once flagged the outliers, unit irrigation costs has been computed in the following three scenarios context. In scenario A, no vertical outliers are assumed then independently from the outlyingness statistics drawn, all observation are kept and the unit irrigation cost are computed as the ratio of total cost on the hectares covered, indeed:

$$UC^{A}_{i,h,k} = \frac{TC_{i,h,k}}{H_{i,h,k}} \quad | \quad abs\left(S_{stdres_{i,h,k}}\right) \leq 2.5 \quad or \quad abs\left(S_{stdres_{i,h,k}}\right) > 2.5 \qquad (2)$$

where $UC^{A}_{i,h,k}$ denotes the unit irrigation cost computed under scenario A for observation i, irrigation investment project type h and territorial aggregation k while $abs\left(S_{stdres_{i,h,k}}\right)$denotes the absolute value of our vertical outlyingness statistics measure.

Differently, in scenario B vertical outliers are cancelled out while no-vertical outliers observations are kept and the unit costs are computed as the previous scenario, indeed:

$$UC^{B}_{i,h,k} = \begin{cases} \frac{TC_{i,h,k}}{H_{i,h,k}} & | \quad abs\left(S_{stdres_{i,h,k}}\right) \leq 2.5 \\ the\ i_{th}\ observation\ is\ cancelled\ out & | \quad abs\left(S_{stdres_{i,h,k}}\right) > 2.5 \end{cases} \qquad (3)$$

Finally, in scenario C all observations are kept but vertical outliers are mean demanded once divided by their absolute value of the robust standardized residuals. Differently, for no-vertical outliers observations unit costs are computed as usual, indeed:

$$UC^{C}_{i,h,k} = \begin{cases} \frac{TC_{i,h,k}}{H_{i,h,k}} & | \quad abs\left(S_{stdres_{i,h,k}}\right) \leq 2.5 \\ \frac{abs\left(\frac{TC_{i,h,k}}{H_{i,h,k}} - E_{h,k}\left[\frac{TC_{i,h,k}}{H_{i,h,k}}\right]\right)}{abs\left(S_{stdres_{i,h,k}}\right)} & | \quad abs\left(S_{stdres_{i,h,k}}\right) > 2.5 \end{cases} \tag{4}$$

where, $E_{h,k}[.]$ denotes the averaging function that apply the mean operator to unit irrigation costs over observations i given irrigation type h and territorial aggregation k.

This outliers weighting scheme it is in line with the form of MM-estimators and it has been shown to produce very reliable summary statistics on unit irrigation costs as pointed out by the practical following application provided in the next section.

4. The cost analysis: the results and policy decisions

This section applies the methodology explained before proving reliable summary statistics on unit irrigation costs, such as, the mean of unit costs, their standard deviation, the maximum value and the minimum value at different territorial levels, namely, at pooled (Table (1)), regional (Table (2)) and country level (Table (3)) and by type of irrigation project, namely, small scale (SS), large scale (LS) and rehabilitation and modernization (RM).

Looking at pooled level, Table (1) shows the first and the most important result of this paper, it is not shown that LS irrigation investment projects are on average more expensive than SS irrigation investment projects but it depends on the scenario chosen. In fact, the mean of unit irrigation costs for LS irrigation investment projects is higher than the mean of unit irrigation costs for SS irrigation investment projects only in scenario B where vertical outliers are deleted.

Table 1: *Unit irrigation costs at pooled level: main summary statistics (this table, shows the number of observations, the mean, the standard deviation, the minimum and the maximum value of unit irrigation costs (UC) computed as in scenarios explained in Section(3) by irrigation investment type)*

	Scenario A		Scenario B		Scenario C	
	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]
UC	225	11954.51 (33169.62) [.41 ;339750]	192	4274.89 (5333.02) [37.91 ; 27060]	225	5307.99 (6427.23) [37.91 ; 44606.92]
UCss	68	26787.76 (55980.29) [.13 ; 339750]	41	3810.17 (4186.94) [37.91 ; 27060]	68	6905.47 (7527.52) [200 ; 43812.24]
UCls	80	5193.91 (6205.70) [.14 ; 28040.57]	76	5095.75 (5704.64) [183.05 ; 27060]	80	5014.48 (5590.95) [183.05 ; 27060]
UCrm	107	4230.24 (9097.42) [.13 ; 78846.16]	100	2772.36 (3883.47) [37.91 ; 26647.3]	107	2936.20 (3999.48) [37.91 ; 26647.3]

Table 2: *Unit irrigation costs at regional level: main summary statistics (this table, shows the number of observations, the mean, the standard deviation, the minimum and the maximum value of unit irrigation costs (UC) computed as in scenarios explained in Section(3) by irrigation investment type)*

	Scenario A		Scenario B		Scenario C	
Afro-Arab region						
	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]
UC	75	7275.88 (10072.16) [.41 ; 64000]	69	6972.65 (7695.34) [37.91 ; 38820]	75	6939.34 (7677.12) [37.91 ; 38820]
UCss	19	9019.92 (14439.58) [.13 ; 64000]	16	6698.65 (5685.41) [558.43 ; 19410]	19	7155.54 (5916.89) [558.43 ; 19410]
UCls	35	5161.12 (6026.75) [.14 ; 27060]	31	5814.87 (6107.34) [192.12 ; 27060]	35	5152.43 (6034.27) [9.81 ; 27060]
UCrm	40	4840.63 (6569.73) [.13 ; 34628.57]	39	4964.75 (6607.93) [37.91 ; 34628.57]	40	4840.63 (6569.73) [0 ; 34628.57]
Middle-East region						
UC	150	14293.83 (39838.63) [42.21 ; 339750]	121	3072.88 (4191.35) [242.85 ; 25222.91]	150	4638.69 (6134.29) [242.85 ; 39647.78]
UCss	49	33677.33 (64207.84) [42.21 ; 339750]	26	2895.61 (3833.23) [470 ; 18077.14]	49	6937.83 (7689.96) [470 ; 38600.26]
UCls	45	5219.41 (6409.05) [251.2 ; 28040.57]	42	4501.95 (5154.73) [355 ; 25222.91]	45	4423.91 (5013.56) [355 ; 25222.91]
UCrm	67	3865.83 (10346.86) [242.85 ; 78846.16]	61	1761.49 (2254.66) [242.85 ; 14089.38]	67	2021.69 (2583.51) [242.85 ; 14089.38]

Differently, when all observations are held or outliers weighting scheme is carried out the mean of unit costs referred to LS irrigation investment projects is lower than the mean of unit costs referred to SS projects[1]. This result imply that, larger is the scale of a given irrigation investment project and smaller could be on average the unit cost paid per hectare. It means that the investor, whether he has the option to choose the type of investment, could save on average money implementing a more efficient water policy by investing in LS projects rather than SS projects if he believes that his information has not atypical values or these values are present but he retains weighable.

Looking at the region level, Table (2) shows the second most important result of this paper. Once vertical outliers are handled, on average unit irrigation costs tend to be lower for Middle East countries rather than Afro-Arab countries[2].

[1] In fact, in scenario C the mean of unit irrigation costs for LS projects is around 1900 USD per hectare cheaper than the average of unit costs for SS projects (6905,14 USD paid on average per hectare versus 5014,48 USD paid on average per hectare). Further, this LS project mean convenience increases amply in scenario A (26787,76 USD versus 5139,91 USD per hectare). Differently, in scenario B the mean of LS projects becomes higher than the mean of SS projects (5095.75 USD versus 3810,71 USD per hectare).

2 In fact, in both scenario B and C the means of unit irrigation cots are lower for Middle-East country rather than in Afro-Arab countries. For example, in scenario B, the mean vary between 1761,49 USD per hectare and 4501,95

Looking at country level, Table (3) confirms, at this territorial level, that once vertical outliers are handled LS irrigation investment projects are not on average more expensive than SS irrigation investment projects but it depends on the chosen country. For example, LS projects are cheaper on average than SS projects in Lebanon, Egypt and Oman in both scenario B and C. This implies that higher is the scale of the irrigation project and lower is on average the cost paid per hectare. It follows that for an international investor are on average more convenient in these countries LS projects rather than SS projects because the cost per hectare for LS projects is smaller.

Table 3: *Unit irrigation costs at country level: main summary statistics (this table, shows the number of observations, the mean, the standard deviation, the minimum and the maximum value of unit irrigation costs (UC) computed as in scenarios explained in Section(3) by irrigation investment type*

	Scenario *A*		Scenario *B*		Scenario *C*	
Lebanon						
	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]	obs.	Mean (Std.Dev.) [min; max]
UC	57	7672.74 (21626.04) [242.85 ; 124635.4]	46	2384.32 (8137.73) [416.66 ; 53333.33]	57	3218.77 (7643.65) [195.02 ; 53333.33]
UCss	16	23302.68 (36535.42) [470 ; 124635.4]	8	9486.11 (18817.3) [470 ; 53333.33]	16	8345.76 (13247.49) [470 ; 53333.33]
UCls	2	15655.54 (17515.09) [3270.498 ; 28040.57]	1	3270.49 (·) [3270.498 ; 3270.498]	2	1635.24 (2312.59) [0 ; 3270.498]
UCrm	39	851.09 (513.71) [242.85 ; 2450]	37	824.85 (445.34) [416.66 ; 2450]	39	851.09 (513.71) [242.85 ; 2450]
Iraq						
UC	40	7184.58 (9107.91) [437.73 ; 44228.57]	35	4799.84 (4434.95) [1011.35 ; 18077.14]	40	4671.53 (4380.21) [422.41 ; 18077.14]
UCss	12	7097.57 (12595.77) [1035 ; 44228.57]	11	3722.02 (4910.58) [1035 ; 18077.14]	12	3411.85 (4803.75) [0 ; 18077.14]
UCls	18	6588.21 (7159.93) [707.94 ; 25222.91]	16	4407.04 (3498.68) [707.94 ; 12900]	18	3970.43 (3523.55) [466.91 ; 12900]
UCrm	17	4919.09 (6500.72) [303.41 ; 26647.3]	15	3769.31 (3423.88) [303.41 ; 14089.38]	17	3962.94 (3249.07) [303.41 ; 14089.38]
Egypt						
UC	26	3822.36 (12349.21) [.41 ; 64000]	20	1751.26 (1358.45) [187.28 ; 4749.34]	26	2558.31 (3004.59) [187.28 ; 15601.41]
UCss	4	17266.11 (31184.74) [.13 ; 64000]	2	2532.14 (1055.61) [1785.71 ; 3278.571]	4	4590.44 (3310.83) [1785.71 ; 9371.253]
UCls	12	993.71 (1003.96) [.14 ; 3671.73]	9	1297.51 (984.09) [192.12 ; 3671.73]	12	977.96 (1019.09) [5.37 ; 3671.73]
UCrm	18	1019.13 (1239.52) [.13 ; 4749.34]	15	1215.69 (1271.65) [187.28 ; 4749.34]	18	1014.04 (1243.78) [.46 ; 4749.34]
Morocco						
UC	25	12041.02 (9747.34) [1116.87 ; 38820]	16	15058.23 (6871.64) [5176 ; 34628.57]	25	10369.99 (8528.41) [1073.01 ; 34628.57]
UCss	10	7656.79 (5514.61) [558.43 ; 19410]	7	7962.71 (2527.52) [6354 ; 13550]	10	6231.84 (3518.91) [1246.54 ; 13550]
UCls	14	8453.78 (6988.23) [558.43 ; 27060]	11	10434.92 (6572.81) [5176 ; 27060]	14	8231.05 (7239.63) [84.71 ; 27060]
UCrm	11	9645.87 (10150.43) [860.57 ; 34628.57]	4	17602.14 (11370.81) [11200 ; 34628.57]	11	7166.04 (10407.04) [480.15 ; 34628.57]
Yemen						
UC	24	3809.99 (4204.03) [355 ; 17759.56]	14	2475.48 (535.88) [1736.66 ; 3177]	24	1785.65 (952.74) [326.57 ; 3177]
UCss	3	12096.99 (2931.36) [8750 ; 14207.65]	0		3	1026.24 (693.44) [413.93 ; 1779.24]
UCls	20	2538.69 (1710.12) [355 ; 8952.381]	14	2475.48 (535.88) [1736.66 ; 3177]	20	1856.01 (1077.93) [176.08 ; 3177]
UCrm	2	2187.43 (1929.67) [822.94 ; 3551.91]	0		2	818.38 (347.74) [572.49 ; 1064.28]
Jordan						
UC	18	65733.87 (92577.48) [251.2 ; 339750]	11	79452.09 (97375.39) [922.5 ; 339750]	18	51232.3 (83149.87) [331.26 ; 339750]
UCss	11	94311.58 (107817.5) [922.5 ; 339750]	9	84103.04 (106995.5) [922.5 ; 339750]	11	8786.72 (9638.03) [92.25 ; 33975]
UCls	3	8046.55 (8743.12) [251.2 ; 17500]	0		3	1141.57 (900.02) [172.90 ; 1951.96]
UCrm	7	17377.53 (27908.79) [1000 ; 78846.16]	4	29261.4 (33440.42) [6000 ; 78846.16]	7	1693.23 (2820.69) [6.75 ; 7884.61]
Mauritania						
UC	8	5516.94 (2925.28) [131.32 ; 7800]	5	7506.27 (290.51) [7084.83 ; 7800]	8	4772.44 (3780.52) [108.41 ; 7800]
UCss	0		0		0	
UCls	2	1468.78 (1891.45) [131.32 ; 2806.25]	0		2	1050.06 (890.73) [420.22 ; 1679.91]
UCrm	6	6866.33 (1588.9) [3666.66 ; 7800]	5	7506.27 (290.51) [7084.83 ; 7800]	6	6255.22 (3075.42) [0 ; 800]
Tunisia						
UC	6	5546.18 (2334.21) [1622.25 ; 8818.45]	4	5709.11 (586.37) [4954.73 ; 6385.16]	6	4048.4 (2612.63) [685.81 ; 6385.16]
UCss	0		0		0	
UCls	4	4647.01 (3096.24) [1622.25 ; 8818.45]	2	4073.65 (1246.02) [3192.58 ; 4954.73]	4	2400.36 (2062.01) [685.81 ; 4954.73]
UCrm	3	4896.37 (1475.92) [3192.58 ; 5782.33]	3	4896.37 (1475.92) [3192.58 ; 5782.33]	3	4896.37 (1475.92) [3192.58 ; 5782.33]
Syria						
UC	6	16977.13 (34211.44) [42.21 ; 86486.48]	4	1881.56 (2930) [42.21 ; 6195.45]	6	2600.47 (3054.25) [42.21 ; 6748.04]
UCss	6	16977.13 34211.44 [42.21 ; 86486.48]	4	1881.56 2930 [42.21 ; 6195.45]	6	2600.47 3054.25 [42.21 ; 6748.04]
UCls	0		0		0	
UCrm	0		0		0	
Oman						
UC	5	8566.35 (5027.95) [4226.82 ; 16594.45]	3	8885.27 (6724.55) [4226.82 ; 16594.45]	5	5499.5 (6642.84) [203.8 ; 16594.45]
UCss	1	16594.45 (·) [16594.45 ; 16594.45]	1	16594.45 (·) [16594.45 ; 16594.45]	1	16594.45 (·) [16594.45 ; 16594.45]
UCls	2	5030.69 1136.83 [4226.82 ; 5834.55]	2	5030.69 1136.83 [4226.82 ; 5834.55]	2	5030.69 1136.83 [4226.82 ; 5834.55]
UCrm	2	8087.97 (3149.49) [5860.945 ; 10315]	0		2	420.85 (306.94) [203.8 ; 637.89]
Algeria						
UC	5	8831.49 (6235.01) [1082.171 ; 15359.74]	3	10299.52 (6236.71) [3331.818 ; 15359.74]	5	6301.48 (7032.19) [47.49421 ; 15359.74]
UCss	1	3331.81 (·) [3331.81 ; 3331.81]	1	3331.81 (·) [3331.81 ; 3331.81]	1	3331.81 (·) [3331.81 ; 3331.81]
UCls	2	13768.24 220.73 [12176.73 ; 15359.74]	1	15359.74 (·) [15359.74 ; 15359.74]	2	7679.87 10860.98 [0 ; 15359.74]
UCrm	2	6644.58 7866.44 [1082.17 ; 12207]	1	12207 (·) [12207 ; 12207]	2	6103.5 8631.65 [0 ; 12207]

Differently, in Iraq. Morocco and Algeria LS projects remain on average more expensive than SS projects in both scenario B and C. This implies that higher is the scale of the irrigation project and higher is on

USD per hectare for Middle East countries while Afro-Arab countries have produced a mean between 4964,75 USD and 6698,65 USD per hectare. Then, means in Middle-East country are lower than means in Afro-Arab country for any type of irrigation investment. Further, in scenario C, the mean of SS irrigation investment project is 6937,83 USD per hectare for Middle-East countries while the mean for Afro-Arab countries is 7155,54 USD per hectare. Furthermore, the mean of RM irrigation investment project is 2021,69 USD per hectare for Middle-East countries while the mean for Afro-Arab countries is 4840,63 USD per hectare. Finally, also considering irrigation projects as whole, is has been confirmed that on average unit costs are lower for Middle East countries rather than Afro-Arab countries (4638,69 USD versus 6939,34 USD per hectare in scenario C and 3072 USD versus 6972,65 USD per hectare in scenario B).

average the cost paid per hectare. It follows that for an international investor are on average more convenient in these countries SS projects rather than LS projects because the cost per hectare for SS projects is smaller.

Finally, for other countries, such us, Mauritania, Tunisia and Syria since some type of irrigation investment projects are missing at total a comparison between SS projects and LS projects in not possible unless new observations or an enlarged dataset is created or provided. Then, for these countries is not possible to formulate investor policy recommendations.

5. Conclusions

The negative influences of outliers on datasets applications is known for a long time. This paper has proposed an advanced econometric method, applied to a real irrigation investment dataset, to handle this issue in a three scenarios context.

This methodology, that can be generalized to any other dataset, has produced reliable summary statistics on unit irrigation costs at different territorial levels and by type of irrigation investment that will help surely governments, international organizations and water management in implementing more efficient water investment decisions and water policies as well as knowing how much to pay on average or at maximum or at minimum for a given irrigation investment project type located in a given country.

Nowadays, the use of these robust statistical methods is becoming more and more important in many fields and applied sciences because outliers are virtually present in any dataset and they must be handled robustly in order to produce correct policy decisions and recommendations.

6. Further developments

These results raise several issues for future applied research. Four are mentioned here. First, it has been assumed that regressors have not outliers and they are properly measured. These two assumptions have implied first that, given the goodness of the estimated regression line, may be present as only vertical outliers and second that the hectare variable is well reported and measured without a measurement error. Then, vertical outliers has been identified under a linear log-log regression with or without a squared regression term, indeed the most suitable functional form case per case, as those observations that have a robust standardized residual larger than 2.25 in absolute value. Although this methodology has produced reliable, realistic and no distorted summary statistics on unit irrigation costs both regressors and the dependent variable might be poorly reported that means a measurement error in the regressors and or in the dependent. Further, measurement error might be correlated that means an additional challenge to handle. Under horizontal outliers and measurement errors the methodology here explained must be tuned and it represents the first future possible development. Second, outliers have been identified assuming no other variables affect the relationship between the total cost and the hectares. This assumption needed to be confirmed in order to avoid the omitted variable bias that can distort any regression consistently. Third, a Bayesian approach rather than traditional regression to identify potential outliers might be carried out taking advantage from the near normality of the log-data transformation. Further, once outliers have been Bayesian identified also a model based directly on the distributions is possible. Fourth, our results are based on 225 initial irrigation projects in two regions. Would there be additional improvements if we were to use more irrigation projects? Therefore, would there be additional improvements if we were use global data rather than two region country data instead? In other words, the issue of systematically selecting data observations from very many other data observations is a difficult challenge that requires further research. In general, having more information is a plus then it follows that enlarging the starting dataset, thought merging or appending also from different data sources should be needed and it represents the most important further development to carry out.

REFERENCES

Edgeworth F. Y. (1887) On Observations Relating to Several Quantities, Hermathena, 6, 279-285.

Huber P. (1964) Robust Estimation of a Location Parameter, Annals of Mathematical Statistics, 35, 73-101.

Maronna R., D. Martin (2006) Robust Statistics, New York: John Wiley and Sons.

Palmegiani G. (2009) Measuring Cultivation Parcel with GPS: a Statistical Evidence, FAO Wye Conference, June 2009, http://www.fao.org/fileadmin/templates/ess/pages/rural/wye_city_group/2009/Paper_2_b3_Palmegiani_Measuring_cultivation_parcels_with_GPS.pdf

Pizzoli E., G. Palmegiani (2007) Rural Development Rural Development Statistics for Policy Monitoring: a Rural-Urban Territorial Classification and Farmers' Income Data, FAO 20th AFrican Commission of Agriculture Statistics (AFCAS), Algers, http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/Workshop_Algiers_December_2007/AW-07-03-1-Rural_Development_Statistics_and_Farmers_Income_Data_for_Policy-Pizzoli.pdf

Rousseeuw P. J., A. Leroy. (1987) Robust Regression and Outlier Detection, New York: John Wiley and Sons.

Rousseeuw P. J., V. Yohai (1987) Robust Regression by Means of S-estimators, Robust and Nonlinear Time Series Analysis, Springer 256-272.

Yohai V. (1987) High Breakdown-point and High Efficiency Estimates for Regression, The Annals of Statistics 15, 642-665.