

Measuring cultivation parcels with GPS: a statistical evidence¹

Gabriele Palmegiani

Affiliation: LUISS University of Rome,

Address: Viale Romania 32, email: gabriele.palmegiani@gmail.com

Abstract: This paper uses African survey data (2005-2006), to study the statistical relevance of measuring surfaces of cultivation parcels using Global Positioning System (GPS) respect to the traditional method using compass and meter. Cultivation parcel selection was not random. Cameroon, Niger, Madagascar and Senegal were involved. More types of GPSs were used.

Before supposing that selection was random, the *unconditional inference* approach is provided for both parametric and non-parametric level. The *paired t-test* and the *Wilcoxon sign rank test* are applied on measurement differences.

After assuming that selection was not random, the *conditional inference* approach for means based on resampling methods is applied. The permutation distribution function of the paired *t*-statistic and the empirical p-values associated are worked out for differences. Two main conclusions are found: first, the conditional inference fully supports the unconditional one, only parcel estimates using GPS60 are statistically equivalent to parcel estimates using traditional method and the lose of accuracy when random sampling is assumed is on the order of 2/1000. Second, parcel estimates using compass tend to be larger than parcel estimates using GPSs. In conclusion, because the GPSs methods are globally cheaper than compass method, is strongly recommended the use of GPS60 to reduce the costs of the agricultural surveys.

Keywords: Random Sampling, Conditional Inference, Independence, Normality, Variance Homogeneity, *t*-tests, Wilcoxon tests, Resampling methods, Permutations tests.

1. Introduction: statistical inference and random sampling for GPS measurements

Statistical inference cannot have nothing to do with the *sampling*. The sampling is that part of statistics concerned with the selection of observations intended to describe some features about the population of interest (target population). Typically, only samples from a given target population are available. Most of the inferential methods require the assumption that the samples have been generated by a random mechanism. It insures that, samples are *random variables*, indeed, a set of values drawn independently from a larger population or *uniform random variables* when all members of the target population have an equal chance of being drawn. Common statistical tests, as *t*-tests assume also that the target population is normally distributed. Often, observations are not a random sample

¹ I wish to express my most sincere gratitude to Mr. Hiek Som (FAO, ESSS) and Mr. Naman Keita (FAO, ESSS) for their availability and supervision in writing this paper. I do warmly appreciate their contributes and comments. For questions or suggestions concerning this paper fell free to contact me.

from a well-defined target population. For example, the one sample extraction without replacement of m balls from an urn containing M balls, with $m < M$, is a different experimental design respect to the individuals recruited to be FAO volunteers. In the first case, the balls are extracted randomly from the urn and each of them is drawn independently from the others. Researchers can make inference on the all sample space variability of the target population, then the statistical inference is called *unconditional* or *simple inference*. Common tests, such as t -tests may be used. In the second case, FAO's volunteers are hardly ever a random sample from a set of all possible candidates having at least a degree, but are a strict selection of resourceful and successful students who have a specific educational background. Besides, for a given division, often volunteers are taken with different skills, then probably they are not independently chosen. Intentional selection exists and some bias must appear. Now, inference can be done only on a restricted part of the sample space associated with the conditioning event of interest, then the statistical inference is called *conditional inference*. In this case, t -tests cannot work well and *resampling methods* for inference and parameters estimation are strongly recommended (Pesarin (2001)).

In practice, researchers may take the random sampling assumption when the selection mechanism does not guarantee randomness. Inferences from data that are chosen according to a given intentional selection are inevitably less secure than where we have random samples. The lose of accuracy may become greater when the samples are smaller. We shall make inference assuming that cultivation parcels are both random and non-random sampled.

The aim of the paper is to verify whether estimation of surfaces of cultivation parcels using traditional method and using types of GPSs gives different results. Then, because the GPS methods are globally cheaper than traditional method, the equivalent GPS measurement, if found, will be strongly recommended to reduce the costs of agricultural surveys.

Given different types of measurements, the *statistical equivalency* between two methods, might be evaluated respect to the surfaces obtained, the time requested, the costs undertaken and the weather conditions presented. Obviously, the equivalence depends on both the size of the parcels considered and the staff ability. It means that *error measurements* might be treated as a linear or nonlinear function of both the parcel extension and the number of training staff hours.

In this paper, the statistical equivalency will be studied only in the results or surfaces obtained. Immediately after this preparatory section, we shall describe the data-set (*Section 2*), then the statistical inference is runned assuming both random parcel selection (*Section 3*) and non-random parcel selection (*Section 4*). For the random sampling case, starting from a simple statistical inference framework (*Subsection 3.1*) we shall select parametric and non-parametric suitable tests. We shall provide both their theory (*Subsection 3.1 & Subsection 3.2*) and their practical application (*Subsection 3.3*). After, relaxing the random selection assumption, a conditional inference approach based on resampling methods is applied. The permutation distribution of the paired t -statistic will be estimated from the statistical units surveyed and unconditional and conditional inference for means will be faced toward (*Section 4*). Final conclusions will be summarized in (*Section 5*) and further developments will be explained in (*Section 6*).

2. The data-set

The GPS survey is entirely built and collected by FAO's Statistical Division in the framework of the project GCP/INT/903/FRA. Cameroon, Niger, Madagascar and Senegal were involved. Measurements on surfaces, perimeters, weather conditions and time requested were taken for different methods of measurements and for each of the 207 cultivation parcels. Since many of these samples are missing for Madagascar, we have chosen to drop out its observations. The total number of observations is reduced to 157. Deleted the possible outliers the applied analysis will be conducted on 126 observations. The covariates of interests (all expressed in squared meters) are the following:

- **S_1** = Compass and meter cultivation parcel surface.
- **S_21_1** = Garmin60 (GPS60) cultivation parcel surface.
- **S_22_1** = Garmin72 (GPS72) cultivation parcel surface.
- **S_24_1** = Magellan400 (MAG400) cultivation parcel surface.

These variables take into account of both bigger and smaller parcels together and only first passages measurements. From these covariates, we shall interested in the following differences:

- **c_g60diff** = the difference between compass and meter cultivation parcel surface and Garmin60 cultivation parcel surface.
- **c_g72diff** = the difference between compass and meter cultivation parcel surface and Garmin72 cultivation parcel surface.
- **c_m400diff** = the difference between compass and meter cultivation parcel surface and Garmin72 cultivation parcel surface.

which will be studied using both the unconditional and the conditional inference approach.

3. Unconditional inference

This section makes statistical inference assuming that cultivation parcel selection was random. It implies that samples are random variables drawn independently from a larger population. We can assume or not normality of the differences. When normality is supposed we shall use tests on means, when normality is not assumed tests on *pseudo-medians* (indeed, medians worked out starting from the ranks of the original samples) will be carried out. Described the general unconditional statistical inference framework (*Subsection 3.1*) two methods will be compared: a parametric level (*Subsection 3.2*) and a non-parametric one (*Subsection 3.3*). In the former, normality is supposed and the *t*-tests may work well. In the latter, normality is relaxed and the *Wilcoxon sing rank test* is runned. The practical application of these tests is contained in (*Subsection 3.4*).

3.1 How can we make simple inference?

Supposing that covariates are random variables, the unconditional statistical inference can be done according to the (Table 1)². As we can see, statistical inference depends on both the aim and the nature of the data that we are facing. On the data hand, choosing the right test to compare measurements implies a selection between two families of tests: *parametric tests* and *non-parametric test*.

Table 1: *Unconditional statistical inference: the aim and the data nature.*

Statistical aim	Data-set features			
	Continous Data		Non-Continous Data	
	Measurements from a Gaussian distribution	Measurements, Ranks or Scores, from a non-Gaussian distribution	Binominal Data (two possible outcomes)	Survival Data
Describe one sample	Mean Standard deviation	Median Interquartile range	Proportion test	Kaplan Meier Survival Curve
Compare one sample to a hypothetical value	One sample <i>t</i> -test	One sample Wilcoxon test	Chi-square test Binominal test	
Compare two unpaired samples	Unpaired <i>t</i> -test	Unpaired Wilcoxon test (Wilcoxon Rank Sum test)	Fisher's test (chi-square test for large samples)	Log-rank test
Compare two paired samples	<u>Paired <i>t</i>-test</u>	<u>Paired Wilcoxon test</u> (Wilcoxon Sign Rank test)	McNemar's	Hazards regression
Compare three or more unmatched samples	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox-regression
Compare three or more matched samples	Repeated-measures ANOVA	Friedman test	Cochrane Q	Hazards regression
Quantify association between two variables	Pearson correlation	Sperman correlation	Contingency coefficients	
Predict value from another measured variable	Linear regression Non-linear regression	Non-parametric regression	Logistic regression	Cox-regression
Predict value from several measured or binominal variables	Multiple linear regr Multiple non-linear regression		Multiple logistic regression	Cox-regression

² When the same test is pointed out, it means the application of the same test under different assumptions.

Parametric tests are based upon the assumption that the samples are drawn from a well defined probability distribution. Often, the Gaussian distribution is assumed. Commonly used parametric tests are listed in the first column of the (*Table 1*).

Because normality is a strong assumption, as starting point it is fundamental verify if normality works well using *histograms of frequencies distributions* or *normal quantile probability plots*. All commonly used non-parametric tests rank the outcome variable from low to high and then analyze the ranks. These tests are listed in the second column of the (*Table 1*) and include the paired and unpaired Wilcoxon test.

Choosing between parametric and nonparametric tests is sometimes easy. We should definitely choose a parametric test if they are sure that our samples are drawn from a population that follows a Gaussian distribution (at least approximately). Instead, when the samples are not normally distributed or the outcome is a rank or some outliers remain is strongly recommended the use of a non-parametric framework. Anyway, cause the *central limit theorem* (CLT), parametric tests work well with large samples even if the population is not Gaussian³.

Data may be also non-continuous. Inference on binominal data (two possible outcomes) or survival data (time to event data) requires the tests listed in the third and fourth column of the (*Table 1*). We do not deal with these latter tests.

On the other hand, the statistical inference depends also by the goal of the analysis.

Given only one sample, for describing, the mean or the standard deviation might be useful. Instead, to make simple inference on the mean, one sample *t*-test should be appropriate.

When samples are two, we need to decide whether to use a *paired test*. To compare three or more samples, the term paired is not adapt and the term repeated samples is used instead. Paired observations are found when two measurements are made on the same statistical unit. In this case, we might expect a correlation between two measurements, either because they were made on the same individual or were taken from the same location. Pairing is effective only when the sample correlations are not weak. When it works, two samples from a given data-set must be handled as paired, then the differences should be taken for inference. The pairing reduces the degrees of freedom of the test statistic. The conclusions may be different, then it would be seriously inappropriate to analyze samples without taking the pairing into account when it is effective (Greene 2008).

When samples are more than two, they may be grouped or matched. In this case, the ANalysis Of VAriance (ANOVA) and repeated ANOVA methods can be used to compare unmatched and matched samples respectively.

³ In most situation, a vast number of possible samples could have been taken from a particular population. Each sample may have a different value for its mean. The distribution of these possible samples means is called the *sampling distribution of the mean*. The CLT states that, for a population with finite mean μ and finite standard deviation σ , the sampling distribution of the mean can often be well approximated by a normal distribution whose mean is μ and whose standard deviation is σ / \sqrt{n} . This result depends strongly on both that the n observations in the sample have been selected independently of each other and on the size of n . When, for one sample $n \geq 30$ or for two samples $n_1 + n_2 \geq 30$, in practice, we do not need to worry too much about the normality assumption. To avoid distribution mistakes, it is often suggested a safer threshold; $n \geq 50$ for one sample and $n_1 + n_2 \geq 50$ for two samples.

Finally, the association between two variables can be studied through correlation measures, instead simple linear regression (SLR) or multiple linear regression (MLR) can be applied for forecasting.

Given that inference framework, to assess the relevance of measuring cultivation parcels using GPS we need to choose the right test.

On the data hand, the data-set is constituted by *continuous* and *independent samples*. Independence assumption is satisfied, because the parcel measuring using a method cannot influence the same measuring using another method. Obviously, samples are continuous, because values observations varies in a continuous way. Then, the choice is reduced at the first two column of (*Table 1*).

On the statistical aim hand, we are interested in comparing couples of measurements. Basically, the comparison may be done in a parametric framework using sample means or in a non-parametric framework using sample medians. Besides, we might work on unpaired or paired samples. Let's see now, how that works.

At parametric level when pairing is not effective, assuming that two samples are drawn from two normal distribution functions with $N(\mu_1; \sigma_1^2)$ and $N(\mu_2; \sigma_2^2)$, the system of the hypothesis is the following:

$$\begin{aligned} H_0; \mu_1 &= \mu_2 \\ H_A; \mu_1 &\neq \mu_2 \end{aligned} \quad (1)$$

Where μ_1 and μ_2 are the unknown theoretical or population means, estimated using the samples counterparts. This hypothesis can be tested using the so called *unpaired two sample t-test* which can be reduced to only one sample unpaired *t-test* on the differences.

When pairing is effective the two populations problem is reduced to only one population problem. Given *i*-th observation and two samples measurements X_{1i} and X_{2i} , the variable of interest becomes the paired differences $d_i = X_{1i} - X_{2i}$. Assuming that differences are normally distributed $N(\mu; \sigma^2)$, the system of the hypothesis becomes:

$$\begin{aligned} H_0; \mu &= 0 \\ H_A; \mu &\neq 0 \end{aligned} \quad (2)$$

Where μ is the unknown theoretical or population mean, estimated using its sample counterpart. This hypothesis can be tested using the so called *paired t-test*.

At non-parametric level, means are substituted by cumulative distribution functions. When pairing is not effective, the system of hypothesis is the following:

$$\begin{aligned} H_0; F_{X_1} &= F_{X_2} \\ H_A; F_{X_1} &\neq F_{X_2} \end{aligned} \quad (3)$$

Where F_{x_1} and F_{x_2} are two unknown population distribution functions, which can be estimated by their empirical distribution counterparts. When pairing is effective, the system of hypothesis becomes:

$$\begin{aligned} H_0; F &= 0 \\ H_A; F &\neq 0 \end{aligned} \tag{4}$$

is the unspecified population distribution function, which can be estimated by its empirical distribution counterpart. We shall use two kind of non-parametric test: the *Wilcoxon sign rank test* and the *permutation paired t-test*. In the first case, a pseudo-moment of these distributions is taken⁴.

In the second case, the location of the empirical counterpart of the distribution F will be studied.

The paring has a significant impact on inference. First, since only a measure of variability is present, tests on variances have not any sense to exist. Second, given the significance level, since the pairing reduces the number of degrees of freedom, the test statistic tends to be higher and the p-value associated smaller, then we are boosted to reject the null hypothesis that measurements methods are, on average, statistically equivalent when it is true.

There are two simple tests for comparing two samples:

- **Student's t -tests:** Can be used when samples are independent, their variances are similar and finite and their distribution is normal or near normal.
- **Wilcoxon tests:** Can be used when samples are independent, the variances are similar and finite, but their distribution is not normal or not near normal distributed.

These tests can be applied to one and two samples problems as well as to paired and to non-paired data. For unpaired problems, the t -test is known as *two sample t-test* and the Wilcoxon test is known as *Wilcoxon rank sum test*. For paired problems, the t -test is known as *paired t-test* and the Wilcoxon test is known as *Wilcoxon sign rank test*. The tests underlined in (*Table 1*) will be used to make simple inference.

3.2 The Student's t -tests

Student was the pseudonym of W.S. Gosset who published his famous paper in *Biometrika* in 1908. He was prevented from publishing under his own name, cause an employment law in place at the time, which allowed his employer to prevent him publishing his ideas as independent work. His distribution, the t -distribution, will be perfected by R.A Fisher, who called the distribution Student's distribution, has revolutionized the study of small sample statistics.

⁴ The moment used by the Wilcoxon sign rank test will be pseudo-medians, that is, medians calculated starting from the ranks of the samples.

Given two random samples, X_1 and X_2 , to test the null hypothesis (*Equation 1*) we can use the so called the *two sample t-statistic*:

$$t = \frac{\text{difference between two means}}{\text{s.e of the mean difference}} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{(\bar{X}_1 - \bar{X}_2)}} = \frac{\bar{X}_1 - \bar{X}_2}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{d} t_{n_1 + n_2 - 2} \quad (5)$$

Where \xrightarrow{d} means “distributed in law”, \bar{X}_1 and \bar{X}_2 are the samples means⁵, S is the sample standard deviation⁶, n_1 and n_2 are instead the samples size. This formula counts the number of standard errors of the mean difference $SE_{(\bar{X}_1 - \bar{X}_2)}$ by which the two sample means are separated. Under the null hypothesis (*Equation 1*), the t -statistic has a Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. But if samples are sufficiently larger⁷, the CLT can be applied:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{(\bar{X}_1 - \bar{X}_2)}} \xrightarrow{a} N(0;1) \quad (6)$$

Where \xrightarrow{a} means “asymptotically distributed”. This happens because the t -distribution depends on the number of degrees of freedom associated with the denominator $SE_{(\bar{X}_1 - \bar{X}_2)}$. Because $n_1 + n_2 - 2$ degrees of freedom have been used to calculate the standard deviation S , the t -statistic has a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. When the samples size increases, also the degrees of freedom increases then the samples standards deviations gives an increasingly good approximation to the populations standards deviations; thus the t -statistic becomes more and more like a standard normal random variable.

⁵ The sample mean for the k -th sample is defined as:

$$\bar{X}_k = n_k^{-1} \sum_{i=1}^{n_k} X_{2k} \quad ; \quad \text{for } k = 1;2$$

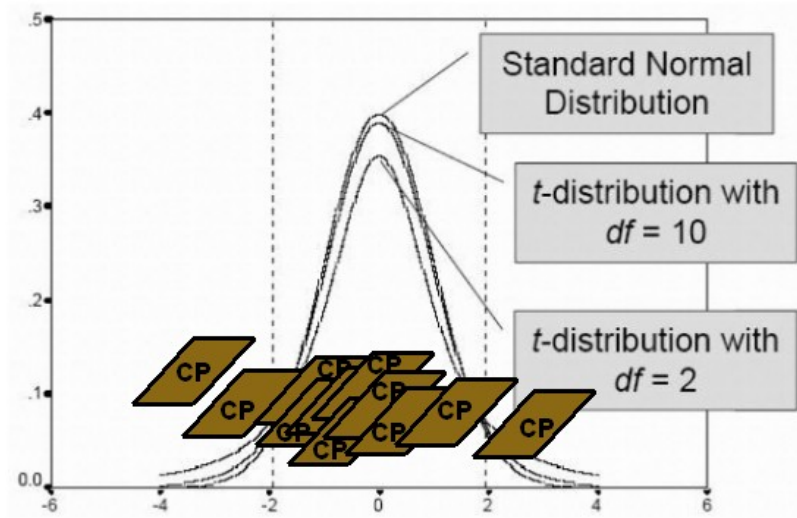
Where X_{2k} are the values of k -th sample.

⁶ The sample standard deviation for the k -th sample is defined as:

$$S_k = \sqrt{\frac{\sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)^2}{n_k - 1}} \quad ; \quad \text{for } k = 1;2$$

⁷ We said $n_1 + n_2 \geq 50$ is safer.

Figure 1: Standard normal distribution and Student-t distribution.



The main difference between a standard normal variable and a Student-*t* distribution is on the *tails* (Figure 1). The Student-*t* distribution is less concentrated around the mean than is the normal distribution and more spread out in the tails, with the difference greatest when the number of degrees of freedom is small than almost ten. Assuming that differences of cultivation parcel estimates are distributed like a normal distribution, means that, these observations (represented using the brown rectangles) are more concentrated around a common mean and less concentrated on the tails.

The total variability of the *t*-distribution has two sources: the sampling variability of the mean difference and the sampling variability of $SE_{(\bar{x}_1 - \bar{x}_2)}$. Since the denominator involves the variability of the samples captured by their standard deviations, it is fundamental for inference. Let's see now, how these work.

When samples have the same variance is reasonable to estimate the standard error of the mean difference treating the variability of the samples together:

$$S = S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

Where S_p is the so called *pooled standard deviation* estimate, S_1 and S_2 are the samples standards deviations. Substituting this formula in (Equation 5) and calculating the samples means, we can obtain the value of the statistic.

A $100(1 - \alpha)\%$ confidence interval for the mean difference is constructed as:

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha; n_1 + n_2 - 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where α is the significance level of the test. It implies that $t_{\alpha; n_1+n_2-2}$ is the percentage point of the t -distribution such that the cumulative distribution function $P_r(t \leq t_{\alpha; n_1+n_2-2})$ equals $1 - \alpha / 2$.

When two samples have different variances we have *heterogeneity in variance*, instead. In this case, to estimate the standard error of the mean difference is better treating the variability of the samples in a separated way. A modified form of t -statistic, known as the Welch test (1949), may be used:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{(\bar{X}_1 - \bar{X}_2)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{d} t_g \quad (7)$$

Where it has been substituted the pooled standard deviation with the samples standards deviations. Under the null hypothesis (*Equation 1*) this statistic can be well approximated by a Student t -distribution with g degrees of freedom, where:

$$g = \left[\frac{c}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1} \right]^{-1}$$

With:

$$c = \frac{S_1^2 / n_1}{S_1^2 / n_1 + S_2^2 / n_2} \in \mathfrak{R}$$

a scalar (it belongs to real space of dimension one).

When pairing is effective, we said that the two sample test is reduced to only one sample test on the paired differences $d_i = X_{1i} - X_{2i}$. In this situation, test on variances have not any sense to exist, because only a measure of variability is present; the standard deviation of paired differences.

To test the null hypothesis (*Equation 2*) we can use the so called *paired t-statistic*:

$$t_{paired} = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{SE_{(\bar{X}_{1i} - \bar{X}_{2i})}} = \frac{\bar{d}}{SE_{\bar{d}}} = \frac{\bar{d}}{S_{\bar{d}} / \sqrt{n}} \xrightarrow{d} t_{n-1} \quad (8)$$

With:

$$\bar{d} = n^{-1} \sum_{i=1}^n d_i \quad ; \quad S_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

Where d_i is the i -th pair difference, \bar{d} is the mean of the paired differences, $S_{\bar{d}}$ is the standard deviation of the paired differences and n is the number of pairs. Under the null

hypothesis (Equation 2), the paired t -statistic follows a t -distribution with only $n-1$ degrees of freedom.

A $100(1-\alpha)\%$ confidence interval for μ can be constructed by:

$$\bar{d} = t_{\alpha; n-1} S / \sqrt{n}$$

Where $\bar{d} = t_{\alpha; n-1} S / \sqrt{n}$ is the percentage point of the t -distribution such that the cumulative distribution function $P_r(t \leq t_{\alpha; n-1})$ equals $1-\alpha/2$. Since:

$$n-1 < n_1 + n_2 - 2$$

the number of degrees of freedom is reduced respect to the two samples case.

3.3 The Wilcoxon tests

The t -tests are based on the main assumption that samples are independent random variables drawn from a larger normal population. We might prefer a non-parametric test if we doubt about the normal assumption of the differences. These tests were proposed by Wilcoxon(1945) for paired and unpaired independent samples. Conversely to the t -test case, here we shall comparing ordered statistics and not samples values.

When the samples are unpaired, the non-parametric alternative to the two sample t -test, is the so called, Wilcoxon rank sum test. Rank-based approaches proceed by transforming the raw data-set into ordered statistics or ranks, one for the smallest value up to the sample size for the largest. To see how the rank sum approach works, consider for example, the following data-set measured on males and females:

$$\text{Raw data} = \begin{pmatrix} 10 & M \\ 20 & M \\ 15 & M \\ 12 & M \\ 9 & F \\ 11 & F \\ 8 & F \end{pmatrix} ; \text{Ranks} = \begin{pmatrix} 3 & M \\ 7 & M \\ 6 & M \\ 5 & M \\ 2 & F \\ 4 & F \\ 1 & F \end{pmatrix}$$

Where the male sample size is greater than the female sample size. The males have ranks of 3, 7, 6, 5 while the females have ranks of 1, 2, 4. On the sum of these ranks the test statistic is built. It implies that, the proper values of the samples do not enter in the analysis. We are discounting unusually original values (the largest and the smallest values) keeping its path.

Given two samples with sizes n_1 and n_2 , with $n_1 > n_2$, the *Wilcoxon rank sum test statistic* is computed as:

$$\begin{aligned}
WRS_2 &= n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum_{i=1}^{n_2} R_i \\
&= n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - G_{n_2}
\end{aligned} \tag{10}$$

Where R_i is the rank of the i -th point observation and G_{n_2} is the sum of all ranks for the smaller sample. The meaning of the test statistic is the following:

$$WRS_2 = \overbrace{n_1 n_2 + \frac{n_2(n_2 + 1)}{2}}^{\text{the maximum value of } G_{n_2}} - \underbrace{G_{n_2}}_{\text{the sum of ranks for the smaller sample}}$$

It measures the distance between the sum of all ranks for the smaller sample and its maximum. Since the distribution of ranks under the null hypothesis (*Equation 3*) has been tabulated we can extract exact p-values also when the samples are relatively small. When the sample sizes increase, a standard normal approximation is possible:

$$\frac{WRS - \mu_{WRS}}{\sigma_{WRS}} \xrightarrow{a} N(0;1) \quad \text{as} \quad n_1 \cup n_2 \rightarrow \infty$$

Where:

$$\mu_{WRS} = \frac{n_1 n_2}{2} \quad ; \quad \sigma_{WRS} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

It works for very small values of the sample sizes. If n_1 and n_2 are both equal to or greater than five this approximation works well. This approximation is better when n_1 and n_2 are both equal to or greater than ten. It becomes very powerful when n_1 and n_2 are both equal to or greater than twenty-five.

When the samples are paired, the non-parametric alternative to the paired t -test, is the so called, *Wilcoxon sign rank test* known also as Wilcoxon matched pairs test. The step ahead of rank-based sign approaches respect to the rank-based sum approaches is to take care about the sign of the rank differences (*Table 2*). To test the null hypothesis (*Equation 4*) we can use the *Wilcoxon signed rank statistic*:

$$WSR^+ = \sum_{i=1}^n 1\{d_i > 0\} R_i \tag{11}$$

Where n is the number of pairs and $1\{\cdot\}$ is the *indicator function*:

$$1\{\cdot\} = \begin{cases} 1 & ; \quad \text{if } d_i > 0 \\ 0 & ; \quad \text{if } d_i < 0 \end{cases}$$

which permits only the summation of all positive ranks of the differences. When the number of pairs n increases, a standard normal approximation works. For $n < 20$, exact probabilities can be calculated, for $n > 20$ the standard normal approximation is used. Due to the fact that the Wilcoxon tests are non-parametric tests, no confidence intervals are allowed.

The advantages of non-parametric tests versus parametric tests are a contentious issue. If the assumptions of the parametric test are fulfilled, then it will be somewhat more efficient, on the order of 5% in large samples, although the difference can be larger in small samples. Anyway, samples independence must be reached. The main disadvantage of the Wilcoxon tests are the problems of *ties*. When several observations share the same value, the average of the tied ranks is used. For example, observations five and six in (Table 2) share the same value of the differences and their rank assigned is a average value. This is not a problem for the large sample normal approximations, but the exact small sample distributions becomes much more difficult to calculate.

Table 2: The rank sign approach: the signed rank differences.

i -th obs.	X_{1i}	X_{2i}	$d_i = X_{1i} - X_{2i}$	$ d_i = X_{1i} - X_{2i} $	rank of $ d_i $	signed rank d_i
1	78	78	0	0	-	-
2	24	24	0	0	-	-
3	64	62	2	2	1	+1
4	45	48	-3	3	2	-2
5	64	68	-4	4	3.5	-3.5
6	52	56	-4	4	3.5	-3.5
7	30	25	5	5	5	+5
8	50	44	6	6	6	+6
9	64	56	8	8	7	+7
10	50	40	10	10	8.5	+8.5
11	78	68	10	10	8.5	+8.5
12	22	36	-14	14	10	-10
13	84	68	16	16	11	+11
14	40	20	20	20	12	+12
14	40	20	20	20	12	+12
14	40	20	20	20	12	+12
15	90	58	32	32	13	+13
16	72	32	40	40	14	+14
						$WSR^+ = 67$ $n = 14$

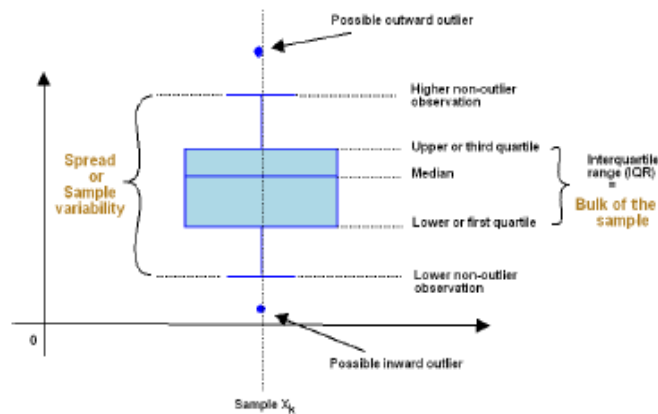
3.4 The statistical analysis

Working on differences, as a preliminary work we have checked both the existence of *outliers* and the satisfaction of all assumptions for applying correctly a t -test.

To check outliers, scatter plots matrix and a box plots are provided. The scatter plot matrix is the multivariate innovation of a scatter plot. A scatter plot visualizes a relation between two covariates, where point observations are represented as pairs in a two-dimensional space. The scatter plot matrix arranges scatter plots in a matrix format. The box plot (Figure 2) is a graph capable of displaying five characteristics of a given sample (labelled by X_k , where k denotes the k -th sample) in the same picture: the

largest non-outlier observation, the upper or third quartile, the median, the lower or first quartile and the smallest non-outlier observation. Outliers are points observations which appear to be inconsistent with the remainder of the data. These point observations having a negative impact on the statistical inference accuracy, must be deleted before the applied analysis. The box plot is a useful tool to inspect these points. Are considered outliers points observations which belong above the largest non-outlier observation or below the smallest non-outlier observation. These abnormal points are labelled by a point.

Figure 2: *The box plot: the meaning.*



Apart possible outliers, this graph is able to explain other sample characteristics. The distance between the upper and the lower non-outlier observation gives information on both the *spread* or *variability* of the sample and its *tail length*. The distance between the third and the first quartiles⁸, the so called *interquartile range*:

$$IQR = Q_3 - Q_1$$

provides information on the *location* of the sample bulk. The median, shown by an horizontal line, is a number, which provides information on both the location of 50% of the data and the sample *skewness*.

As we were expecting, the scatter plot matrix (*Figure 3*) shows that differences are concentrated around the zero, but some pairs are too far from the bulks of the samples. These points may be outliers or measurements errors. This evidence is confirmed by the box plot (*Figure 4*) where the bulks are crushed by some influential values represented by the isolated points. Dropping these inconsistent values, the size of the samples is reduced and both the scatter plot matrix (*Figure 5*) and the box plot (*Figure 6*) enhance their exploratory power. As we were expecting, the sample differences produces similar graphs, but some departures can be inspected. The scatter plot matrix (*Figure 5*) shows that the point concentration is more diluted when Magellan400 is involved.

⁸ Quartiles divides the sorted samples into quarters. For example, the first quartile, Q_1 , is a number, it is greater than 25% of the sample cases, and lower than the remaining 75%.

Table 3: *Sample differences: summary statistics.*

Sample Differences	Observations	Mean	Standard dev.	Minimum	Maximum
c_g60diff	88	17.6208	227.6341	-529	577
c_g72diff	126	102.1954	293.5093	-742	884
c_m400diff	126	144.014	277.6434	-436	1156.34

Figure 3: *Outliers check of the differences: the scatter plot matrix.*

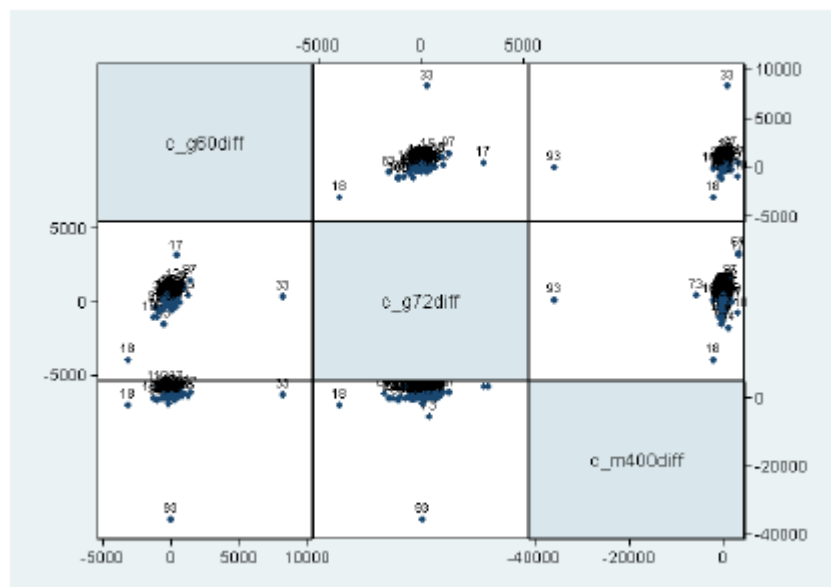
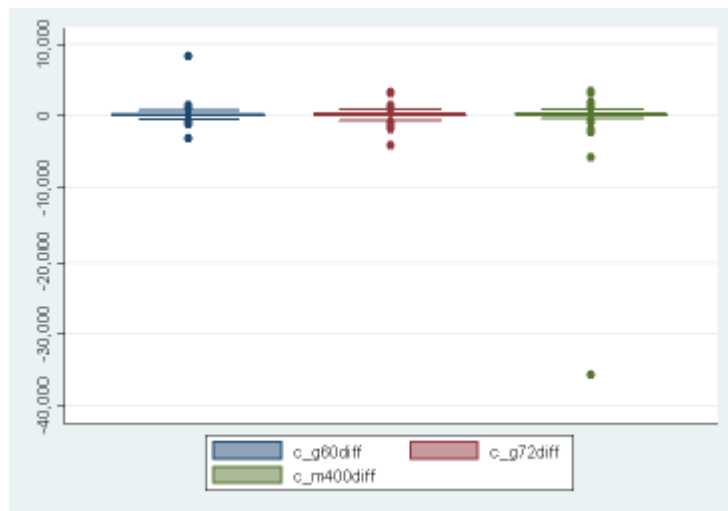


Figure 4: *Outliers check of the differences: the box plot.*



The box plot (Figure 6) shows that sample differences do not seem to differ much in their: spread, location, *skewness* and tails length. It is confirmed by inspecting their summary statistics (Table 3). Anyway, the better results are reached for the surface difference between the traditional method and Garmin60 method. Its median near to zero is located in the middle of the sample bulk. Then, since the distance between the upper non-outlier and the lower non-outlier observation is the smallest, the variability should be the shortest. For all samples, remains the problem of possible outliers in the differences. We would have to drop other point observations, but we risky to losing significant sample variabilities. The permanence of outliers suggests the use of non-parametric tests.

Figure 5: Outliers check of the differences: the scatter plot matrix of filtered samples.

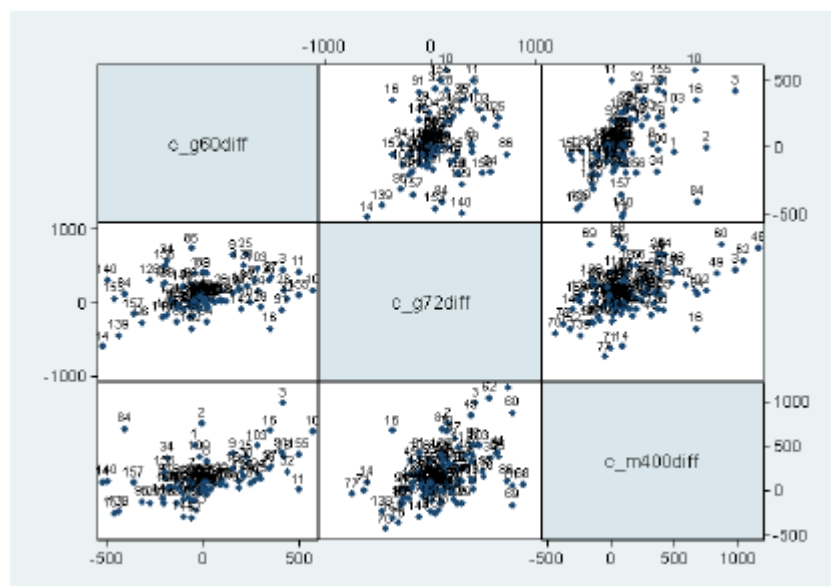
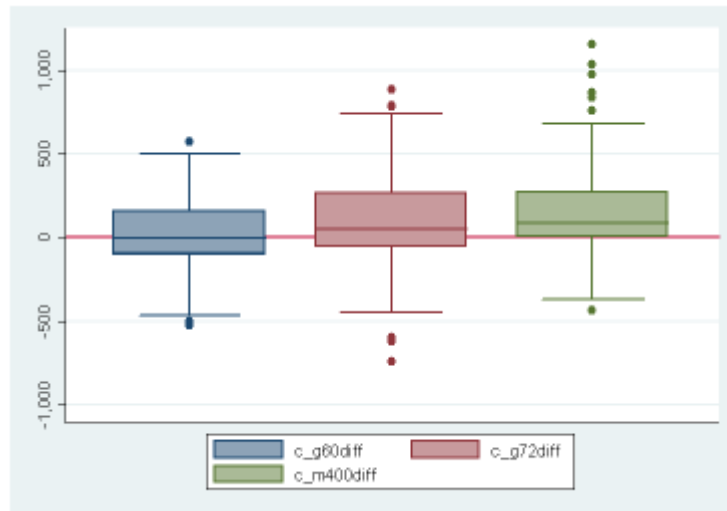


Figure 6: Outliers check of the covariates: the box plot of filtered differences.



Dropped the outliers we have verified for differences if the assumptions for applying correctly a t -test are satisfied.

First, we have tested if normality works well. To check normality, histogram plots of the frequency distribution (*Figure 7*) and normal quantile probability plots on defined residuals (*Figure 8*) have been used. If the normality assumption is satisfied, the frequency distributions should be distributed like a normal and the quantiles of the residuals should be linearly related to the quantiles of the normal distribution. These figures suggest that both Garmin60 and Garmin72 difference may be considered near normally distributed, but for Magellan400 difference the normality is further. This departure from normality, although tempered from the CLT, suggests the use of non-parametric tests.

Second, samples must be independent. Here independence is guaranteed, because is reasonable to assume that measurements cannot influence each other. It means that, for example the measurement using compass cannot influence the measurement using Garmin72. It implies that the variance of the mean difference between two methods equals the sum of two sample variances.

Third, samples variances should not differ significantly. But since the samples must be considered as paired, tests on variances have not any sense to exist, only a measure of variability is present. Otherwise, when the samples are unpaired, each of them has a measure of variability and tests on variances based on F test are strictly necessary to decide between a t -test based on the pooled standard deviation or a modified t -test based on the Welch approximation (*Equation 7*).

Figure 7: Normality check of the differences: histogram plots of frequency distribution.

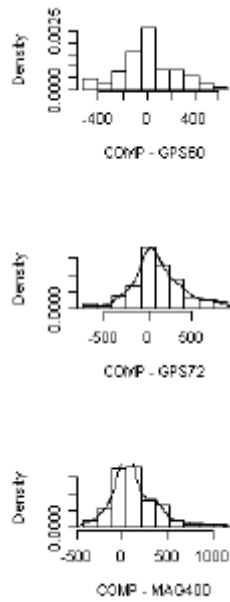
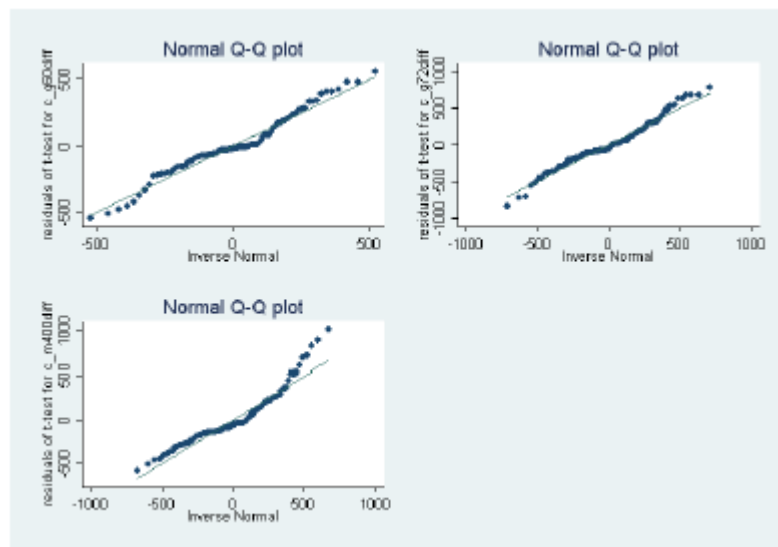


Figure 8: Normality check of the differences: normal quantile probability plots.



The statistical inference must be performed according to these difference features summarized in (Table 4). Usually tests, as t -tests, can work well only when all of these assumptions (at the right of Table 4) are satisfied. The permanence of outliers and the departure from normality, suggests the use of non-parametric methods. We shall provide conclusions for both parametric and non-parametric methods. All tests are conducted at 5% significance level.

At parametric level differences are assumed normally distributed, although some departure from normality appears. To assess the statistical equivalency for

unconditional means, the paired t-test (*Equation 8*) on samples differences is applied (*Table 5*).

Table 4: *t*-test assumptions and features found.

<i>t</i> -test assumptions	Features found
1) Normality (measurements must be normal or near normal distributed)	1a) c_g60diff can be considered near normal. 1b) c_g72diff can be considered near normal. 1c) c_m400diff cannot be considered normal.
2) Independence (measurements cannot be influence each other)	2a) It is satisfied, because measurements cannot influence each other.
3) Random Sampling (measurements must be random variables) (random selection avoids conscious or unconscious bias)	We may assume that the selection was: 3a) Random → <i>t</i> -tests work well 3b) Not random → <i>t</i> -tests do not work well → use permutation tests
4) Variance Homogeneity (measurements must have the same variance)	We must assume that samples are paired: → tests on variances have not any sense → the standard deviation of paired differences S_d is assumed as the only measure of variability
5) No Outliers (all possible outliers must be deleted before the applied analysis)	5a) Some evidence of outliers remains.

In Panel A, traditional method and Garmin60 method are compared. Since the p-value is greater than the significance level ($0.4697 > 0.05$), we cannot reject the null hypothesis (*Equation 2*) that the true difference in means is equal to zero, then we can conclude stating that parcel estimates using Garmin60 *are not statistically different* from parcel estimates using traditional method.

In Panel B, traditional method and Garmin72 method are faced toward. Since the p-value is now smaller than the significance level ($0.0001514 < 0.05$), we can reject the null hypothesis and we can conclude stating that parcel estimates using Garmin72 are statistically different from parcel estimates using traditional method.

In Panel C, traditional method and Magellan400 method are compared. As was happening for Garmin72, the p-value is smaller than the significance level ($4.601e-08 < 0.05$), we can reject the null hypothesis and we can conclude stating that parcel estimates using Magellan400 are statistically different from parcel estimates using traditional method.

The p-value of the test communicates the different being, but nothing about the level and the sign of the parcel estimates departure. Then, to move from statistical equivalency towards the relatively size of the parcels, we need to take care of the sign of mean of the differences.

When the measurement methods are statistically different (it means, the p-value is not significative), this value and its sign is exploratory. Otherwise, for equivalent measurements this value and its sign loses its explicative power, because the parameters estimates are in their confidence intervals. In other words, when statistical equivalency works nothing can be stated on the relatively size of the parcels, they must be considered identical.

Table 5: Paired *t*-test on sample differences: the parametric result.

Paired <i>t</i> -test	
Panel: A	
data: S.1 and S.21.1	
$t_{paired} = 1.738$; $df = 87$; p-value = 0.4697	
alternative hypothesis: true difference in means is not equal to 0	
95 percent confidence interval:	
-30.61026	65.85185
mean of the differences :	
17.62080	
Panel: B	
data: S.1 and S.22.1	
$t_{paired} = 3.9084$; $df = 125$; p-value = 0.0001514	
alternative hypothesis: true difference in means is not equal to 0	
95 percent confidence interval:	
50.44549	153.94531
mean of the differences :	
102.1954	
Panel: C	
data: S.1 and S.24.1	
$t_{paired} = 5.8224$; $df = 125$; p-value = 4.601e-08	
alternative hypothesis: true difference in means is not equal to 0	
95 percent confidence interval:	
95.06151	192.96658
mean of the differences:	
144.0140	

As we can see (Table 5), for all cases, the mean of the differences is positive. It implies that traditional method produces, on average, larger estimates respect to the GPSs methods. These estimates are: larger respect to the Garmin72 (the mean of the differences is +102.1954) and brandly larger respect to the Magellan400 (the mean of the differences is +144.014).

Due to both the permanence of outliers and the departure from normality appears, the use of non-parametric methods are suggested. Sample means are substituted by sample medians and the Wilcoxon sign rank test (Equation 11) on samples differences is applied (Table 6).

In Panel AW, traditional method and Garmin60 are face toward. Since the p-value is greater than the significance level ($0.7977 > 0.05$), we cannot reject the null hypothesis (Equation 4) that the true difference in medians is equal to zero, then we can conclude stating that parcel estimates using Garmin60 are not statistically different from parcel estimates using traditional method.

In Panel BW and CW since the p-value is smaller than the significance level, we can reject the null hypothesis concluding that both Garmin72 and Magellan400 parcel estimates should be considered statistically different from parcel estimates using the traditional method.

Let's inspect now, the relatively sizes of the parcels using the value and the sign of pseudo-medians of the differences. In Panel BW and CW, the pseudo-median of the differences is positive. Then, parcel estimates using traditional method are *slightly larger* respect to the Garmin72 (the pseudo-median of the differences is +91), and *larger* respect to the Magellan400 (the pseudo-median of the differences is +109.565). The pseudo-median departure pointed out in Panel AW, has not exploratory power.

Table 6: *Wilcoxon sign rank test: the non-parametric result.*

Wilcoxon Sign Rank test
Panel: AW
data: S.1 and S.21.1
$W^+ = 1886.5$; p-value = 0.796
alternative hypothesis: true location shift is not equal to 0
pseudo-median of the differences:
5.99996
Panel: BW
data: S.1 and S.22.1
$W^+ = 5428$; p-value = 0.0001076
alternative hypothesis: true location shift is not equal to 0
pseudo-median of the differences:
91
Panel: CW
data: S.1 and S.24.1
$W^+ = 6276$; p-value = 3.018e-08
alternative hypothesis: true location shift is not equal to 0
pseudo-median of the differences:
109.565

The non-parametric approach based on the Wilcoxon sign rank test, supports brandly the parametric conclusions (*Table 7*).

Table 7: *The paired t-test and the Wilcoxon sing rank test (random sampling is assumed).*

Paired t -test: the results

1) Statistical equivalency:

- * traditional method is *statistically equivalent* to the Garmin60 method.
- * traditional method is *statistically different* to the Garmin72 method.
- * traditional method is *statistically different* to the Magellan400 method.

2) Parcel estimates:

- * using traditional method are *larger* than parcel estimates using Garmin72 method.
- * using traditional method are *brandly larger* than parcel estimates using Magellan400 method.

Wilcoxon Sign Rank test: the results

1) Statistical equivalency:

- * traditional method is *statistically equivalent* to the Garmin60 method.
- * traditional method is *statistically different* to the Garmin72 method.
- * traditional method is *statistically different* to the Magellan400 method.

2) Parcel estimates:

- * using traditional method are *slightly larger* than parcel estimates using Garmin72 method.
- * using traditional method are *larger* than parcel estimates using Magellan400 method.

On the statistical equivalency hand, traditional method is found, statistically equivalent to the Garmin60 method. Instead Garmin72 and Magellan400 are discovered, statistically different from traditional method. On the parcel estimates hand, the traditional method tends to produce larger parcels estimates respect to all GPSs measurements methods.

4 Conditional inference

In this section, the data will be reanalysed using conditional test procedures, indeed, statistical tests where the distribution of the test statistic under the null hypothesis is determined *conditionally* on the data at hand.

When the samples are not generated by a random mechanism, covariates are not longer random variables and parametric statistical tables (such as t or F tables) are not valid because their are based on *theoretical distributions* which assume random sampling. Fisher(1935) was the first to understand that, classical parametric tests comparing observed statistics to theoretical distributions were inappropriate. The sample space variability is intentionally reduced and *empirical distributions* of the test statistics, indeed, distributions of the test statistics calculated directly from the statistical units surveyed, should be used instead. The resampling is the tool to depart from theoretical distributions to empirical distributions.

These concepts together with the resampling issue are briefly explained in (*Subsection 4.1*). After that, the paired permutation t -test is theoretically described in (*Subsection 4.2*) and applied in (*Subsection 4.3*). Final conclusions are gather in (*Section 5*) and further developments are suggested in (*Section 6*).

4.1 How can we make conditional inference

Up to now, the keeping of random sampling assumption has had a great influence. Samples were random variables and the test statistic was itself a random variable with a defined theoretical distribution associated with it. For example, under the null hypothesis, the t_{paired} test statistic was as a random variable distributed as a Student- t distribution with $n-1$ degree of freedom. The observed value of the test statistic was found from the samples and its critical value from the statistical tables. Finally, comparing the observed and the critical value of the test statistic we have formulated our test decision.

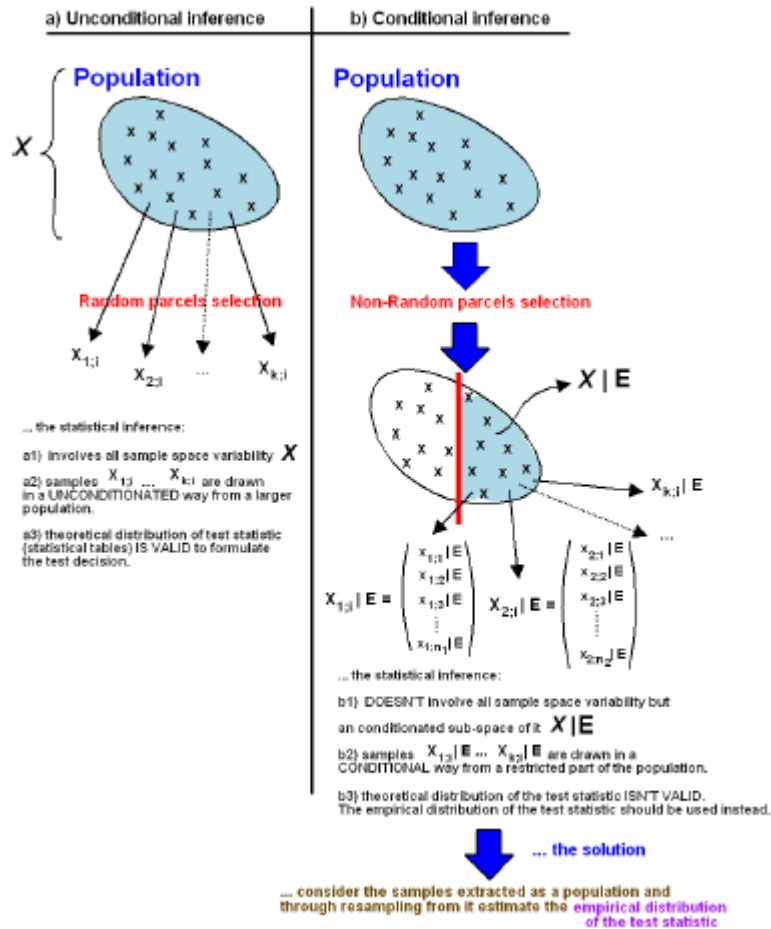
When random sampling assumption is not satisfied this procedure fails and resampling methods are necessary for statistical inference. About that, Edginton (1987) has written: *... in the absence of random sampling the statistical inferences are restricted to the subjects actually used in the experiment, and generalization to other subjects must be justified by a nonstatistical argument.* Two issues are raised. First, intentional selection implies a conditional restriction of the sample space variability. Second, population statistical remarks cannot be directly inferred from the statistical units surveyed without any artificial transformation applied of the samples extracted.

Let's see now, the resampling role for these two issues (*Figure 9*). Supposing that there is a target population whose individuals are elements of a sample space denoted by \mathcal{X} and assuming that we are interested in studying the variable of interest X . It is supposed observable on each element of \mathcal{X} and that associated with (X, \mathcal{X}) there exists a possibly unknown population distribution indicated with P_r . The notation $\Theta = (X, \mathcal{X}, P_r)$ summarizes the statistical model associated with the unconditional problem.

When intentional selection appears: first, the reference space is obtained by considering the restriction of \mathcal{X} to the sub-space associated with the conditioning event of interest⁹, which is

Figure 9: *Unconditional and conditional inference*

⁹ The event of interest \mathbf{E} represents any conscious or unconscious selection bias introduced in the experiment. For example, for GPS cultivation parcel selection, \mathbf{E} could be local government's bonds, FAO's staff instructions or natural environment impediments.



pointed out by $\mathcal{X}|E$. The statistical model becomes: $\Theta|E = (X, \mathcal{X}|E, P_r|E)$, where $P_r|E$ is the unknown conditional probability distribution of the population. Second, because we are observing realizations of $\mathcal{X}|E$ with probability law $P_r|E$ and not realization of \mathcal{X} with probability law P_r . Moreover, we are also interested in *generalizing* our inference to the entire target population and, due to lack of assumptions, we are not allowed to use unconditional methods based on all \mathcal{X} . It means that, test statistical tables are not useful and empirical distributions of the test statistic must be used instead. Resampling is the tool used for generalizing to other statistical units the conditional statistical remarks.

There are at least three major types of resampling. Each of them, was developed by different people at different periods for different purposes:

- **Randomization tests:** Was developed by Fisher (1935-1960). They are also known as *permutation tests* because the inference foresees the shuffling of the samples elements as a cards deck. In his later years, Fisher lost interest in the permutation method because there were no computers in his days to implement such time consuming and laborious calculus. The aim is to make inference estimating the empirical distribution of the test statistic, called the *permutation distribution*.

- **Jackknife:** Was invented by Quenouille (1949) and later developed by Tukey (1958). The aim is to estimate the distribution of a population by *deleting* one observation at a time. This distribution can be used, for estimating the bias and the standard error of a given estimator.
- **Bootstrap:** Was introduced by Efron (1979, 1981, 1982) and further developed by Efron & Tibshirani (1993). The aim is to estimate the distribution of a population by resampling *with replacement* for estimating the standard error and the bias of a given estimator.

Because we are interested in inference and not in estimators' reliability, we shall have to dealing with permutation tests. More specifically, we shall consider the permutation test procedure associated to the paired t -test, which is called the *permutation paired t -test*.

4.2 Permutation tests

In non-random hypothesis testing, permutation tests are often applied as a non-parametric test based on resampling, but unlike to the ordinary bootstrap sample replicates are repetitively drawn *without replacement* from the samples observed.

Permutation tests exist for any type of test statistic. Under the null hypothesis (*Equation 3*) or (*Equation 4*), estimating the permutation distribution of the test statistic, we can compute exact p-values as a proportion of test statistic replicates that are, in absolute value, at least large as the observed test statistic calculated for the original samples.

The main assumption of these tests is that the sample observations are *exchangeable* under the null hypothesis, that is, the joint distribution of the samples remains unchanged under rearrangements of their elements positions when the null hypothesis is true. This implies two consequence: first, observations viewed individually must be identically distributed, second, to compare the location of sample distributions equal variance assumption is required (Good (2000)).

The test statistic chosen, may take into account to one and or two sample as well as to paired and to un-paired data. We shall consider the permutation distribution associated to the paired t -statistic.

When pairing is not effective, to test the null hypothesis (*Equation 3*), we may use the resampling counterpart to the unpaired t -test, known as *unpaired permutation t -test*.

Suppose that, two independent samples $X_{1;n_1} = (x_{1,1}, \dots, x_{1,n_1})'$ and $X_{2;n_2} = (x_{2,1}, \dots, x_{2,n_2})'$ with $n_1 \neq n_2$ are observed from the unknown distributions F_{X_1} and F_{X_2} . To carry out the permutation test, first of all, we need to work out the unpaired t -test statistic for all sample replicates. To do that, we need to create a set capable of interchanging the values attached to each statistical unit. Cause independence, we can consider the ordered pooled sample:

$$Z = X_{1;n_1} \cup X_{2;n_2} = (x_{1,1}, \dots, x_{1,n_1}, x_{2,1}, \dots, x_{2,n_2})'$$

indexed by:

$$v = \{1;2;\dots;n_1;n_1 + 1;\dots;n_1 + n_2\} = \{1;2;\dots;N\}$$

It implies that:

$$Z_i = \begin{cases} X_{1,i} & ; \text{ if } 1 \leq i \leq n_1 \\ X_{2,i-n_1} & ; \text{ if } n_1 + 1 \leq i \leq n_1 + n_2 \end{cases} \quad (12)$$

indeed, Z sample gathers sample elements in a defined order. Now, any partition of Z , let's say, $Z^* = \{X_1^*; X_2^*\}$ corresponds to a permutation γ of the integers v , where $Z_i^* = Z_{\gamma(i)}$.

In this unbalanced situation, we can generate:

$$P_{N;n_1;n_2} = \binom{n_1 + n_2}{n_1} = \frac{(n_1 + n_2)!}{n_1!n_2!} = \frac{N!}{n_1!n_2!}$$

permutations. According to the Permutation Lemma, Efron(1993 pag.84), is reasonable assuming that all permutations are equally likely. It implies that, under the null hypothesis (*Equation 3*) a partition of Z has probability:

$$P_r(Z^* = z^* | H_0) = \left[\frac{N!}{n_1!n_2!} \right]^{-1}$$

of having been drawn. Where z^* is a particular realization of the Z^* partition. If $\hat{\theta}$ is our test statistic calculated for the original samples, it can be viewed as a function of them, indeed, $\hat{\theta}(X_{1;n_1}; X_{2;n_2})$. Using the index vector v we can rewrite it as a function of the ordered sample and the index vector, that is, $\hat{\theta}(X_{1;n_1}; X_{2;n_2}) = \hat{\theta}(Z; v)$. It follows that, the *permutation distribution* of our test statistic is the distribution of the test statistic replicates:

$$\begin{aligned} \{\hat{\theta}^*\} &= \left\{ \hat{\theta}(Z; \gamma_j(v)) \quad ; \quad j = 1, 2, \dots, \frac{N!}{n_1!n_2!} \right\} \\ &= \{ \hat{\theta}^{(j)} | \gamma_j(v) \} \end{aligned} \quad (13)$$

Where $\hat{\theta}^{(j)}$ is the j -th replicates of the test statistic and $\{\hat{\theta}^*\}$ a set of test statistic values which can be displayed using a histogram plot.

Thus, the cumulative distribution function of the permutation distribution is an equally weighed function which cumulates the values of the j -th test statistic replicates:

$$F_{\hat{\theta}^*}(t) = P_r(\hat{\theta}^* \leq t | H_0) = \left[\frac{N!}{n_1!n_2!} \right]^{-1} \sum_{j=1}^N 1\{\hat{\theta}^{(j)} \leq t\} \quad (14)$$

Besides, the achieved significance level or p-value of our test statistic $\hat{\theta}$ is the probability that the observed statistic is not greater than the test statistic replicates:

$$P_r(t) = P_r(\hat{\theta}^* \geq \hat{\theta} | H_0) = \left[\frac{N!}{n_1!n_2!} \right]^{-1} \sum_{j=1}^N 1\{\hat{\theta}^{(j)} \geq \hat{\theta}\} \quad (15)$$

where, the observed test statistic $\hat{\theta}$ is a function of Z and v , indeed, $\hat{\theta} = \hat{\theta}(Z;v)$. The achieved significance level for the lower tail or for the two tails test based on the observed statistic $\hat{\theta}$, can be calculated in similar way. Only when $P_r(\hat{\theta}^* \geq \hat{\theta} | H_0) \leq \alpha$, we shall be boosted to reject the null hypothesis, stating that, parcel estimates are conditionally statistically different.

When pairing is effective, the permutation procedure does not change too much. The main effect is the reduction of the total number of the permutations.

In fact, being the test statistic calculated for the number of pairs n , we have only n paired differences to interchange and not longer $n_1 + n_2 = N$. Because each paired differences can be interchanged two times, taking the positive or the negative sign of it, we can generate 2^n permutations. Then, a paired two sample experiment with n pairs has only 2^n possible permutations:

$$\underbrace{P_{n;2} = 2^n}_{\text{total number of paired permutations}} < \underbrace{\frac{N!}{n_1!n_2!} = P_{N;n_1,n_2}}_{\text{total number of unpaired permutations}} \quad (16)$$

it implies that the inverse of the total number of permutations, indeed, the probability of each permutation is greater for the paired case:

$$P_r(Z_p^* = z_p^* | H_0) = 2^{-n} > \left[\frac{N!}{n_1!n_2!} \right]^{-1} = P_r(Z^* = z^* | H_0) \quad (17)$$

These changes affect partially the permutation procedure. The pooled sample is worked out as:

$$Z = X_{1;n} \cup X_{2;n} = (x_{1;1}, \dots, x_{1;n}; x_{2;1}, \dots, x_{2;n})'$$

indexed by:

$$v = \{1; 2; \dots; n; n+1; \dots; 2n\}$$

and structured as:

$$Z_i = \begin{cases} X_{1;i} & ; \text{ if } 1 \leq i \leq n \\ X_{2;i-n} & ; \text{ if } n+1 \leq i \leq 2n \end{cases} \quad (18)$$

In this situation, only 2^n partitions may be formulated from the pooled ordered sample Z , then, if $\hat{\theta}_p(Z;v)$ is our paired test statistic, the *paired permutation distribution* of $\hat{\theta}_p^*$ is the distribution of the paired replicates:

$$\begin{aligned} \{\hat{\theta}_p^*\} &= \{\hat{\theta}(Z; \gamma_j(v)) \quad ; \quad j = 1, 2, 3, \dots, 2^n\} \\ &= \{\hat{\theta}_p^{(j)} \mid \gamma_j(v)\} \end{aligned} \quad (19)$$

Then, the cumulative distribution function of the paired permutation distribution:

$$F_{\hat{\theta}_p^*}(t) = P_r(\hat{\theta}_p^* \leq t \mid H_0) = 2^{-n} \sum_{j=1}^{2^n} 1\{\hat{\theta}_p^{(j)} \leq t\} \quad (20)$$

As was happening before, the achieved significance level or p-value of our paired test statistic $\hat{\theta}_p^*$ is the probability that the observed statistic is not greater than the test statistic replicates:

$$P_r(t) = P_r(\hat{\theta}_p^* \geq \hat{\theta}_p \mid H_0) = 2^{-n} \sum_{j=1}^{2^n} 1\{\hat{\theta}_p^{(j)} \geq \hat{\theta}_p\} \quad (21)$$

The achieved significance level for the lower tail or for the two tails test based on the observed statistic $\hat{\theta}_p^*$, can be found in similar way. Only when $P_r(\hat{\theta}_p^* \geq \hat{\theta}_p \mid H_0) \leq \alpha$, we will reject the null hypothesis, stating that, parcel estimates are conditionally statistical different.

In practice, unless that the sample size is very small, is not feasible work out the test statistic for all possible unpaired or paired permutations. An approximate permutation test can be effectively implemented by randomly drawing a large number of samples without replacement but not all possible permutations. At least 99 and at most 999 random permutations should be sufficient (Davison & Hinkley (1997)). We shall consider 999 random replications.

Finally notice that, cause resampling, both the permutation distributions and the empirical p-value associated are random variables, since their values change each time that the simulation is implemented.

4.3 The conditional inference analysis: the permutation distribution of paired t -statistic

This section contains the practical implementation of the paired permutation test described previously. Because, replications of the paired t -statistic will be used, we shall compare unconditional and conditional inference for means.

The logical sequence that explains the paired permutation t -test is summarized by (Figure 10), while its result is displayed in (Figure 11).

As we can see (*Figure 10*), the test goes through four main steps. First, the ordered sample Z capable of gathering sample elements in a specific order must be created. This set of values could be a pooled ordered matrix whether the samples were picked up in a more than two dimensions.

Second, we need to work out the test statistic for the original samples. We have already done that in (*Table 5*). These values (indicated by t_{paired}) will be used as p-values' cutoff points.

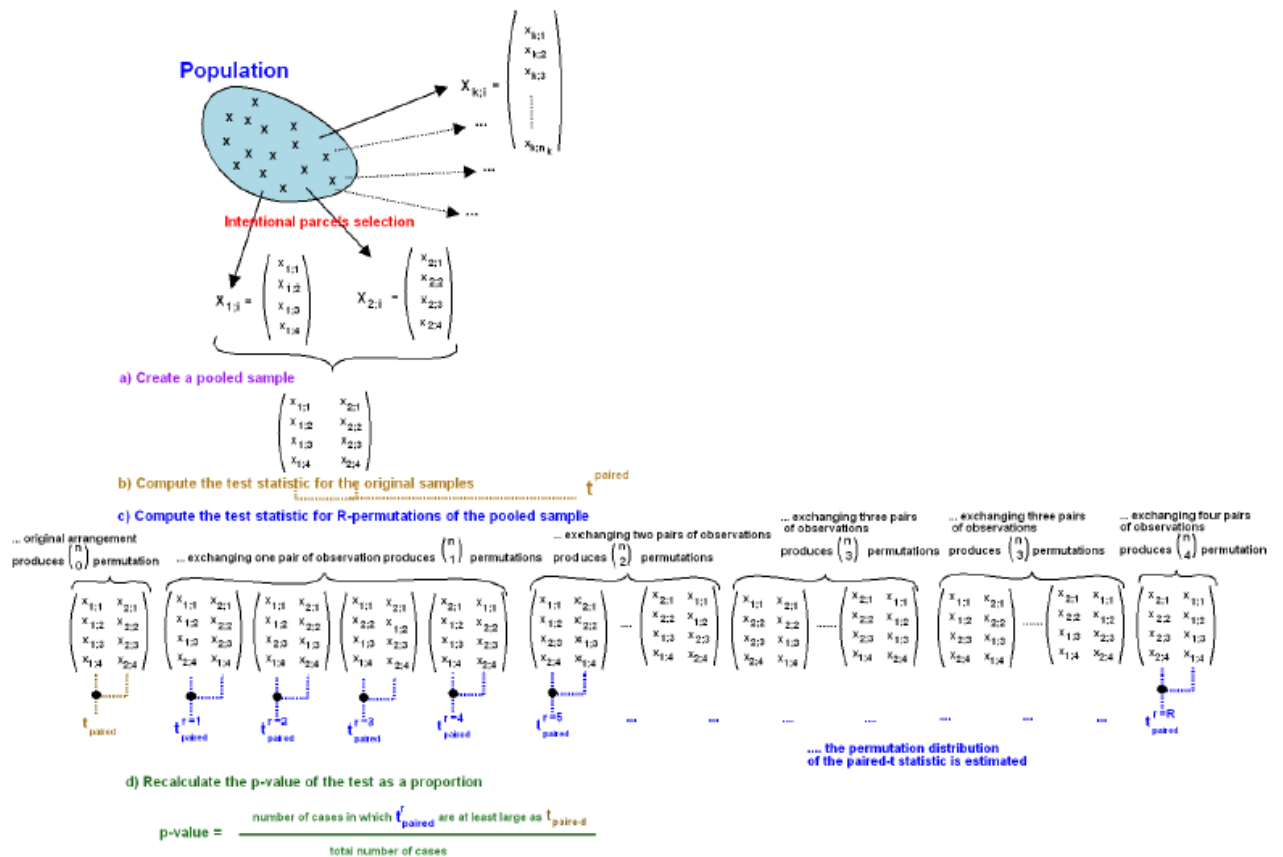
Third, is necessary estimating the test statistic for the sample replicates. When h pairs are interchanged, $\binom{n}{h}$ permutations are generated. Obviously, if any pair or all pairs are interchanged only one permutation is created, otherwise more than one permutation is produced. Summing up all of these combination terms, we can obtain the total number of paired permutations:

$$\sum_{h=0}^n \binom{n}{h} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n-1} + \binom{n}{n} = 2^n \quad (22)$$

where n is the number of pairs. The paired t -statistic is worked out only a large number of these total permutations. The paired t -test may have a different value for each permutation. The distribution of these test statistic replicates is the permutation distribution of the paired t -test, it represents the empirical counterpart of the paired Student- t distribution.

Fourth, the p-value of the test is recalculated as a proportion of the cases in which the paired t -statistic replicates are, in absolute value, at least large as the observed paired t -statistic.

Figure 10: *The logical sequence of the paired permutation t-test.*



This sequence of steps produces the permutation distributions displayed in (Figure 11), from these, the empirical p-values are gathered in (Table 8).

These histogram estimates (Figure 11) are the empirical distribution functions of paired t -test statistic worked out *conditionally* to the data at hand. Under the null hypothesis (Equation 4), the empirical p-value of the test will be given by the area of the histogram estimate outside the absolute value interval of the observed statistic (this area is emphasised in gold brown).

In Panel AR, the empirical p-value is worked out for the differences between traditional method and Garmin60 method. Since it is greater than the significance level ($0.472 > 0.05$) we cannot reject the null hypothesis (Equation 4) and then we can conclude stating that the parcel estimate using Garmin60 are not conditionally statistical different from parcel estimates using traditional method.

In Panel BR and CR, the same procedure is applied for the differences between Garmin72 method and Magellan400 method. Here, the empirical p-value is smaller than the significance level ($0.001 < 0.05$), then we can reject the null hypothesis, stating that parcel estimate using Garmin72 or Magellan400 are conditionally statistical different from parcel estimates using traditional method.

The unconditional and conditional inference on means tend to move in the same direction (Table 9).

The empirical p-values bear us to the same final conclusion of theoretical ones: only the Garmin60 cultivation parcel surfaces *are statistically equivalent* to the cultivation parcel surfaces measured using the traditional method. For the largeness of the

cultivation parcels nothing can be state using the conditional approach, then the unconditional results remain valid.

Figure 11: *The approximated permutation distributions of paired t-test.*

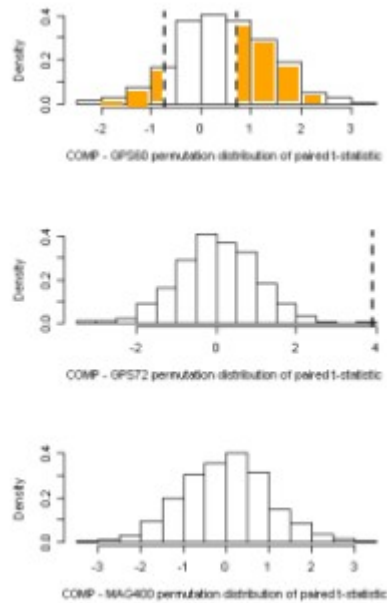


Table 8: *Permutation paired t-test: the conditional inference result.*

Permutation paired <i>t</i> -test
Panel: AR
data: S.1 and S.21.1 Observed test statistic: $\hat{\theta} = 0.7262$; Empirical p-value = 0.472 alternative hypothesis; $H_A : F \neq 0$
Panel: BR
data: S.1 and S.22.1 Observed test statistic: $\hat{\theta} = 3.9084$; Empirical p-value = 0.001 alternative hypothesis; $H_A : F \neq 0$
Panel: CR
data: S.1 and S.24.1 Observed test statistic $\hat{\theta} = 5.8224$; Empirical p-value = 0.001 alternative hypothesis; $H_A : F \neq 0$

Paired t -test	Permutation paired t -test
Panel: A	
data: S.1 and S.21.1	
$t_{paired} = 0.7262$; $df = 87$; p-value = 0.4697	$\hat{\theta} = 0.7262$; Empirical p-value = 0.472
$H_A : \mu \neq 0$	$H_A : F \neq 0$
95 percent confidence interval:	
-30.61026 65.85185	
mean of the differences:	
17.62080	
Panel: B	
data: S.1 and S.22.1	
$t_{paired} = 3.9084$; $df = 125$; p-value = 0.0001514	$\hat{\theta} = 3.9084$; Empirical p-value: 0.001
$H_A : \mu \neq 0$	$H_A : F \neq 0$
95 percent confidence interval:	
50.44549 153.94531	
mean of the differences :	
102.1954	
Panel: C	
data: S.1 and S.24.1	
$t_{paired} = 5.8224$; $df = 125$; p-value = 4.601e-08	$\hat{\theta} = 5.8224$; Empirical p-value: 0.001
$H_A : \mu \neq 0$	$H_A : F \neq 0$
95 percent confidence interval:	
95.06151 192.96658	
mean of the differences:	
144.0140	

5 Conclusions

The unconditional and conditional statistical inference done on the available survey dataset for Cameroon, Niger and Senegal, allows drawing some concluding remarks:

- Statistical inference is *strictly necessary* to assess the relevance of GPSs measurements. It cannot have nothing to do with the sampling. Unconditional or simple statistical inference can be formulated only when parcel selection is supposed random. Both parametric and non-parametric approaches are possible. Instead, when parcel selection is assumed non-random, resampling methods for parameter estimation and inference should be preferred;
- The measurement of cultivation parcels using GPS may be *significant* to reduce the costs of agricultural surveys. The significance works when methods measurements are found statistically equivalent. It implies that, the more expensive traditional method may be substituted by the cheaper equivalent GPS method;
- On the statistically equivalence hand, using both unconditional and conditional inference procedures, traditional method is found *statistically equivalent* to the

Garmin60 method. Garmin72 and Magellan400 methods are discovered statistically different instead. The loss of accuracy when we accept the null hypothesis using the unconditional approach is on the order of 2/1000;

- On the parcel estimates hand, only the unconditional approach can be used. Both the parametric and non-parametric approaches applied suggest us that the traditional method tends to produce larger parcel estimates than GPSs methods.

In conclusion, because GPSs methods are globally cheaper than traditional method, is strongly recommended the use of Garmin60 to reduce the costs of agricultural surveys.

6 Future research

As basic hint for the upcoming papers, we suggest to take care about the random mechanism which has generated the samples. When the selection mechanism does not guarantee randomness, we recommend taking apart the classical statistical tests and move on to the permutation tests. Instead, if the selection mechanism guarantees randomness both classical tests and permutation tests may be applied.

Taking in mind that, the arguments presented in this paper can be developed in two broad directions: changing the statistical aim or changing the permutation procedure.

On the one hand, changing the statistical aim keeping the permutation procedure unchanged represents the immediate extension of this work. We might be interested in:

- Study the **statistical equivalency** taking into account of **smaller and bigger parcels separately**. It may allow us, first, to verify how the size of the parcels affect the statistical equivalency, second, to find out the *surface threshold* apart which the equivalency is not satisfied;
- Study the **statistical equivalency** for the **time requested** to do the measurements. It may permit us, first, to verify how the time affect the statistical equivalency, second, to discover the *time threshold* apart which the equivalency is not satisfied.

On the other hand, we may change just the permutation procedure. Then to study the statistical relevance of different measurements we may consider:

- Permutation tests which take care of *any differences* of the two samples, not only in means. The permutation distribution of the **Kolmogorov-Smirnov test statistic** represents the optimal tool to measure this global departure;
- **Multivariate permutation tests** where more than two samples are jointly faced toward. These tests represent an open field for the research. The main type of multivariate permutation test is based on *nearest neighbors*. The nearest neighbour is an algorithm used in data mining, statistical pattern recognition, image processing to classify variables based on minimum distance from the query instance.

References

- Agresti A. (2002): *Categorical Data Analysis*, John Wiley & Sons, 2nd edition.
- Davison A. C. & Hinkley D.V. (1993): *Bootstrap Methods and their Application*, Cambridge University Press, Oxford, 1997.
- Edgington E. S. (1987): *Randomization Tests*, Ney York, Macel Dekker.
- Efron B. (1979): Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, 7, 1-26.
- Efron B. (1981): Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, *Biometrika*, 63, 589-599.
- Efron B. (1982): The jackknife, the bootstrap, and other resampling plans, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.
- Efron B.& Tibshirani R.J. (1993): *An introduction to the bootstrap*, New York: Chapman & Hall.
- Fisher R. A. (1936): *The Design of Experiments*, Oliver and Boyd, Edinburgh.
- Good P. (2000): *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, New York, Springer-Verlag.
- Green W. H. (2008): *Econometric Analysis*, Prentice Hall, 6th edition.
- Gosset W. S. (1908): The probable error of the mean, *Biometrika*, 6, 1-25.
- Kennedy P.E. (1995): Randomization tests in Econometrics, *Journal of Business and Business*, 7, 1-14.
- Palmegiani, G., Pizzoli E.: (2007b): Rural Development Statistics (RDS) for Policy Monitoring: a Rural-Urban Territorial Classification and Farmers' Income Data, AFCAS, Alger. www.fao.org/statistics/meetings/afcas2007
- Peracchi, F. (2000): *Econometrics*, Wiley.
- Pesarin, F. (2001): *Multivariate Permutation Tests: with Applications to Biostatistics*, Wiley.
- Quenouille M. (1949): Approximate tests of correlation in time series, *Journal of the Royal Statistical Society Series B*, 11, 18-84.
- Tukey J.W. (1958): Bias and confidence in not quite large samples, *Annals of Mathematical Statistics*, 11, 18-84.
- UNECE, FAO, OECD and World Bank (2005) *Rural Household's Livelihood and Well Being: Statistics on Rural Development and Agriculture Household Income*, Handbook, United Nations, www.fao.org/statistics/rural.
- Wooldridge J. M. (2001): *Econometric Analysis of Cross Section and Panel Data*, MIT press.
- Wilcoxon F. (1945): Individual comparisons by ranking methods, *Biometrics*, 1, 80-83.