THE INTERNATIONAL SOCIETY OF PRECISION AGRICULTURE PRESENTS THE
13th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE
July 31-August 4, 2016 • St. Louis, Missouri USA

Processing yield data from two or more combines

Maldaner, L. F.; Molin, J. P.; Canata, T. F.

Av. Pádua Dias, 11, 13418-900. University of São Paulo - "Luiz de Queiroz" College of Agriculture. Department of Biosystems Engineering. Piracicaba, SP, Brazil

A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA

Abstract. *Erroneous data affect the quality of yield map. Data from combines working close to each other may differ widely if one of the monitors is not properly calibrated and this difference has to be adjusted before generating the map. The objective of this work was to develop a method to correct the yield data when running two or more combines in which at least one has the monitor not properly calibrated. The passes of each combine were initially identified and three methods to correct yield data were tested: a) Machine by machine - (1) select the combine with more data (larger harvested area in the field); (2) compute the average yield value from each combine; (3) a correction factor is generated at each point with the ratio between the average of the combine and the nearest combine; (4) yield data from the nearest combine are multiplied by the correction factor; (5) if more than two combines are involved, identifies the nearest combine and repeat step (4) and (5) to all combines. b) Track by track - this method is similar to the previous, however the average yield values are extracted only at points within the pass closest to the combine 1 and the pass of the nearest combine. c) Point by point - a correction factor is generated through the medium of yield ratio of the closest points between the combine with the largest harvested area and the nearest combine; yield data will be multiplied by this factor and these steps will be repeated for all combines. It is very important to have the total production and field area for control and comparison. The closer are the two values, the greater will be the efficiency of post processing of data. The three methods were evaluated by using raw data from corn, cotton and wheat harvesting and were able to correct the data with distinct characteristics. The model should be selected according to each area and with user needs. Therefore, there is no standard method to be used for the correction of yield data.*

Keywords. *erroneous data, correction data, yield mapping*

Introduction

Precision agriculture techniques seek to aggregate information necessary to monitor the variability in the crop, as well as to support decision making for possible interventions. The yield map is a generalized concept, and possibly the most crucial information for a successful precision agriculture system. Maps generated with georeferenced data identify and alert for the locations where interventions must be directed.

However there are various errors in yields maps related to the characteristics of the harvester, to the yield measuring system, the variations within the field, the operator and to the procedure for obtaining the map. Blackmore and Moore (1999) mentions the presence of errors in yield maps well as measuring the moisture and yield sensor errors, filling time error of the harvester in the headwaters, GNSS positioning errors, driver errors and file write error.

The presence of errors in the data used in the preparation of yield maps is very detrimental to the quality of the generated map, and can even lead to erroneous interpretations even when the volume of errors in relation to the total of collected data is small. Thus, it is extremely important that these errors are removed before any analysis of the maps.

Various methods have been studied by several authors (MOLIN and GIMENEZ, 2000; MENEGATTI and MOLIN, 2003, 2004; PING and DOBERMANN, 2005; SIMBAHAN et al., 2004; ARSLAN and COLVIN 2002; BLACKMORE and MOORE, 1999; SPEKKEN et al., 2013) to remove these erroneous data in a data set generated by yield monitors.

However, there are other types of errors found in yield maps. One important type which was not yet reported is the error caused by generation of data set by various yield monitors in the same field with different calibrations. This is a common harvest scenario in large grain fields. The yield data from combines which are working close to each other can differ widely if one of the monitors is not properly calibrated. This difference must be treated before the map generation.

The objective of this work was to develop a method to correct the yield data when these came from different combines working in the same field and at least one of the yield monitors were not properly calibrated.

Methods

From a spatial data set consisting of points a model is proposed for the normalization of data where there is at least one file generated by a monitor not calibrated correctly. The file must contain three attributes: the latitude and longitude data, in either geographical coordinates (datum wgs84) or UTM (Universal Transverse Mercator) formats, the yield data and the yield monitor identification.

The geographic coordinates are converted into coordinates UTM, allowing the calculation of distances between the points. The first step is to recognize each path of the combine. A modification over Menegatti and Molin (2004) was applied in order to identify the ending points from each path.

After the identification of path, the model identifies the monitor with more data (larger area harvested on the field) (C_i). A correction factor (F_c) is generated by the relationship between the data values C_j and the nearest combine (C_j). In this work, three methods will be tested to generate the F_c here named: *Machine by Machine*, *Track by track* and *Point by point*. The correction of the data C_j is done through the equation 1:

$$C_{jcor} = C_j \times F_c \tag{1}$$

where F_c is the correction factor, C_j is the value of the yield of each point of the nearest monitor C_i to be corrected and C_{jcor} is the value of the corrected yield.

In *Machine by Machine* method, the correction factor is generated through the relationship between the mean value of the combine C_i yield data and the mean values of the points C_j combined (Figure 1B). The *Track by Track* method identifies the two closest paths between the C_i combine and C_j combine (Figure 1C). The correction factor is generated through equation 2:

$$\mathrm{F_c} = \frac{\mathrm{Med_i}}{\mathrm{Med_j}} \tag{2}$$

where F_c is the calculated correction factor and Med_i and Med_j values are given by:

$$Med_i = \frac{\sum_{i=1}^{n} V_P(T_k C_i)}{n} \tag{3}$$

$$Med_j = \frac{\sum_{i=1}^{n} V_q(T_g C_j)}{n} \tag{4}$$

where *n* is the total number of points within the path, V_P is the values of yield of each point within the *k* path of the *i* combine and V_q is the value of yield of each point within the *g* path of *j* combine.

The *Point by Point* method identifies the two closest path between the combines C_i and C_j and then is calculated the distances between the points of the path of the combine C_i and points of the path of the combine C_j. Identifies the pairs of points between the two paths through the shortest distance calculated. Points near path between the two forming a pair (Figure 1D). The correction factor is generated by the sum of the relationships between pairs of points:

$$F_c = \frac{\sum_{i=1}^{z} \frac{Pt_i(T_k C_i)}{Pt_j(T_g C_j)}}{z} \tag{5}$$

where F_c is the calculated correction factor, Pt_i is the point of yield value within the path the combine C_i, Pt_j is the point of yield value within the path the combine C_j and *Z* is the number of pairs of points between the two paths.

If there are more than two combines near, the process of identifying the nearest combine and the calculation of the correction factor is repeated for all combines. When more than two combines in field, only one of which is near to the combine C_i, the values of the nearest combine and then identifies which are first corrected combine is closest to the corrected data set has already generated and then if the correction factors between them. This process is repeated until all combines are fixed with yield value.

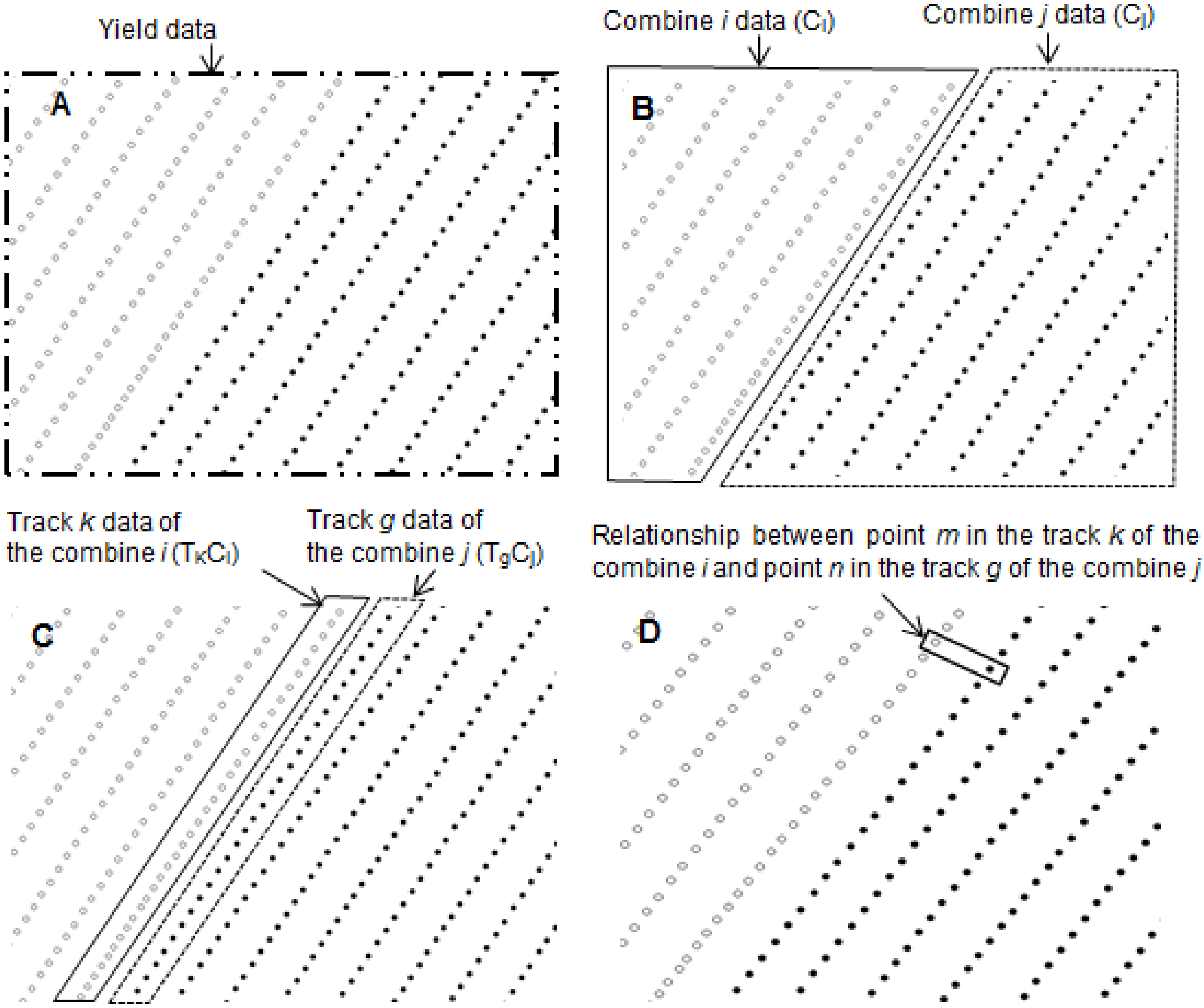


Fig 1. Identification of yield data (A) used to generate the correction factor for the three proposed methods: *Machine by Machine* (B), *Track by Track* (C) and *Point by Point* (D).

Case study

For the assessment of the three methods we used yield data for three different areas with three annual crops (corn, cotton, wheat). Each area contains yield data collected from two yield monitors. The amount of data collected by each combine for each area is presented in Table 1.

Table 1. Number of data collected by each combine.

		Number of data		
Variable	Crop area (ha)	Combine C_i	Combine C_j	Total
Corn	172.10	41043	37400	78473
Cotton	152.88	45482	32961	84297
Wheat	72.41	26463	15873	42336

A high variation in productivity of the field can be seen more clearly in Figure 2, where yield values for the three plots are shown. It can be seen that there are variations between yield values in small distances, which shows inconsistency with the expected spatial dependence of yield. This indicates the occurrence of one or more yield monitors within the plot not be properly calibrated. Table 1 shows the variation of yield values for the areas under study.

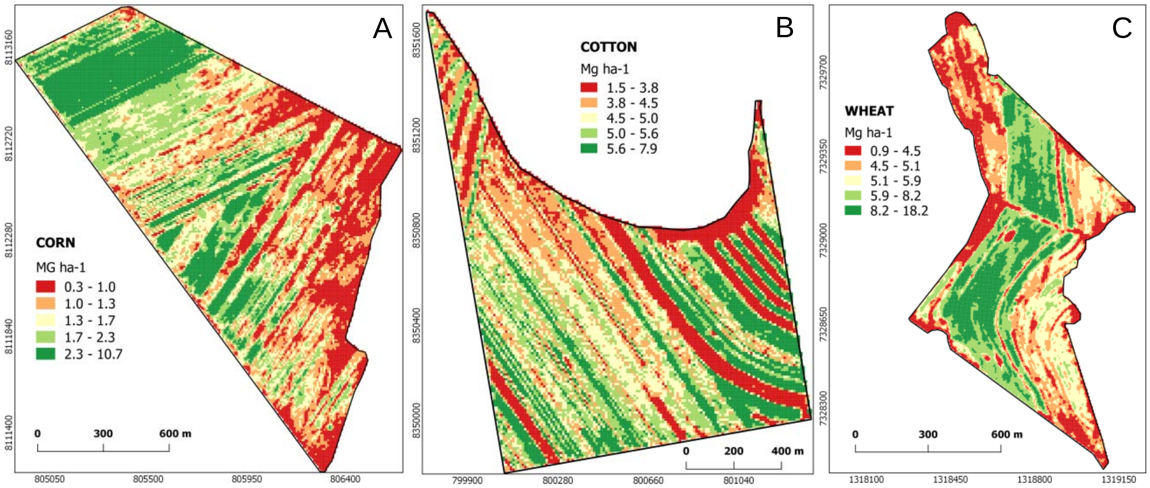


Fig 2. Yield data generated by different combines for the three different areas: maize (A), cotton (B) and wheat (C).

The values of skewness and kurtosis suggest that the distribution is not normal, which indicates that the average yield being influenced by extreme yield value. The median values deviate from the values of arithmetic mean and also the minimum and maximum values reinforce the observation of high variability, with high values of the coefficient of variation (CV). High values of CV may be considered as the first indicator of the existence of heterogeneity in the data (RIBEIRO JUNIOR, 1995; GOOVAERTS, 1999 e FROGBROOK et al., 2002).

Table 2. Descriptive statistics of the raw dates.

Variable	Count	Mean	StDev	Minimum	Median	Maximum	Skewness	Kurtosis	Variance	CV (%)
		Mg ha^{-1}								
CORN	78443	1.87	1.03	0.34	1.59	6.67	1.39	1.53	1.07	55.19
COTTON	84297	4.81	1.55	0.68	5.07	9.48	0.04	-1.52	2.39	32.10
WHEAT	42336	6.09	2.00	0.36	5.38	23.98	0.67	0.24	4.04	32.97

CV – coefficient of variation.

To correct yield data, the first step was the separation of the path by identifying the ends of the path of the combine. Then it identified the combine with more yield data within the field (C_i) and the nearest combine (C_j).

Correcting the yield data by applying the correction factors generated by the three proposed methods was performed. It was noticed that after correction of yield value the mean value of yield decreased for areas of corn and wheat. In the area with cotton, the application of the correction factor over the yield data increased the mean value yield (Table 3).

For the three sets of data analyzed (cotton, corn, wheat), the applied methods reduced the standard deviation, meaning that there was a reduction in the amplitude of the data. The application of the correction factor generated by the three methods reduced the variation in the data, showing the effectiveness of the correction.

Table 3. Descriptive statistics of the yield data after application of the correction.

Method of correction	Count	Mean	StDev	Minimum	Median	Maximum	Skewness	Kurtosis	Variance	CV (%)
		Mg ha^{-1}								
CORN										
Combine by combine	78443	1.42	0.61	0.20	1.32	3.99	0.64	-0.11	0.38	43.33
Track by track	78443	1.22	0.54	0.14	1.16	3.49	0.34	-0.68	0.29	44.29
Point by point	78443	1.28	0.55	0.16	1.21	3.49	0.37	-0.68	0.30	43.06
COTTON										
Combine by combine	84297	6.04	0.76	0.68	6.10	9.48	-0.93	2.70	0.58	12.57
Track by track	84297	6.16	0.77	0.68	6.23	9.48	-0.95	2.81	0.59	12.50
Point by point	84297	6.29	0.81	0.68	6.36	9.87	-0.82	2.49	0.65	12.86
WHEAT										
Combine by combine	42336	4.78	0.79	0.36	4.83	13.82	-0.37	3.76	0.64	16.71
Track by track	42336	4.85	0.82	0.36	4.91	14.40	-0.33	4.04	0.67	16.85
Point by point	42336	4.69	0.79	0.36	4.75	13.21	-0.35	3.24	0.63	16.84

CV – coefficient of variation.

Comparing the three correction methods, there was no significant difference between the mean values of yield in both sets of data analyzed, nor in the values of the coefficient of variation of the data, showing that the methods of correction were similar for the studied areas.

The spatial dependence analysis to the data was carried out before and after the three methods of correction, with the use of the software VESPER 1.6. In the selection of models of variogram were considered the smallest RMSE values (root mean square error) adjustment of theoretical models to experimental variograms.

For the three sets of analyzed data, the application of the correction factor (for the three methods) reduced the value of nugget effect (Table 4). According Cambardella et al., (1994), the nugget effect is a variogram parameter that indicates the unexplained variability in the model, considering the sampling distance used. The data correction process contributed to the characterization of the spatial dependence, reducing the variability not explained by models adjusted to the raw data for the areas under study.

The spatial dependence exists when there is increased semi variance within a certain range, called "range" (a), which is the range within which the samples are spatially correlated. After reaching a stabilization of semi variance occurs in a value called "still" (C0 + C1). After this value is no more spatial dependence between samples (GOMES et al., 2009). There was reduction in the range values after application of the correction factors in the corn yield data. For cotton and wheat yield data there was a greater spatial continuity of data after application of correction factors, that is, greater range of value in the variogram.

The assessment of the spatial dependence is performed through interpretation of some indicator of spatial dependence. In this work it was used the spatial dependence ratio given by Cambardella et al., (1994). The original yield data of corn crop presented high spatial dependence (IDE = 2.32 %), and even after application of the correction factor the data continued presenting high index of spatial dependence.

Table 4. Model and parameters of variograms generated for the three areas studied

DATA	MODEL	RMSE	C0	C1	A1	IDE
			CORN			
Original	Gaussian	0.082	0.81	34.19	10000.00	2.32
Combine by combine	Gaussian	0.015	0.18	1.47	2186.90	11.16
Track by track	Gaussian	0.008	0.11	0.75	1499.50	12.76
Point by point	Gaussian	0.008	0.12	0.87	1638.60	11.95
			COTTON			
Original	Exponential	0.170	1.90	0.47	86.03	80.32
Combine by combine	Spherical	0.019	0.32	12.02	50000.00	2.56
Track by track	Gaussian	0.035	0.39	0.88	1491.10	30.85
Point by point	Gaussian	0.030	0.45	2.06	2386.60	17.79
			WHEAT			
Original	Spherical	0.865	0.94	3.13	248.00	23.14
Combine by combine	Spherical	0.026	0.55	0.46	4909.10	54.70
Track by track	Spherical	0.025	0.58	0.20	1898.60	74.20
Point by point	Spherical	0.030	0.54	0.93	10000.00	36.99

RMSE – root mean square error, C0 - nugget effect, C1- Still, A1- range, IDE – Index of spatial dependence (%).

The original yield data of cotton presented low spatial dependence, but after applying the correction factors generated by the methods *Machine by Machine* and *Point by point* the data presented a high spatial dependence. The original yield data of wheat had high spatial dependence, however after application of the correction factors to the data presented moderate spatial dependence.

The application of the original data correction factors assisted in the visualization of real spatial variability of low and high yield value in both areas detected with the raw data, but for the cotton yield data the application of correction factors not completely removed the variations in productivity between the two combines in the field.

This occurred because there were areas within the field where the yield values from the combine C_i were higher than the yield values from the combine C_j and areas within the field in which they combine C_i yields were lower than the values from the harvester C_j, which indicates the need for different correction factors within the same field.

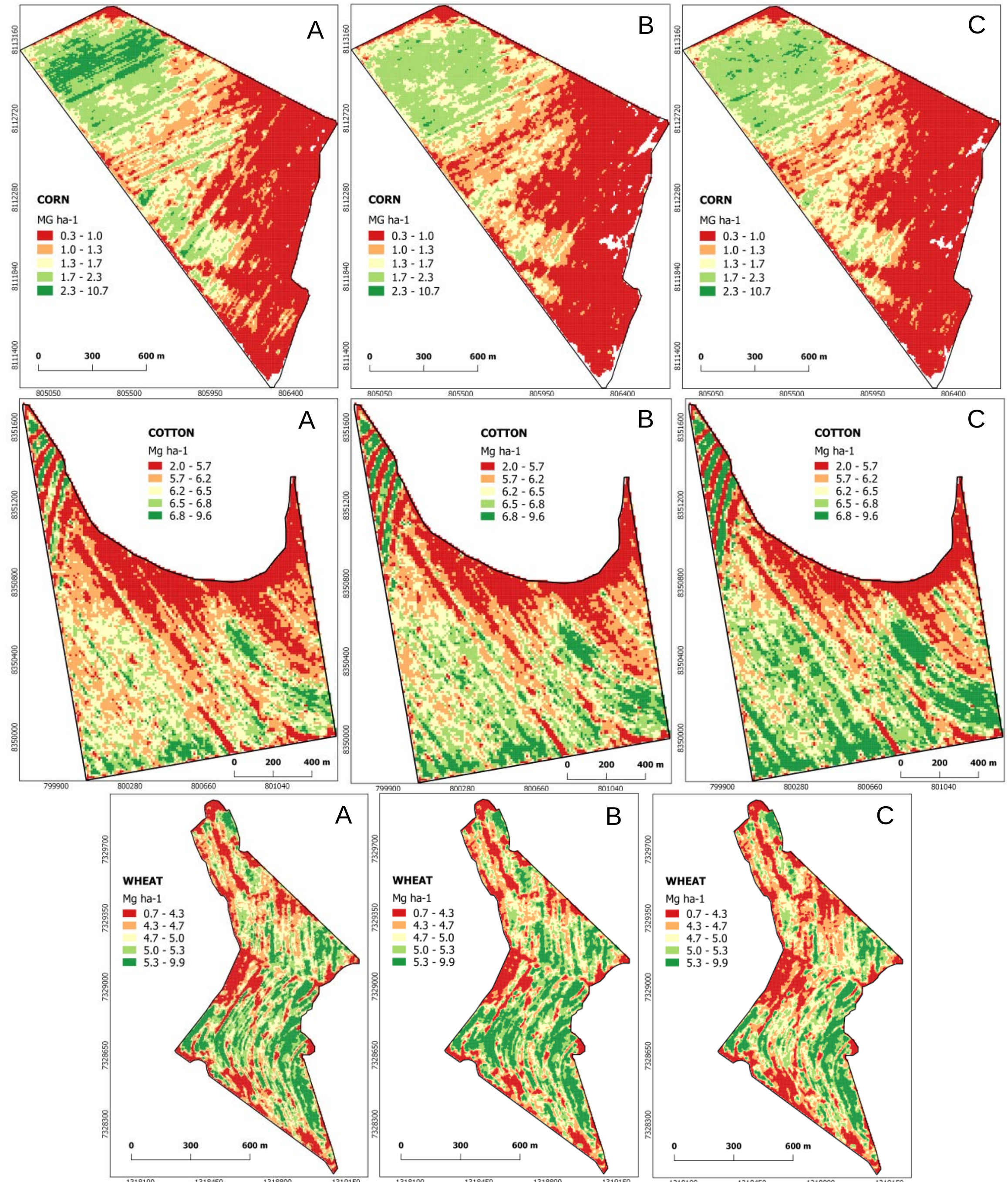


Fig 3. Yield data generated after the application of the three correction methods (*Machine by Machine* (A), *Track by Track* (B) and *Point by Point* (C).

Conclusion

The proposed methods of yield data correction were efficient in eliminating differences in yield when the relationship of the variation in the yield between two combine are equal across the field. The methods proposed were not effective when there is the relationship of the variation of yield between two combines is different, and there is the need of elaborating new methods for this case.

Acknowledgements

For financial support provided by CAPES (Higher Education Personnel Improvement Coordination).

References

Arslan, S., Colvin, T. S. (2002). Grain yield mapping: yield sensing, yield reconstruction, and errors. Precision Agriculture 3, 135–154.

Blackmore, B. S., Moore, M. (1999). Remedial correction of yield map data. Precision Agriculture 1, 53–66.

Cambardella, C.A.; Moorman, T.B.; Novak, J.M.; Parkin, T.B.; Karlen, D.L.; Turvo, R.F.; Konopa, A.E. (1994). Field-scale variability of soil properties in central Iowa soils. Soil Science Society of American Journal, v.58, n.5, p.1501-1511.

Jin, J.; Tang, L. (2010). Optimal coverage path planning for arable farming on 2D surfaces. Transactions of the ASABE, v. 53, n. 1, p. 283.

Frogbrook, Z. L. Oliver, M. A., Salahi, M., & Ellis, R. H.(2002). Exploring the spatial relations between cereal yield and soil chemical properties and the implications for sampling. Soil Use and Management, v. 18, n. 1, p. 1-9.

Gomes, E. G., Grego, C. R., Mello, J. C. C. B. S., Valladares, G.S., Mangabeira, J. A. de C., Miranda, E. E. (2009). Dependência espacial da eficiência do uso da terra em assentamento rural na Amazônia. Production, 19(2), 417-432.

Goovaerts, P.(1999). Geostatistics in soil science: state-of-the-art and perspective. Geoderma, v.89, p.1-45.

Menegatti, L. A. A.; Molin, J. P. (2003) Metodologia para identificação e caracterização de erros em mapas de produtividade. Revista Brasileira de Engenharia Agrícola e Ambiental, v.7, n.2, p.367-374.

Menegatti, L.A.A., Molin, J.P. (2004). Removal of errors in yield maps through raw data filtering. Revista Brasileira de Engenharia Agrícola e Ambiental, Campina Grande, v.8, n.1, p.126-134.

Molin, J.P.; Gimenez, L.M.(2000). Desenvolvimento de um algoritmo para redução de erros em mapas de rendimento obtidos em agricultura de precisão. In: Congresso e Mostra de Agroinformática, 1, Ponta Grossa, Paraná, 2000. InfoAgro 2000. Ponta Grossa.

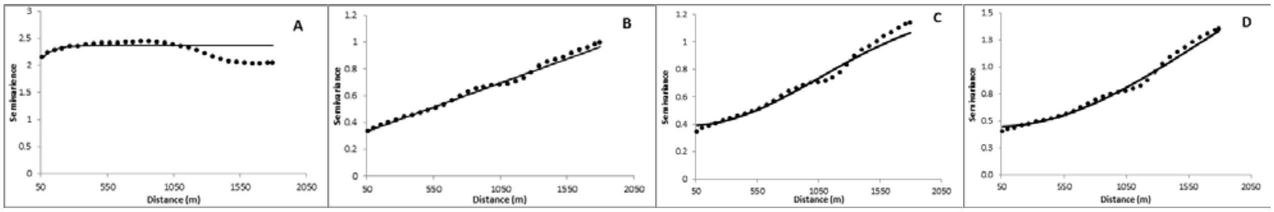
Ribeiro Junior, P. J. (1995). Métodos geoestatísticos no estudo da variabilidade espacial de variáveis do solo. 1995. 99 f. Dissertação (Mestrado em Agronomia)-Escola Superior deAgricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1995.

Simbahan, G. C., Dobermann, A., Ping, J. L. (2004). Screening yield monitor data improves grain yield maps. Agronomy Journal 96, 1091–1102.

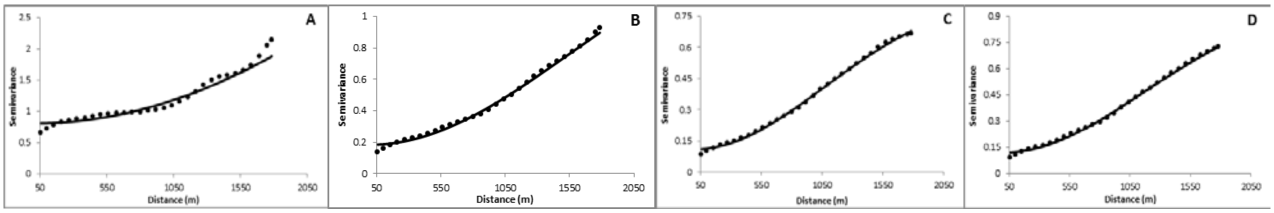
Spekken, M.; Anselmi, A. A.; Molin, J. P.(2013). A simple method for filtering spatial data. In: 9th European conference on precision agriculture, 2013, Lleida. Precision Agriculture`13. Wageningen: Wageningen Academic Publishers, 2013. v. 1. p. 1-823.

Appendix

Cotton



Corn



Wheat

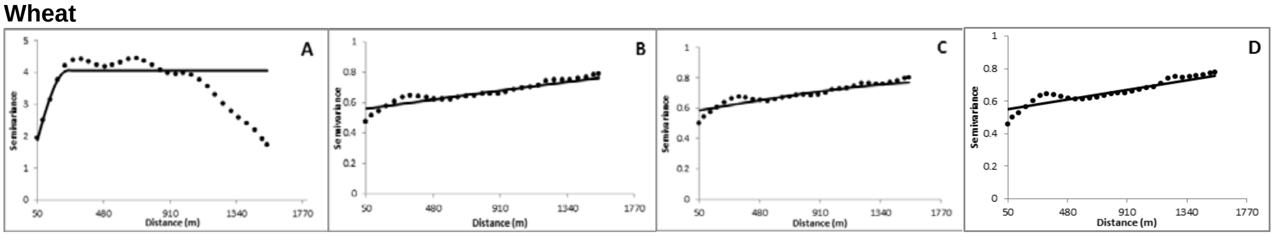


Fig 4. Variogram of original data (A) and of the data generated after the application of the method of correction *Machine by Machine* (B), *Track by Track* (C) and *Point by Point* (D).