Revista Brasileira de Climatologia
Brazilian Journal of Climatology

OPEN ACCESS

DOI: 10.5380/abclima

IDENTIFYING POTENTIAL REGIONS FOR A PRECIPITATION INDEX INSURANCE PRODUCT IN PARANÁ – BRAZIL: A HIERARCHICAL CLUSTERING APPROACH

IDENTIFICANDO REGIÕES EM POTENCIAL PARA UM PRODUTO DE SEGURO DE ÍNDICE CLIMÁTICO NO PARANÁ - BRASIL: UMA ABORDAGEM DE AGRUPAMENTO HIERÁRQUICO

IDENTIFICACIÓN DE REGIONES POTENCIALES PARA UN PRODUCTO DE SEGURO DE ÍNDICE DE PRECIPITACIÓN EN PARANÁ - BRASIL: UN ENFOQUE DE AGRUPAMIENTO JERÁRQUICO

Daniel Lima Miquelluti
Universidade de São Paulo
danielmiq@usp.br

Vitor Augusto Ozaki
Universidade de São Paulo
vitorozaki@usp.br

Abstract: In this article the availability and quality of public databases for soybean yields and daily rainfall in the state of Paraná in Brazil is assessed in order to verify the feasibility of an index insurance product. The multiple imputation by chained equations (MICE) method is utilized to fill missing values in the rainfall dataset and study the existence of spatial and temporal patterns in the data by means of hierarchical clustering. The results indicate that Paraná fulfills data requirements for a scalable weather index insurance with MICE and hierarchical clustering being effective tools in the pre-processing of precipitation data.

Keywords: Index-insurance. Hierarchical clustering. MICE.

Resumo: Neste artigo é avaliada a disponibilidade e a qualidade de bancos de dados públicos sobre a produção de soja e a precipitação diária no estado do Paraná no Brasil a fim de verificar a viabilidade de um produto de seguro de índice climático. O método de imputação múltipla por equações encadeadas (MICE) é utilizado para preencher valores ausentes no conjunto de dados de precipitação

e estudar a existência de padrões espaciais e temporais nos dados por meio de agrupamento hierárquico. Os resultados indicam que o Paraná cumpre os requisitos de dados para um seguro de índice climático escalável, com o MICE e o agrupamento hierárquico sendo ferramentas eficazes no pré-processamento dos dados de precipitação.
Palavras-chave: Seguro paramétrico. Agrupamento hierárquico. MICE.

Resumen: En este artículo, se evalúa la disponibilidad y la calidad de las bases de datos públicas para los rendimientos de la soja y las precipitaciones diarias en el estado de Paraná en Brasil con el fin de verificar la viabilidad de un producto de seguro paramétrico. El método de imputación múltiple por ecuaciones encadenadas (MICE) se utiliza para completar los valores faltantes en el conjunto de datos de lluvia y estudiar la existencia de patrones espaciales y temporales en los datos mediante agrupación jerárquica. Los resultados indican que Paraná cumple con los requisitos de datos para un seguro de índice meteorológico escalable con MICE y la agrupación jerárquica como herramientas efectivas en el procesamiento previo de datos de precipitación.
Palabras-clave: índice-seguro. Agrupación jerárquica. MICE.

Submetido em: 01/04/2020
Aceito para publicação em: 06/07/2021
Publicado em: 22/09/2021

INTRODUCTION

One of the flagships in the recent agricultural policy in Brazil, crop insurance has been advertised as one of the pillars of the 2016/2017 and 2017/2018 Agricultural and Livestock Plan (MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO, 2016; MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO, 2017). However, since its development in Brazil, this type of insurance has not achieved its intended endings with the protected area under 10% of the agricultural land (OZAKI, 2013). The low uptake is credited to the government insufficient investments in subsidies for the crop insurance program, however as noted by Oñate et al. (2016) one of the most subsidized crop insurance programs in Brazil, Proagro Mais, has failed to reduce uncertainty and risks. Also, as historic yields are not always available, insurers tend to use data provided by the Brazilian Institute of Geography and Statistics (IBGE), which are aggregated at the municipality level, thus pushing away high yield farmers and attracting the ones with low yields (when compared to the municipality average yield).

Relying on subsidies to increase crop insurance uptake seems not to be a good alternative as tax payers' and several countries' perception of farm subsidies worsens (Edwards, 2018). The benefits of this type of subsidy have shown to favor only the ones receiving it and not the entire community (DRABENSTOTT, 2015; BABCOCK, 2015; KIRWAN & ROBERTS, 2016). Therefore, subsidy free alternatives should be sought in order to improve the financial security of farmers.

This does not mean the government should end all crop insurance programs, but improve their self-sustainability. In this sense, one promising product is parametric insurance, which has lower premium costs when compared to traditional insurance. The absence of in situ claim adjustment and moral hazard monitoring greatly reduces the administrative costs of this type of insurance, permitting a subsidy free crop insurance (JENSEN & BARRETT, 2017). Another advantage of index insurance products is the rapid and payment of indemnities, also due to the non-existence of local loss assessment.

The basis of index insurance development is systemic risk, one of the factors halting conventional crop insurance expansion. The correlation of losses among policyholders causes significant increase in the indemnities, renting conventional crop insurance infeasible in the long run. Given that crops are exposed to a series of widespread risks, such as drought, floods

and windstorms it is clear that traditional crop insurance will not provide the adequate protection.

One of the key aspects of parametric insurance design is data, especially of high quality and from a sustained source. In the context of index insurance, high quality means a long, consistent and unbiased historical record. However, as noted in Collier et al. (2010) the data needs for a weather index insurance (WII) depends on the characteristics of the weather event insured.

Opposed to the traditional lines of insurance, parametric insurance relies on the spatial correlation of risks (systemic risks), so one of the first steps when designing this type of insurance is to determine the area affected by the event as this will indicate the necessary spatial resolution (RAO, 2011). Each event presents a spatial behavior, so the topography of the target region must be carefully studied, as a rough terrain alters weather patterns.

Aside from spatial correlation, temporal correlation is also important as weather events tend to follow a pattern in time. Such phenomena are observed in South America with the occurrence of El Niño and La Niña (ENSO), or in Asia with the monsoons. Data must have the proper temporal resolution to capture these seasonal patterns.

Just as important as historical weather data are historical records of loss and their cause, which will provide information of the impacts of different levels of the weather risk thus enabling the determination of an index trigger. Ideally, when developing a WII, one should be able to estimate the probability distribution function and correlations (presumably high) of each of these variables. A general benchmark for the minimum length of climatic data is 30 years (COLLIER ET AL., 2010).

In Brazil, parametric insurance was introduced in 2017 by Swiss Re for a single large producer of corn, cotton and soybean in the states of Bahia, Mato Grosso and Minas Gerais. However, the literature in the subject is still inexistent, even the Brazilian literature in crop insurance is also fragile. This is due, in part, to the data scarcity which was mitigated in 2016 by the release of a Crop Insurance Atlas by the Brazilian Ministry of Agriculture, Livestock and Supply. Therefore, the objective of this work is to assess if the state of Paraná is suitable for this type of product, regarding the data requirements and the existence of yield and rainfall spatial patterns. The chosen crop is soybean, given that Paraná is the second largest producer in Brazil with a total of 19,073,706 tons produced in 2017, being also the second in average

yields (3,663 kg/ha in 2017) according to the Brazilian Institute of Geography and Statistics (IBGE, 2019).

MATERIALS AND METHODS

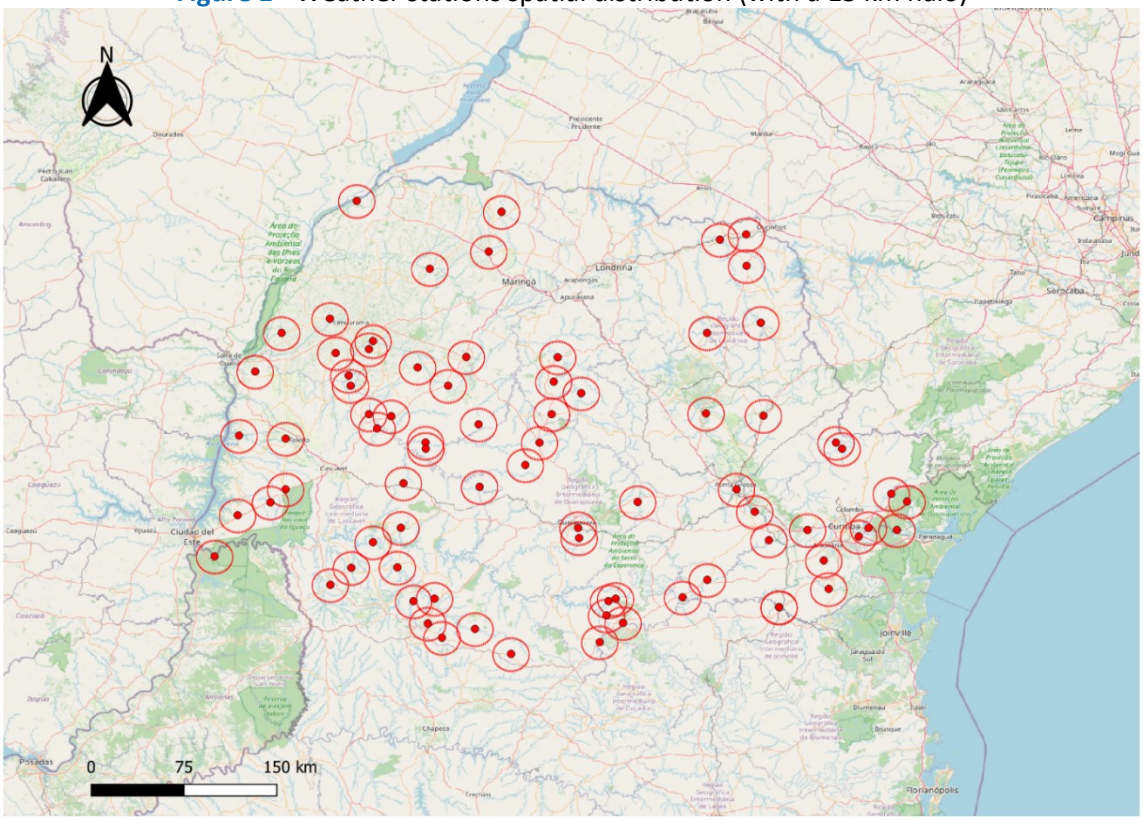
Daily precipitation data in Brazil are available from the National Water Agency (ANA) Hydrometeorological Network (MINISTÉRIO DO DESENVOLVIMENTO REGIONAL, 2005) and the National Institute of Meteorology (INMET), being that the former presents a more comprehensive distribution of weather stations from several sources in the state of Paraná. Therefore, only precipitation data from the National Hydrometeorological Network (RHN) was collected, spanning from 01/06/1973 through 31/12/2015 for a total of 1163 weather stations. This series was later aggregated in monthly totals.

Also, the series of annual soybean yields (in kg ha^{-1}) for each of the 399 municipalities in the state of Paraná, from 1980 through 2016, were obtained from the National Institute of Geography and Statistics (IBGE, 2019).

Data cleaning and yield detrending

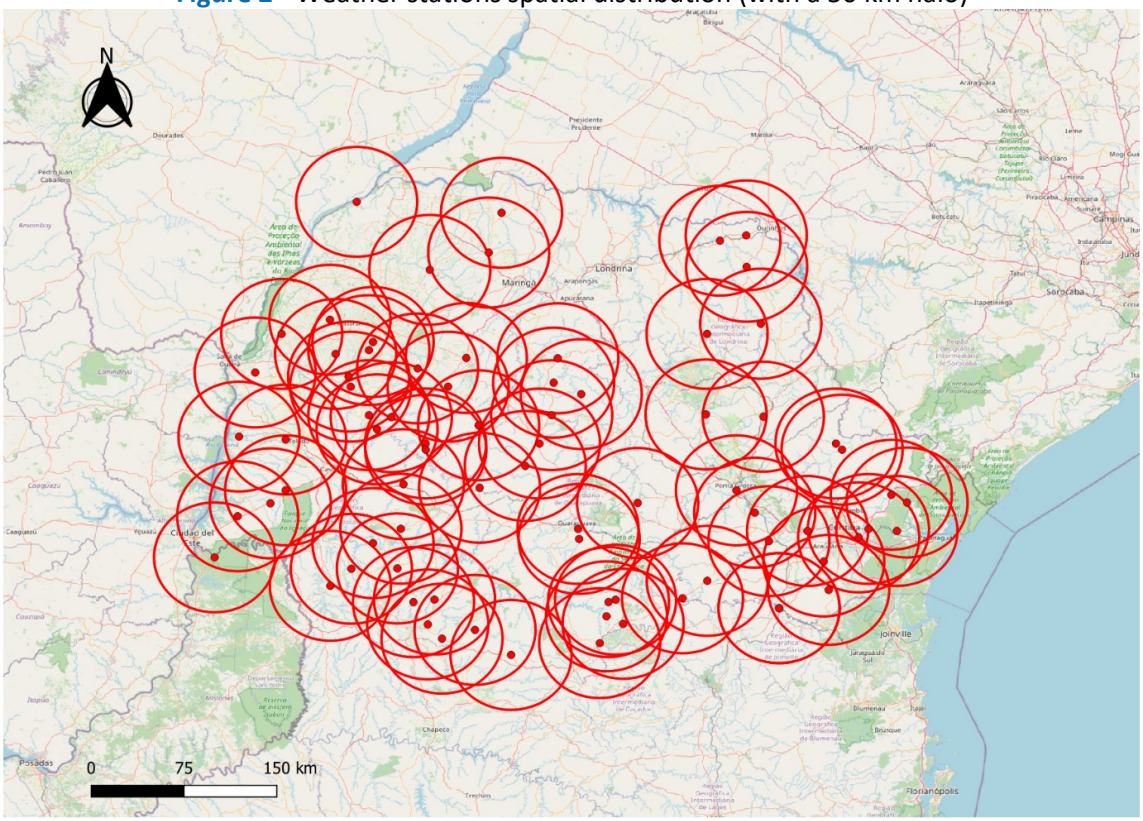
From the initial set of 1163 weather stations and 399 municipalities the ones with 15% or less of missing data were filtered, resulting in 78 stations and 174 municipalities. Values of precipitation were capped at 150mm to account for operational errors in the weather stations. Observing the spatial distribution of weather stations with less than 15% of missing data with a 15km halo there is an indication that a micro scale index insurance is not recommended (Figure 1). However, when a 50km halo is added there is only a portion of the state without coverage, mainly around the city of Londrina (Figure 2). This indicates an aptitude for parametric insurance at the meso and/or macro scales, targeted to cooperatives and other larger risk aggregators (COLLIER, 2010). At larger scales, weather index insurance permits the identification of large events and decreases the impact of basis risk. Microinsurance is possible for the municipalities with a weather station (78) and surrounding locations up to 15 km, however this greatly reduces the scalability of WII in Paraná.

Figure 1 - Weather stations spatial distribution (with a 15 km halo)



Source: Elaborated by the authors (2020)

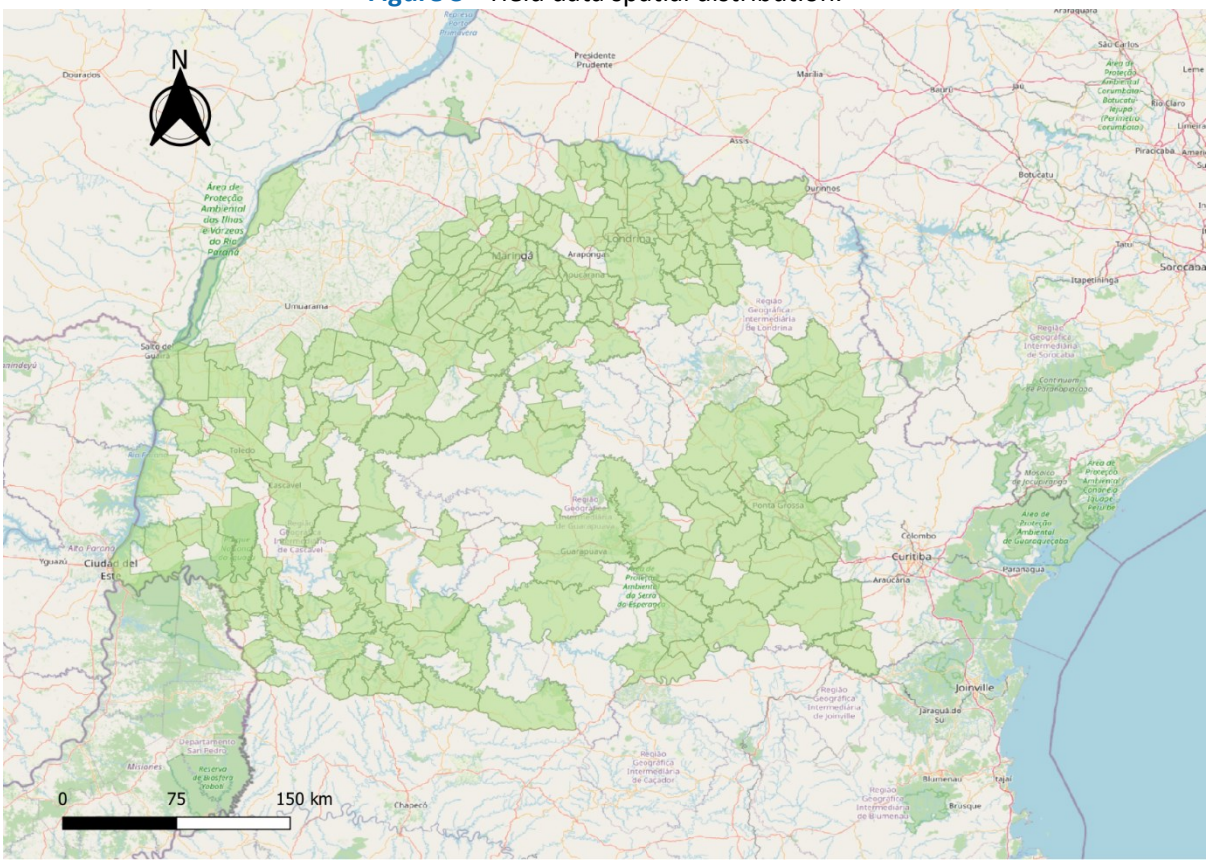
Figure 2 - Weather stations spatial distribution (with a 50 km halo)



Source: Elaborated by the authors (2020)

The yield data coverage is more disperse with some gaps, especially in the northwest and east portion of the state (Figure 3). In the northwest this lack of data reflects the characteristics of the region, with sandy soils and warm climate, being thus restrictive to the growth of soybean. Another reason for the low presence of soybean is the predominance of ranching in this region. This author also notes that in the east the presence of soybean is limited. Nevertheless, the available data represents the bulk of soybean producers in the state with approximately 70% of the state total production in 2016 (IBGE, 2019).

Figure 3 - Yield data spatial distribution.



Source: Elaborated by the authors (2020)

Crop yield data are subject to changes in practices and technology, which are not of interest for this study, therefore the yields were detrended. A linear regression was adjusted to the yield data with time as the explanatory variable, then the last observed yield was corrected using the model residuals for each year (GALLAGHER, 1987; DUARTE ET AL., 2018). The detrended yields are defined by the following equation:

$$\tilde{y}_t = \widehat{y_{2016}}\left(1 + \frac{\hat{e}_t}{\hat{y}_t}\right)$$

where $\tilde{y}_t$, $\hat{y}_t$ and $\hat{e}_t$ are, respectively, the corrected yield, the fitted yield and the residual for year t, $\widehat{y_{2016}}$ is the fitted yield for 2016.

This initial filtering is based on the five characteristics required in order to obtain a suitable dataset for weather index insurance design (COLLIER, 2010). The first is historical length, general standard of 30 years of data, permitting a better estimation of the probability distributions of derived indexes. The second is spatial specificity, which is dependent on the type of index insurance product to be designed. Farm-level products require at least one weather station each 15km, while meso and macro level products will perform adequately with one weather station each 50-100km. The third characteristic is temporal specificity, regarding the availability of data on a timely basis. For the purpose of drought monitoring, the daily precipitation from ANA will be sufficient. The fourth characteristic is completeness, which is why weather stations with more than 15% of missing data was removed from the dataset and Multiple Imputation by Chained Equations (MICE) is employed to fill the missing gaps on the data. The last characteristic is validity, giving both the insurer and the client confidence that the data comes from a source that cannot be tampered by any of the involved parties. This is why only data from public institutions were considered in this work.

Imputation for precipitation data

Given the existence of missing data Multiple Imputation by Chained Equations (MICE) was applied, a method that combines imputation for multivariate data (RUBIN, 1988) and Fully Conditional Specification, which was developed under several names, being chained equations the one implemented here using the R software (VAN BUUREN, 2011).

While multiple imputation considers a single imputation model for each variable with missing values, the chained equations technique permits the use of separate and univariate imputation models for each of these variables (BARTLETT ET AL., 2015). In this way, hundreds of variables may be imputed with a high degree of flexibility (HE ET AL., 2010). Continuous variables may be modeled through linear regression and binary variables through logistic regression for example (CHEVRET ET AL., 2015). However, MICE does not have the same

theoretical basis as other methods such as multivariate normal imputation, what does not seem to be an issue (WHITE ET AL., 2011).

A natural question when using imputation methods is whether the missing rate may be too high to use multiple imputation methods such as MICE. Research shows that these methodologies are unbiased when data is missing at no higher than 50%, being unstable for higher percentages, especially if the data distribution is asymmetrical (LEE & CARLIN 2012; HAJI-MAGHSOUDI ET AL., 2013). However, this does not imply that multiple imputation should be discarded as it exhibits superior performance to other methods even for a 75% data loss, despite biased estimates (MARSHALL ET AL., 2010).

For a partially observed random sample of the multivariate distribution $\mathrm{P}(\mathrm{Y}|\theta)$, completely specified by the vector of k unknown parameters θ and representing the complete data Y, the posterior distribution of θ and then the predictive distribution of Y are obtained through a Gibbs sampler of the form:

$$\begin{aligned}
\theta_1^{\star(t)} &\sim \mathrm{P}\left(\theta_1 \middle| Y_1^{(obs)}, Y_2^{(t-1)}, \dots, Y_k^{(t-1)}\right) \\
Y_1^{\star(t)} &\sim P\left(Y_1 \middle| Y_1^{(obs)}, Y_2^{(t-1)}, \dots, Y_k^{(t-1)}, \theta_1^{\star(t)}\right) \\
&\vdots \\
\theta_k^{\star(t)} &\sim \mathrm{P}\left(\theta_k \middle| Y_k^{(obs)}, Y_1^{(t-1)}, \dots, Y_{k-1}^{(t-1)}\right) \\
Y_k^{\star(t)} &\sim P\left(Y_k \middle| Y_k^{(obs)}, Y_1^{(t-1)}, \dots, Y_{k-1}^{(t-1)}, \theta_k^{\star(t)}\right)
\end{aligned}$$

where $Y_j^t = (Y_j^{(obs)}, Y_j^{*(t)})$ is the jth imputed variable at iteration t and $Y^{(obs)}$ is the portion of Y that is observed.

The predictive mean matching (PMM) imputation method was chosen within MICE given that precipitation is generally skewed, thus not normally distributed. Nevertheless, simulations have shown that normal imputation models do work with non-normal data (GRAHAM & SCHAFER, 1999). Imputations made through PMM better resemble the observed values than methods based on the normal distribution (WHITE ET AL., 2011). This follows from the way PMM work as it uses the predicted value for a given missing value to identify similar observations. These identified observations are used to create a matching set is containing q matches, from which PMM then draws a random observation. Therefore, PMM uses the real

observed values to fill the missing data and thus preventing extrapolation beyond the range of the data (LITTLE, 1988).

In order to capture seasonal changes, latitude, longitude, and month and year binaries were chosen as covariates. With this specification, the MICE procedure assumes Y being normally distributed[1] and estimates a linear multiple regression. This yields a $\widehat{\boldsymbol{\beta}}$ vector of parameters (of length *k*), with an estimated covariance matrix $\boldsymbol{V}$ and root mean-squared error $\widehat{\sigma}$, from fitting this model to $Y^{(obs)}$.

The next step is to draw the imputation parameters $\sigma^{\star}$, $\boldsymbol{\beta}^{\star}$ from the exact joint posterior distribution of σ, $\boldsymbol{\beta}$. The parameter $\sigma^{\star}$is drawn as $\sigma^{\star} = \widehat{\sigma}\sqrt{(n_{\text{obs}} - k)/g}$, where n_{obs} if the number of observed values, *g* is a random draw from a χ^2 distribution with $n_{obs} - k$ degrees of freedom. Then, $\boldsymbol{\beta}^{\star}$ is drawn as $\boldsymbol{\beta}^{\star} = \widehat{\boldsymbol{\beta}} + \frac{\sigma^{\star}}{\widehat{\sigma}}\boldsymbol{u}_1\mathbf{V}^{1/2}$, where $\boldsymbol{u}_1$ is a vector of *k* independent random draws from a standard Normal distribution and $\mathbf{V}^{1/2}$ is the Cholesky decomposition of **V**.

For each missing value Y_i with covariates $\boldsymbol{X}_i$ PMM identifies the *q* individuals with the smallest values of $\left|\widehat{\boldsymbol{\beta}}\boldsymbol{X}_o - \boldsymbol{\beta}^{\star}\boldsymbol{X}_i\right|$ $(o = 1, \ldots, n_{obs})$. Of these *q* closest individuals, one is chosen at random ($Y_{h'}$), and the imputed value of Y_i is $Y_{h'}$. Thus, the imputed value is an observed value of *Y* whose prediction is closely matched by the perturbed prediction.

The size of the matching set is chosen by the researcher with values like $q = 1$ in leading to estimated standard errors that are too low and t-statistics that are too large (MORRIS ET AL., 2014). Whereas values ranging from $q = 3$ over $q = 10$ showed a small advantage (SCHENKER & TAYLOR, 1996; MORRIS ET AL., 2014). The size of the matching set is dependent on sample size and may have poor performance in small samples as the difference between similar observations is increased.

PMM has shown similar performance to correctly specified parametric models and better than poorly specified ones characterized by non-normality (SCHENKER & TAYLOR 1996, MORRIS ET AL., 2014) and skewness (MARSHALL ET AL., 2010) considering that the method does not have a strong theoretical backing (KENWARD & CARPENTER, 2007).

1 This assumption does not affect the quality of the imputations as this regression is simply a metric for matching (Little, 1988).

Finally, for this analysis the number of repeated imputations was *m*=5, *q*=5 and the number of iterations was *t*=20. The quality of the imputations was checked using the Kolmogorov-Smirnov test in order to check departures from the original distribution of data (BONDARENKO & RAGHUNATHAN, 2016).

The application of MICE has been successful in several areas, including precipitation data imputation in Brazil by de Carvalho et al (2017).

Clustering procedures

Prior to the application of hierarchical clustering, precipitation data was aggregated monthly and the standardized precipitation index (SPI) with a three-month scale was calculated, thus capturing drought events during the crop season (ZARCH ET AL, 2015). We chose the Ward's clustering method with an Euclidean distance matrix since it has already proved successful in defining homogenous precipitation regions in Brazil (KELLER FILHO, 2005). The optimal number of clusters was obtained through the majority rule of 30 indices, an algorithm implemented in Charrad et al. (2014).

RESULTS AND DISCUSSION

Imputations

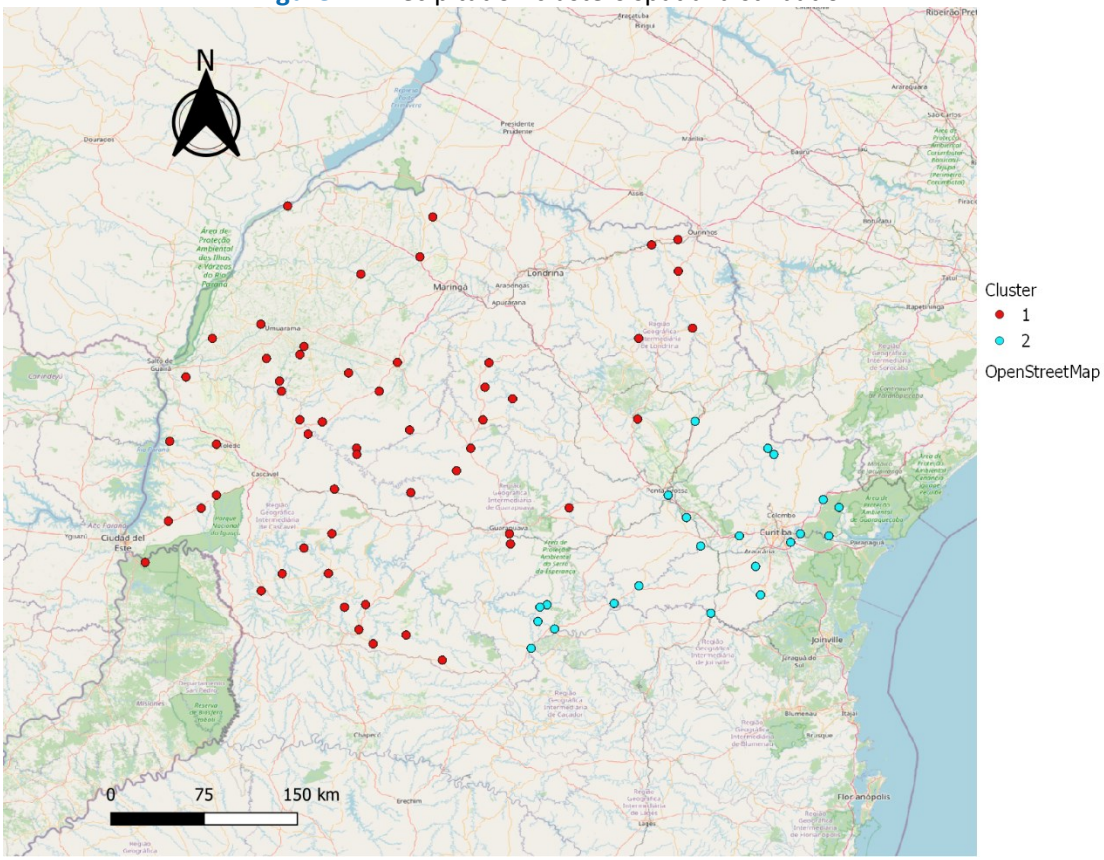
According to the Kolmogorov-Smirnov test the distribution of the imputed precipitation does not differ from the original dataset (D = 0.005964, p-value = 0.1016), therefore the procedure did not alter the underlying structure of the data. This result reinforces the use of MICE as a valid imputation procedure for precipitation data in Brazil (DE CARVALHO ET AL., 2017). It must be noted that albeit its effectiveness, MICE should be used with caution in datasets with 50% or more of missing values. Also, the specification of the correct imputation model and quality of predictors plays a large role in the quality of the imputations (WHITE ET AL, 2011).

a. Precipitation clusters

According to the majority rule, the optimal number of clusters was two, with nine votes, followed by three clusters with six votes (Table 1). Given that it is used a different approach to the clustering methodology than in Keller Filho et al (2005), where several statistical

parameters are calculated from five-day accumulated precipitation, and here the three-month SPI is used, the results do not completely match but are very similar regarding the characteristics of the clusters. Cluster 1 represents areas in the west, center and north of the state, with higher total precipitation in the year aggregate but greater variability among years. Whereas cluster 2 represents the center and east of Paraná, with a lower total precipitation but with less variability (Figure 4).

Figure 4 - Precipitation clusters spatial distribution



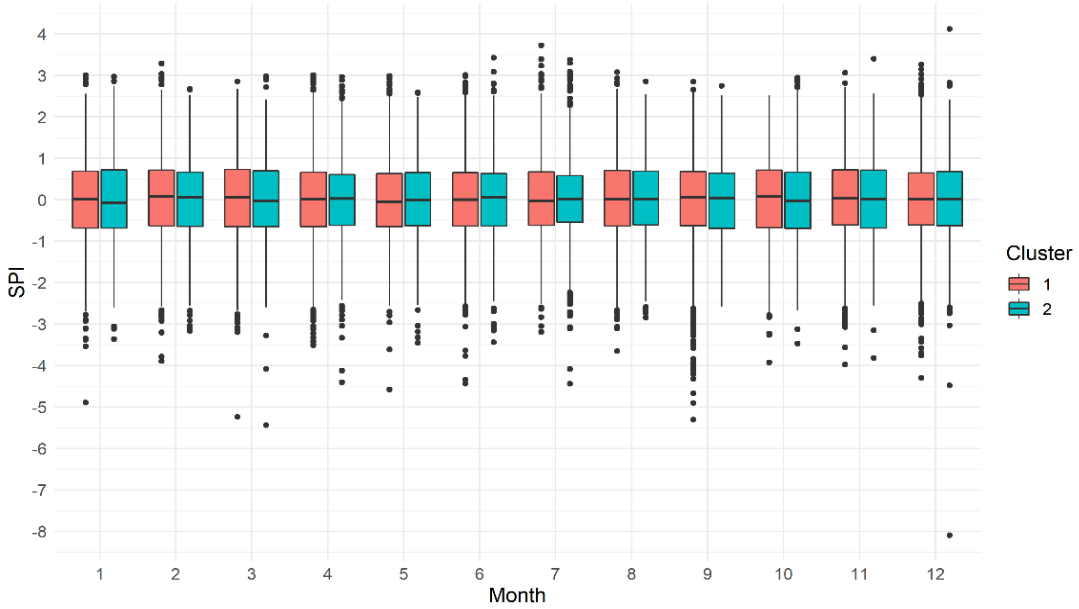
Source: Elaborated by the authors (2020)

Ideal number of clusters for each variable

Regarding the SPI values for each month and cluster, it is interesting to observe that there is little difference in the median of monthly SPI, albeit statistically significant according to the cluster analysis (Figure 5). When carefully analyzed, it can be observed that cluster 1 has a greater number of observations in the lower ranges of SPI, indicating the occurrence of moderate and severe droughts. This can be explained by the greater variability in

precipitation, and the occurrence of droughts in the north and northeast of the state as identified by Fritzsons et al. (2011).

Figure 5 - Monthly SPI boxplot, per cluster.

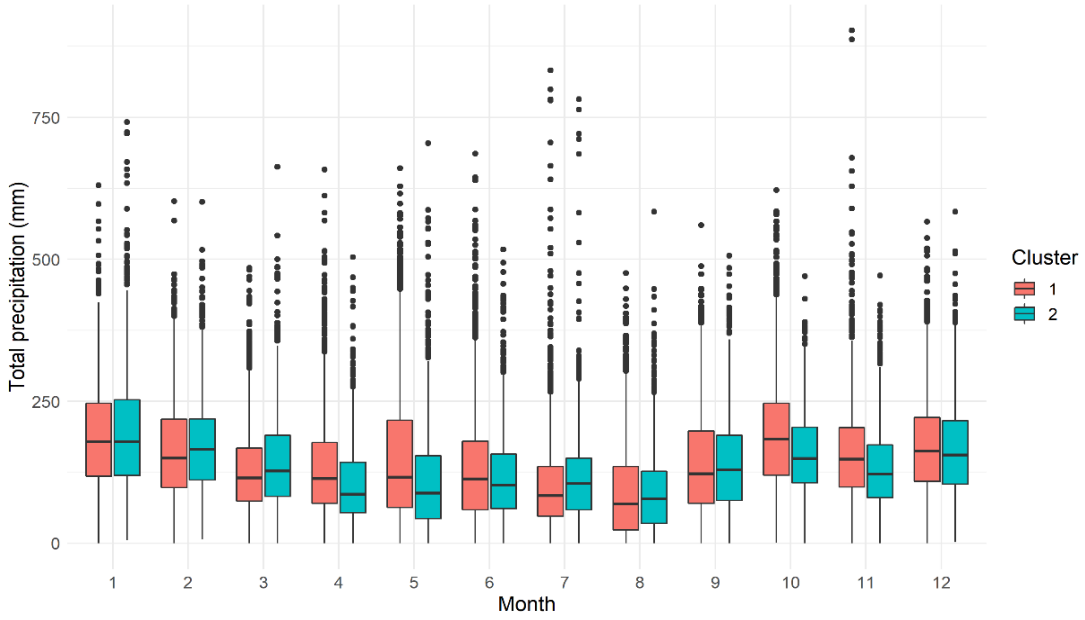


Source: Elaborated by the authors (2020)

When analyzing only the period in which soybeans are grown in the state, October through March, cluster 1 presents variable conditions, as there is a surplus in precipitation during the growth and reproductive stages with a decrease in precipitation in the end of the growth period (Figure 6). However, there must be caution with the occasional occurrence of drought, which can be mitigated using irrigation or risk management products such as crop insurance. Despite the decrease in precipitation from January through April/May, the total precipitation in this period is sufficient for cultivars ranging from 450 to 700 mm of water requirements.

For Cluster 2, the opposite is observed with lower levels of precipitation from October through December and higher levels in January and February. However, in these areas, there is a steeper descent in precipitation levels, being the region adequate for cultivars requiring from 450 to 650 mm of water. It must be noted that areas represented in cluster 2 have a lower variability, thus, it suffers less from drought and excessive rain periods (Figures 6 and 7).

Figure 6 - Monthly precipitation (total) boxplot, per cluster.

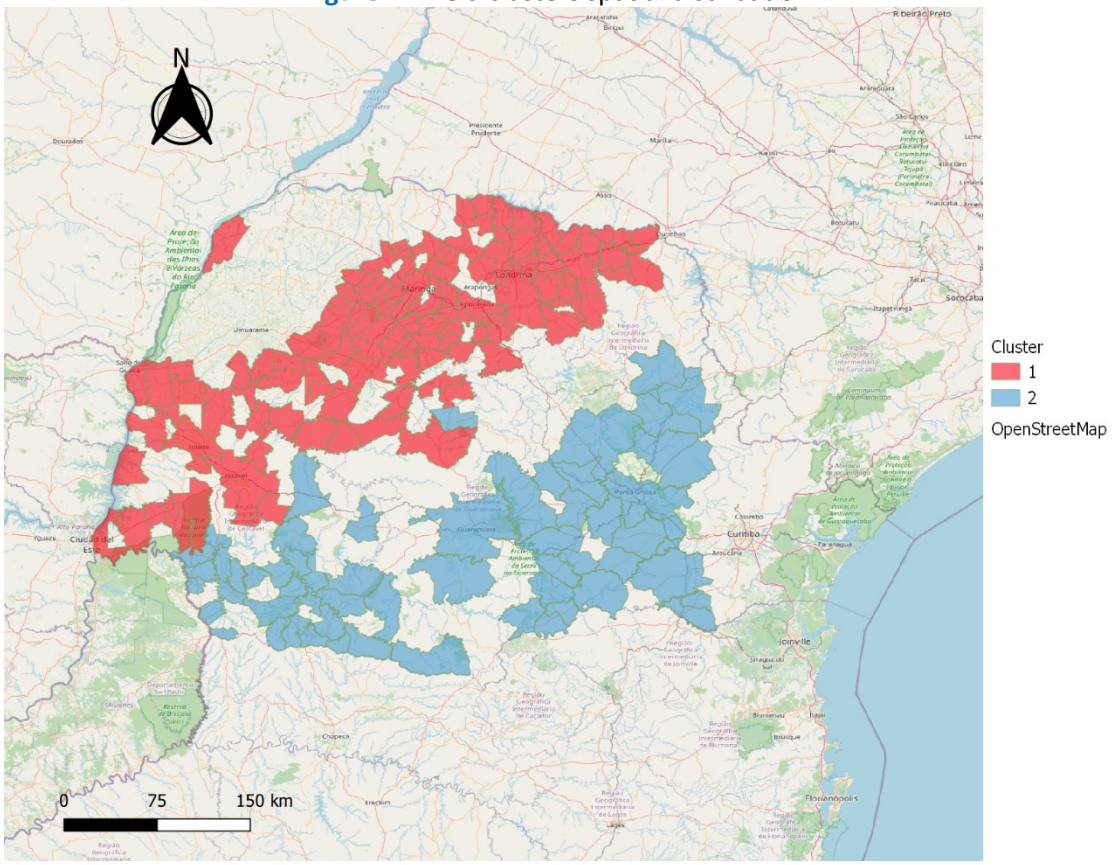


Source: Elaborated by the authors (2020)

Yield clusters

According to the majority rule, the optimal number of clusters was two, with eleven votes, followed by three clusters with eight votes (Table 1). Similarly to the precipitation clusters, cluster 1 represents the west and northwest of the state while cluster 2 comprehends the south, center and east of Paraná (Figure 7). Thus, the only difference from the rainfall clusters is that the yield cluster 1 has less presence in the center and south of Paraná.

Figure 7 - Yield clusters spatial distribution.



Source: Elaborated by the authors (2020)

Table 1 - Ideal number of clusters for each variable

	Variable	Ideal number of clusters*								
		0	1	2	3	5	6	7	8	10
Votes	Rainfall	2	1	9	6	0	1	1	2	1
	Yield	2	0	11	8	2	0	0	1	2

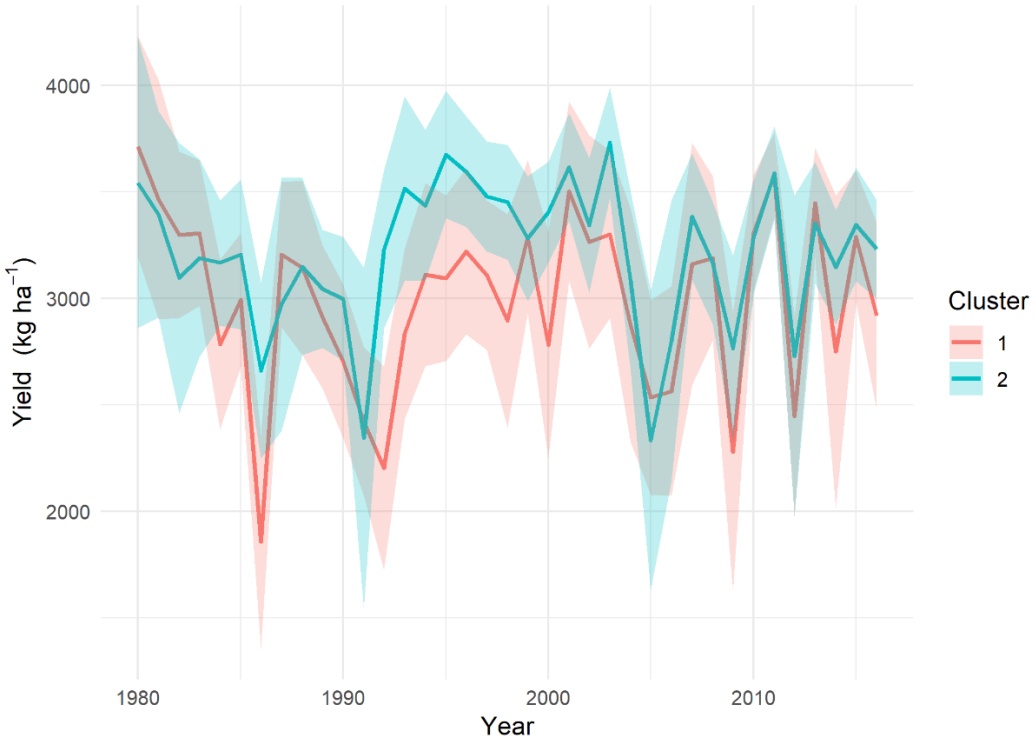
*Only numbers with at least one vote are presented

Source: Elaborated by the authors (2020)

Both clusters present a similar yield level from the beginning of the series through 1990 and from 2001 onwards, however, in the period comprised between 1991 and 2000 cluster 1 has lower yields (Figure 8). Also, in years where losses occurred (1986, 1991, 1992, 2005, 2009, 2012), cluster 1 municipalities suffered greater losses, increasing cluster variability and decreasing the mean and median of the whole period (Table 2). The latter can be explained by the presence of municipalities in the northern portion of the state in cluster 1, as said in the previous section this region has sandy soils and higher temperatures, being more

susceptible to drought. Other researchers such as Felema et al. (2016) also study the spatial behavior of soybean yields in Paraná. However, while the results from this paper agree to some measure, no comparisons are made as both studies used only two years of data against the 37-year time series of the present study.

Figure 8 - Soybean yields time series with 95% confidence intervals, per cluster.



Source: Elaborated by the authors (2020)

Table 2 - Yield clusters descriptive statistics.

Cluster	Mean	Median	Standard deviation	Coefficient of variation (%)
1	2993,46	3051,21	597,05	19,95
2	3209,30	3282,11	533,26	16,62

Source: Elaborated by the authors (2020)

Yield and precipitation clusters relationship

When comparing with the results found for the precipitation clusters, the need to consider other environmental variables is exemplified. Regardless of the precipitation cluster 2 having lower precipitation levels, other factors such as soil type and temperature lead to greater yields in this region. The southwest of Paraná is the only region with high precipitation

and high yields. Intersecting the clusters would lead to a further separation, with three separate regions, the southeast with lower precipitation levels and high yields, the west and center with good precipitation levels (but with higher variability) and lower yields and the southwest as described above. These “new clusters” could present separate regions for the design of a weather index insurance products, with each region having a fine-tuned product.

This analysis does not encompass soil and other weather variables, which are also important in the determination of the suitability of cultivars for each region. The northwest of Paraná presents sandy soils and higher temperatures, therefore, farms in this region suffer more from drought periods as these soils have a lower water holding capacity and the increase in temperature leads to a higher evapotranspiration. On the contrary, for the south portion of the state, soils are rich in clay, altitudes are higher and temperatures lower, this coupled with a low variability in precipitation results in a lower risk of drought related yield losses (LIMA ET AL., 2012). Consequently, when choosing adequate risk management strategies and in the design of crop insurance products, such as weather index insurance, these variables must be taken in account.

CONCLUSION

Verifying the availability and quality of data sources is one of the first steps when designing a weather index insurance product. This step is particularly difficult in large developing countries such as Brazil, where the weather agencies do not have the necessary funds to maintain a large net of weather stations. Given this lack of resources, the existing stations also suffer from missing data, a problem that generally implies in pricier insurance. In this paper the quality of precipitation and yield data in Paraná-Brazil is evaluated and a proven method to deal with missing data is presented.

Despite the variability of soil and temperature conditions it is found that the state of Paraná presents a great opportunity for index insurance based on precipitation data. There is a good coverage of suitable weather stations and the clusters found indicate the scalability of WII and the existence of spatially correlated weather events. The sharp decrease in weather stations from the original set to the filtered one is due to the lack of historical data in many of the stations, as the number of operational stations is around 900, thus the weather station coverage should improve with time.

It is found that MICE proved a reliable method to fill gaps in precipitation data with up to 15% of missing observations, therefore it should be considered by insurers as an alternative to the practice of loading insurance premium in cases where data is not complete. This would provide a more attractive product without losing precision in the pure risk estimates as the method does not change the probability distribution of data.

This article presents a beginning of the exploration of weather index insurance design in the Brazilian literature, as the economic viability of index insurance was not verified, focusing only in the technical aspects required for its operation. Thus, additional studies are required to determine if WII is a viable option to the crop insurance market in Paraná and how it compares to existing crop insurance products.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

BABCOCK, Bruce A. The concentration of US agricultural subsidies. **Iowa Ag Review**, v. 7, n. 4, p. 4, 2015.

BARTLETT, Jonathan W. et al. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. **Statistical methods in medical research**, v. 24, n. 4, p. 462-487, 2015.Bondarenko, I.; Raghunathan, T. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Statistics in medicine, v. 35, n. 17, p.3007-3020, 2016.

BONDARENKO, Irina; RAGHUNATHAN, Trivellore. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. **Statistics in medicine**, v. 35, n. 17, p. 3007-3020, 2016.

CHARRAD, Malika et al. NbClust: an R package for determining the relevant number of clusters in a data set. **Journal of statistical software**, v. 61, n. 1, p. 1-36, 2014.

CHEVRET, S.; SEAMAN, Shaun; RESCHE-RIGON, M. Multiple imputation: a mature approach to dealing with missing data. **Intensive care medicine**, v. 41, n. 2, p. 348-350, 2015.

COLLIER, Benjamin; SKEES, Jerry; BARNETT, Barry. Weather index insurance and climate change: Opportunities and challenges in lower income countries. **The Geneva Papers on Risk and Insurance-Issues and Practice**, v. 34, n. 3, p. 401-424, 2009.

COLLIER, Benjamin; BARNETT, Barry; SKEES, Jerry R. State of Knowledge Report — Data Requirements for the Design of Weather Index Insurance. **Bill & Melinda Gates Foundation**, [*s. l.*], 2010. Disponível em: http://globalagrisk.com/Pubs/2010_GlobalAgRisk_State_of_Knowledge_Data_sept.pdf. Acesso em: 21 out. 2020.

DE CARVALHO, José Ruy Porto et al. Model for multiple imputation to estimate daily rainfall data and filling of faults. **Revista Brasileira de Meteorologia**, v. 32, p. 575-583, 2017.

DRABENSTOTT, Mark. Do farm payments promote rural economic growth?. **Ag Decision Maker Newsletter**, v. 9, n. 6, p. 2, 2015.

DUARTE, Gislaine V. et al. Modeling of soybean yield using symmetric, asymmetric and bimodal distributions: implications for crop insurance. **Journal of Applied Statistics**, v. 45, n. 11, p. 1920-1937, 2018.

EDWARDS, Chris. **Agricultural subsidies**. 2018.

FELEMA, João et al. Um estudo da produtividade do feijão, do milho e da soja na agricultura paranaense, nos anos de 2000 e 2010: uma análise espacial. **Ensaios Fee**, v. 36, n. 4, p. 817-842, 2016.

FRITZSONS, Elenice et al. Análise da pluviometria para definição de zonas homogêneas no Estado do Paraná. **Raega-O Espaço Geográfico em Análise**, v. 23, 2011.

GALLAGHER, Paul. US soybean yields: Estimation and forecasting with nonsymmetric disturbances. **American Journal of Agricultural Economics**, v. 69, n. 4, p. 796-803, 1987.

GRAHAM, John W.; SCHAFER, Joseph L. On the performance of multiple imputation for multivariate data with small sample size. **Statistical strategies for small sample research**, v. 50, p. 1-27, 1999.

HAJI-MAGHSOUDI, Saiedeh et al. Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons. **International journal of health policy and management**, v. 1, n. 1, p. 69, 2013.

HE, Yulei et al. Multiple imputation in a large-scale complex survey: a practical guide. **Statistical methods in medical research**, v. 19, n. 6, p. 653-670, 2010.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA (IBGE). IBGE Automatic Recovery System = Sistema IBGE de Recuperação Automática, 2019. Available in: https://sidra.ibge.gov.br/home/ipp/brasil. [Accessed: dec. 2019].

JENSEN, Nathaniel; BARRETT, Christopher. Agricultural index insurance for development. **Applied Economic Perspectives and Policy**, v. 39, n. 2, p. 199-219, 2017.

KELLER FILHO, Thadeu; ASSAD, Eduardo Delgado; LIMA, Paulo Roberto Schubnell de Rezende. Regiões pluviometricamente homogêneas no Brasil. **Pesquisa Agropecuária Brasileira**, v. 40, p. 311-322, 2005.

KENWARD, Michael G.; CARPENTER, James. Multiple imputation: current perspectives. **Statistical methods in medical research**, v. 16, n. 3, p. 199-218, 2007.

KIRWAN, Barrett E.; ROBERTS, Michael J. Who Really Benefits from Agricultural Subsidies? Evidence from Field-level Data. **American journal of agricultural economics**, v. 98, n. 4, p. 1095-1113, 2016.

LEE, Katherine J.; CARLIN, John B. Recovery of information from multiple imputation: a simulation study. **Emerging themes in epidemiology**, v. 9, n. 1, p. 1-10, 2012.

LIMA, V.C.; LIMA, M.R.; MELO, V.F. **Knowing the main soils of Paraná: approach for elementary and middle school teachers** = Conhecendo os principais solos do Paraná: abordagem para professores do ensino fundamental e médio. Sociedade Brasileira de Ciência do Solo (Eds.). Núcleo Estadual Paraná, Brasil, 18, 2012.

LITTLE, Roderick JA. Missing-data adjustments in large surveys. **Journal of Business & Economic Statistics**, v. 6, n. 3, p. 287-296, 1988.

MARSHALL, Andrea; ALTMAN, Douglas G.; HOLDER, Roger L. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. **BMC medical research methodology**, v. 10, n. 1, p. 1-10, 2010.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO – MAPA. **Departamento De Gestão De Riscos.** General Report 2017 - Rural Insurance Premium Grant Program (PSR) = Relatório Geral 2017 – Programa de Subvenção ao Prêmio do Seguro Rural (PSR), 2017. Available at: http://www.agricultura.gov.br/assuntos/riscos-seguro/seguro-rural/documentos-seguro-rural/RelatorioGeralPSR2017.pdf [Accessed Dec 22, 2019]. (in Portuguese).

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO – MAPA. PLANO AGRÍCOLA E PECUÁRIO 2016/2017. MAPA **Indicadores**. Accessed December 07, 2019. http:// http://www.agricultura.gov.br/assuntos/sustentabilidade/plano-abc/arquivo-publicacoes-plano-abc/PAP1617.pdf.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO – MAPA. PLANO AGRÍCOLA E PECUÁRIO 2017/2018. MAPA Indicadores, 2018. Accessed December 07, 2019.

MINISTÉRIO DO DESENVOLVIMENTO REGIONAL (Brasil). Agência Nacional de Águas (ANA). **Rede Hidrometeorológica Nacional**. [*S. l.*], 2005. Disponível em: https://www.snirh.gov.br/hidroweb/apresentacao. Acesso em: 14 out. 2020.

MORRIS, Tim P.; WHITE, Ian R.; ROYSTON, Patrick. Tuning multiple imputation by predictive mean matching and local residual draws. **BMC medical research methodology**, v. 14, n. 1, p. 1-13, 2014.

OÑATE, Carlos Andrés; OZAKI, Vitor Augusto; BRAVO-URETA, Boris. **Impact Evaluation of the Brazilian crop insurance public program "Proagro Mais"**. 2016.

OZAKI, Vitor Augusto. Qual o custo governamental do seguro agrícola?. **Revista de Economia e Sociologia Rural**, v. 51, p. 123-136, 2013.

RAGHUNATHAN, Trivellore E. et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. **Survey methodology**, v. 27, n. 1, p. 85-96, 2001.

RAO, Kolli N. Weather index insurance: Is it the right model for providing insurance to crops?. **ASCI Journal of Management**, v. 41, n. 1, p. 86-101, 2011.

RUBIN, Donald B. An overview of multiple imputation. In: **Proceedings of the survey research methods section of the American statistical association**. Citeseer, 1988. p. 79-84.

RUBIN, Donald B. **Multiple imputation for nonresponse in surveys**. John Wiley & Sons, 2004.

SCHENKER, Nathaniel; TAYLOR, Jeremy MG. Partially parametric techniques for multiple imputation. **Computational statistics & data analysis**, v. 22, n. 4, p. 425-446, 1996.

VAN BUUREN, Stef; BOSHUIZEN, Hendriek C.; KNOOK, Dick L. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in medicine**, v. 18, n. 6, p. 681-694, 1999.

VAN BUUREN, Stef. Multiple imputation of discrete and continuous data by fully conditional specification. **Statistical methods in medical research**, v. 16, n. 3, p. 219-242, 2007.

VAN BUUREN, Stef; GROOTHUIS-OUDSHOORN, Karin. mice: Multivariate imputation by chained equations in R. **Journal of statistical software**, v. 45, n. 1, p. 1-67, 2011.

WHITE, Ian R.; ROYSTON, Patrick; WOOD, Angela M. Multiple imputation using chained equations: issues and guidance for practice. **Statistics in medicine**, v. 30, n. 4, p. 377-399, 2011.

ZARCH, Mohammad Amin Asadi; SIVAKUMAR, Bellie; SHARMA, Ashish. Droughts in a warming climate: A global assessment of Standardized precipitation index (SPI) and Reconnaissance drought index (RDI). **Journal of hydrology**, v. 526, p. 183-195, 2015.