A study on heteroskedasticity assumptions effects over the crop yield insurance premiums

VICTOR FERNANDO SILVA
Universidade de São Paulo
JOÃO VINÍCIUS DE FRANÇA CARVALHO
Universidade de São Paulo

Resumo
O seguro de produtividade agrícola é uma das principais ferramentas de gestão de risco do agronegócio do Brasil, movimentando mais de R$2 bilhões anuais. O modelo tradicional do seguro agrícola consiste em indenizar o produtor em anos de quebra de safra, cobrindo a diferença entre o valor dos recursos efetivamente colhidos e a cobertura securitária contratada. Para tarifá-lo, estima-se um modelo de regressão e, a partir dos seus resíduos, são calculadas as possíveis indenizações. Contudo, não há consenso na literatura sobre a melhor estrutura destes resíduos. Este trabalho tem como objetivo avaliar as principais suposições empregadas na precificação do cálculo de prêmios de seguro agrícola no Brasil: homoscedasticidade ou heterocedasticidade proporcional. Foram obtidos dados oficiais da Produção Agrícola Municipal, do IBGE, de três diferentes culturas (soja, trigo e arroz) de 1.876 municípios brasileiros, dispostas anualmente entre 1974-2018. Os resultados sugerem que, para a maioria dos municípios brasileiros, as duas hipóteses são aceitas para as culturas de soja, arroz e trigo. Boa parte dos municípios onde não se pode aceitar as hipóteses são pouco expressivos no mercado de seguro agrícola e possuem rendimentos mais voláteis ao longo do tempo. Por fim, a adoção da hipótese proporcional mostrou-se, na maioria das vezes, mais adequada para reduzir a taxa de prêmios de seguro.
Palavras-chave: seguro agrícola, heteroscedasticidade, volatilidade, subsídio, custo do seguro

Abstract
The crop yield insurance is one of the main risk management tools in the Brazilian agribusiness, generating over BRL 2 billion yearly. The traditional crop insurance model consists in compensating agricultural producers in crop shortfall years, covering the gap between the harvested and covered yield. To price this insurance policy, a regression model is estimated over the historical data and from its residuals, the shortfall and claims are calculated. However, there is no agreement over the optimal structure of these deviations. Our objective is to evaluate the main assumptions used in crop yield insurance pricing in Brazil: homoskedasticity and proportional heteroskedasticity. The official data of agricultural production from IBGE for three crops (soybeans, wheat and rice) for 1.876 Brazilian municipalities, presented yearly, between 1974 and 2018. The results suggest that for most Brazilian municipalities, both assumptions may be accepted for the three crops. Many municipalities where the assumptions cannot be accepted are not significant in the crop insurance market and its historical yields are more volatile. Finally, the adoption of the proportional heteroskedasticity reflects in reduction of crop insurance rates.
Keywords: crop insurance, heteroskedasticity, volatility, subsidy, insurance cost.

Classificação JEL: G22; Q02, Q14.
Área 11: Economia Agrícola e do Meio Ambiente

1. Introduction

The crop yield insurance is one of the main risk management tools in the Brazilian agribusiness. The total premium collected by insurance companies surpassed BRL 2 billion in 2018. The amount subsidized by the government program *Programa de Subvenção Federal ao Prêmio de Seguro Rural* (PSR) represents 43% of this value. Soybeans were the most supported crop in the program, with BRL 155 million (Superintendência de Seguros Privados, 2019).

In the U.S., the Risk Management Agency (RMA) is the responsible for the counties' ratemaking, the equivalent to the Brazilian municipalities (Harri et al., 2011). In Brazil, although the crop insurance pricing is designed by private companies, the Ministry of Agriculture, Livestock and Supply (MAPA), through PSR, subsidies part of these premiums, reaching up to 40% of the individual premium. With a BRL 368 million budget, PSR secured approximately BRL 12.5 billion in agricultural production (Wedekin, 2019). From the tax standpoint, is important to assure that the public resources destined to the insurance operation, and financed by the taxpayers, are correctly applied in the underwriting as well as in the pricing.

The usual model of crop insurance consists in reimburse the producer in crop shortfall years, covering the difference between the harvested financial amount and the coverage purchased. In order to calculate the premiums, a regression model is estimated for the historical yields and from its residual values, the likely claims are calculated. Thus, the methodological treatments to these residuals are fundamental aspect to the pricing process and deserve attention.

Two primaries heteroskedasticity assumptions have been maintained in the area-yield insurance literature (Harri et. al, 2011): (i) that these residuals are homoskedastic (the variance of the series is constant), and (ii) that the changes in the yield standard deviation are proportional to the changes in the yield, keeping the coefficient of variation constant; the latter is called *proportional heteroskedasticity*. Although these are extensively used in international literature (Deng et al., 2007; Ker & Coble, 2003; Miranda, 1991), and more recently in Brazil (Duarte et al., 2018), there are several papers suggesting the heteroskedasticity is a structure present in the historical yield series and it varies with the region (Harri et al., 2011; Just & Pope, 1978).

Once there is no agreement about the residual's variability pattern and given the economic relevance of this subject in Brazil, our main objective in this paper is to estimate the residuals behavior in function of the yields. This approach is different from the literature once it does not suppose any previous variance structure, as in previous papers (Deng et al., 2007; Ker & Coble, 2003; Miranda, 1991). In order to evaluate the robustness, we evaluate premiums from different methodologies.

2. Theoretical background

In this section we present the evolution of the crop insurance pricing methods. Initially, an outlook of the Brazilian crop insurance market and how the government agricultural risk management programs work. Following, how the literature treats the crop yield insurance contracts and particularly claims. Finally, we present the main methodological aspects on the time series treatment.

2.1. Gerenciamento social de risco agrícola no Brasil

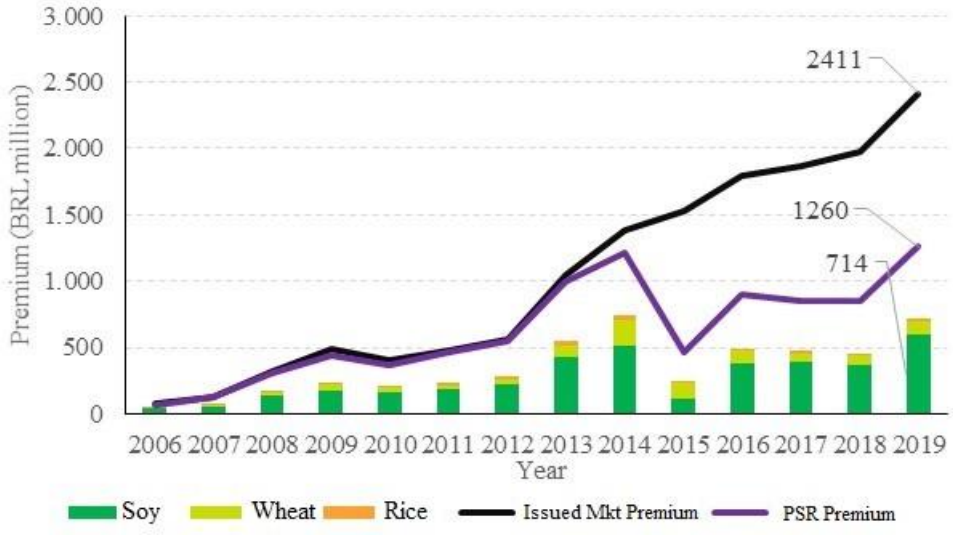
There are two main risk management programs in activity in the country, *Programa de Subvenção ao Prêmio de Seguro Rural* (PSR) and *Programa de Garantia da Atividade Agropecuária* (Proagro), both agriculture support programs financed with government budgetary resources. There is no participation of private companies in Proagro.

Proagro is not an insurance mechanism and was the only way of protecting financial contracts available to producers against eventual losses caused by climate hazard events. The Proagro hiring in made by farmers with the agents of Proagro (banks or credit unions) directly in the costing financing contracts. The farmer pays a fee called "additional", a percentage of the total value to be charged via Proagro (BACEN, 2020).

Although the risk mitigation alternatives adopted by the country since 1954 (with the *Rural Insurance Stability Fund* -FESR- creation), only in the last ten years the rural insurance industry started to develop (Wedekin, 2019). Since PSR creation, in 2003, this market expanded quickly. In 2006, BRL 88 million was issued in premiums, while in 2018 the issued premium surpassed BRL 2 billion.

Figure 1 shows that between 2006-2014 start there is a growing trend in the rural insurance market, sharing the same pattern as PSR. PSR presents stagnation in 2015, when an expressive reduction happens (over 50%). Finally, the soybean crop is the most expressive in the program, receiving more than half of PSR premiums in history.

Figure 1 – Time evolutions of the crop insurance premium (in BRL million), PSR and soybean, wheat and rice crops, between 2006-2019



Source: SES/Susep and MAPA.

As presented in Figure 1, the rural insurance private market expanded by 63 times in the period. Ozaki (2010) shows the South region concentrates most of the resources offered by the government. Besides that, Ozaki points the lack of quality and quantity of the information in the public databases, which could help to improve public politics for the rural sector.

De Medeiros (2013) evaluates PSR as positive for the federal government, once the hiring of an insurance policy transfers the rural activity risks for the private market, relieving National Treasury from the constant debt renegotiation with the farmers, which happened in Proagro.

2.2. Crop insurance contracts pricing

As in every insurance contract, the actuarially fair premium should reflect the future claims expectations, and to form it, the average of past experiences (Bowers et al., 1997). Specifically, over the premiums of this sector in Brazil, the individual farmer historical yield or the municipal agricultural production, surveyed by the Brazilian Institute of Geography and Statistics (IBGE) in the Municipal Agricultural Production (PAM) research. The periodicity of this research is yearly and it covers the whole country (Instituto Brasileiro de Geografia e Estatística, 2018).

Halcrow (1949) was the precursor in the proposition of actuarial structures for pricing crop yield insurance, assuming the residuals distribution should follow the Normal distribution. The interest for the pricing of this kind of contract comes, in part, from the U.S. Congress, which in 1938 introduced several measures of rural politics, including the risk management of the farms, the *Farm Bill* (Wedekin, 2019). Before this, the crop insurance was offered exclusively by private insurance companies in the country.

The base for pricing this insurance structure is the county yield, the sum of the whole production in the county over the planted area (*area-yield*). The rationale to use the county

yields rather than the individual has the intention of preventing the adverse selection, problem raised by Halcrow (1949) and reinforced by Skees & Reed (1986). Once the farmers have more information about your own yield than the insurer, only the farmers that have more claims to receive than premiums to pay in the long run will hire the coverage. The insurer, experiencing more claims, assumes the insurance contracts rate should be greater, attracting more and more adverse risks, generating the anti-selection spiral (Miranda, 1991).

Skees, Black, & Barnett (1977) expanded the crop insurance pricing methodology to the regional yield basis. The model was developed to price the Group Risk Plan (GRP) insurance contracts, developed by the United States Department of Agriculture (USDA), responsible for the crop insurance pricing in the counties. This model consists into estimating a regression model for the historical yield and, using the residuals of this model, estimate the losses based on the historical losses (loss-cost method), or estimate a distribution for these residuals. Both techniques are recurrently used in the literature. Recently, both methods were refined, loss-cost (Harri et al., 2011; Ker & Tolhurst, 2019) and estimating residuals probability distribution (Duarte et al., 2018; Xiao et al., 2017) in order to improve the premium calculation process.

Part of the calculation algorithm proposed by Skees et al. (1997) requires an estimation for the yield based on the available information. In Brazil, IBGE discloses Harvest statistics since 1974 through PAM.

Historically, crop yields have been growing throughout the world. There is an agreement that the main factor for this growing is the agricultural technological development (Ker & Coble, 2003). Therefore, the yield evolution makes the yield time series present positive slope over time. According to Ye, Nie, Wang, Shi, & Wang (2015), weather is more accountable for the pattern presence in the residuals, when they are not purely random (white noise).

Several trend models have been used to describe the data behavior. Among the deterministic models, is presented the *splines* model, which is used by RMA in its policies. Just & Weninger (1999) used polynomial models, while Deng et al. (2007) used the log-linear model. Yet in stochastic models, Ozaki & Silva (2009) used an autoregressive model to fit the data, despite that Harri et al. (2009) have presented limitations in the usage of stochastic models for the yield time series.

Ye et al. (2015) evaluated the performance of the main regression models (linear, log-linear, ARIMA, pure MA, Savitzky-Golay, exponential, non-parametric Lowess) used in the literature to detrend the series. The authors pointed to statistically significant differences between one model or another. Yet, in average, ARIMA models tend to generate higher premium rate than others.

2.4. Yield series heteroskedasticity

In literature there are two main assumptions over the residuals and the form of the heteroskedasticity. The first admits these residuals to be homoskedastic, i.e., its variance is independent of the historical yield. Coble, Heifner, & Zuniga (2000) and Miranda (1991) found evidence that the series are homoskedastic and used this assumption to model the crop yields. In opposition, as second assumption, the standard deviation of these residuals increases proportionally to the yield. This form of heteroskedasticity is denominated *proportional*, and was also verified in several papers (Ker & Coble, 2003; Skees et al., 1997).

Harri et al. (2011) developed a methodology to estimate the heteroskedasticity level of the yield series. The study was presented as a back test to the methodology used by RMA, which tacitly supposed the residuals have a proportional heteroskedasticity. Authors found evidence that homoskedasticity nor proportional heteroskedasticity should be accepted without previous analysis. RMA adopted Harri et al. (2011) recommendations for the heteroskedasticity analysis in their pricing.

2.5. Residuals probability distribution

The probabilistic distribution of the residuals has been another topic for debate in the literature. Due to the reduced number of yield observations (usually historical data has 30 years in average), any statistical analysis over this data becomes more complex (Ozaki et al., 2008). Several ways to identify these patterns were, and still are, proposed. Some of the used techniques are parametric distributions, non-parametric kernel distributions, and more recently, probability distributions non-parametric generated by computational simulations *(bootstrap).*

Botts & Boles (1958) lifted the first assumption, that the yield should follow the Normal distribution. Day (1965), on the other hand, found evidence suggesting non-normality in the residuals. Atwood, Shaik & Watts (2003) and Gallagher (1987) found evidence the yield could follow, respectively, distributions Gama e Logistic, both skewed and kurtotic. Skees et al. (1997) applied a beta distribution for skewness and heavy left-tail. In Brazil, Duarte et al. (2018) modelled the soybean yields in Parana using a bimodal distribution *(odd-log-logistic-F).*

About non-parametric distributions, Goodwin & Ker (1998) used the non-parametric kernel estimators to approach the yield time series behavior. Unlike the parametric, non-parametric estimation is a set of techniques used to estimate the residuals distribution without making any assumptions over the data shape (Altman, 1992). Thereby, the estimated curve shape is revealed by its own observations (Ozaki et al., 2008).

3. Data and methods

3.1 Dataset treatment

We used data from the *Sistema IBGE de Recuperação Automática* platform (SIDRA) from IBGE. We also used PAM data, which is yearly updated and discloses information about planted area and produced amount of temporary and permanent crops.

In order to model the yield time series, planted area, and produced amount, we gathered data for three crops in municipal aggregation level: soybeans, rice and wheat. These crops figure between the seven biggest crops in PSR, being soybeans the largest, with 42.8% of the premium itself, being followed by winter corn with 21.6%. Rice and wheat hold respectively 8% and 1.9%. In long run, these percentages have low fluctuation. This data is available through SIDRA platform, which contains data from 1947 to 2020 (47 years).

The average production yield in year t ($\boldsymbol{Yt}$) is defined as the quotient between produced amount (measured in tons) and planted area (in acres).

$$\boldsymbol{Y_t = \frac{Produced\ amount\ in\ year\ t}{Planted\ area\ in\ year\ t}}, \qquad \mathbf{t} \in \mathbf{1, 2, \ldots, T} \qquad (\mathbf{1})$$

Once the planted area variable started being registered in 1988, all of time series used begin in this year. Although the soybean cultivation is a yearly event, which occurs with no interruptions during this whole period, there are observations disclosed by IBGE as *non-available values*. IBGE exemplifies these as the following: "*Beans production in certain municipal was not researched or this county did not exist in the research year*". When information is omitted, the gap will be replaced by the average between every other observation for other counties weighted by the inverse of the distance between the two counties. Thus, the omitted value is replaced by the following relation:

$$\boldsymbol{Y_i^* = \sum_{k=1}^{n} \frac{Y_k}{\omega_{i,k}}} \qquad (\mathbf{2})$$

where $\boldsymbol{Y_i^*}$ is the missing observation for the county i, $\boldsymbol{Y_k}$ is the yield observation for the county k (used to replace i) and $\boldsymbol{\omega_{i,k}}$ is the weight given to the observation to county k, according to the distance between the counties. Thus, this parameter is defined as:

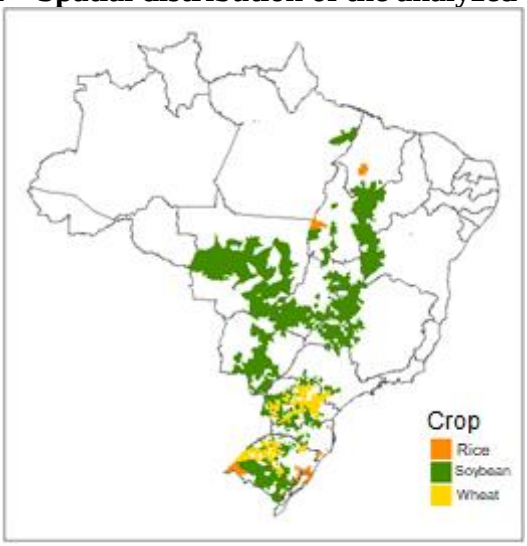
$$\boldsymbol{\omega_{i,k} = \frac{d_{i,k}}{\sum_{k=1}^{n} d_{i,k}}} \qquad (\mathbf{3})$$

where $\boldsymbol{d_{i,k}}$,the distance between i and k is measured by the latitude and longitude distances, which are equivalent to triangle sides and the theoretical distance (in a straight line) equals the hypotenuse.

$$\boldsymbol{d_{i,k} = \sqrt{(lat_i - lat_k)^2 + (long_i - long_k)^2}} \quad (\mathbf{4})$$

So, the counties closer to the missing data county have a larger influence in the missing value calculation. Initially, the three crops' analysis was restricted to the counties part of *edaphoclimatic* regions (also known as *soybean regions*). This delimitation was made based on Agricultural Zoning for Climatic Risk from MAPA for 2011 harvest. According to MAPA, these 3.578 counties present soil and weather appropriate conditions for soybeans cultivation. Based on these counties, we removed every county which have average planted area under 10.000 acres between 2014 and 2018, in order to prevent the influence of big producers over the expected yields for that county. Figure 2 presents the counties which will be analyzed in this study.

Figure 2 – Spatial distribution of the analyzed counties



Source: own elaboration, based in the selection algorithm.

3.2 Non-parametric trend models (*kernel*)

In literature, several authors explored different trend models to work with crop yield data, and usually interested in explaining benefits and differences among them (Goodwin & Ker, 1998; Harri et al., 2009; Ker & Coble, 2003; Ye et al., 2015). The consensus that crop yields, generally, has grown throughout the world due to the technological development in farms (Shao et al., 2010; Zeng et al., 2014). However, there are no answer yet if this growth is really caused by anthropogenic factors or it is an outcome of natural events.

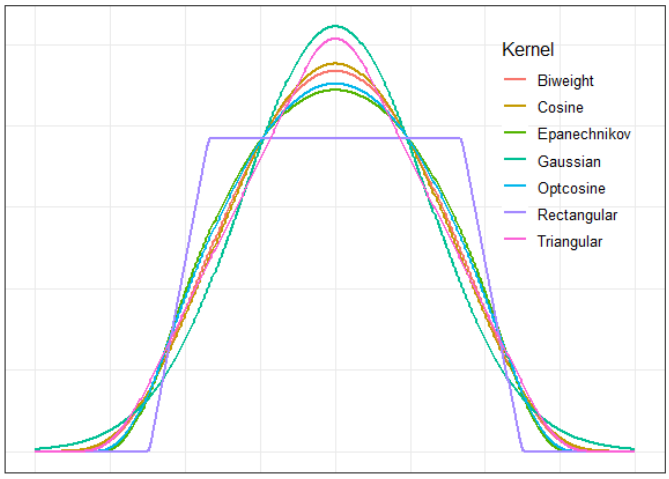
Thus, this is the reason why it is necessary to adopt models to transform these time series into stationary and with no autocorrelation. Before the choice between deterministic and stochastic models, Harri et al. (2009) found weaknesses in stochastic models for this yield data treatment. This way, we are going to use a *kernel* trend model. A *kernel* function is defined as non-negative, defined in real numbers set, and integrable. In addition, a *kernel* function also can be understood as a "window" functions, once it assumes a real and positive value between certain minimum and maximum values of the domain (inside its window)

For the trend model, we used a *Gaussian kernel* as interpolation method between two points, exactly the most used in literature. The Gaussian kernel is a good start point in kernel choosing process, because once the data distribution is not known, the points are interpolated in a smooth way, using a Normal distribution. The definition is:

$$\boldsymbol{K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2}} \quad (\mathbf{5})$$

Figure 3 illustrates the interpolation between two points using different kernel types.

Figure 3 – Comparison of different kernel functions interpolating two points



Source: own elaboration.

The main objective of our paper is to find the residuals of the functional relation between yield and time, conditioned to the latter. Thus, the conditional expected values between the two variables are defined as:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{m}(\boldsymbol{X}) = \int_{-\infty}^{+\infty} \boldsymbol{y} \boldsymbol{dF}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}) = \int_{-\infty}^{+\infty} \boldsymbol{y} \boldsymbol{f}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}) \boldsymbol{dy} \quad (6)$$

where $\boldsymbol{f}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y})$is the conditional distribution for Y given X. In order to model the trend, we used the Nadaraya-Watson estimation. This was the same method used by Racine & Li (2004) and it is based on local averages, denominated *m*, using a kernel as weighting function. This estimator is defined as:

$$\widehat{\boldsymbol{m}}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{K}_{\boldsymbol{h}}(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{i}}) \boldsymbol{y}_{\boldsymbol{i}}}{\sum_{\boldsymbol{j}=1}^{\boldsymbol{n}} \boldsymbol{K}_{\boldsymbol{h}}(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{j}})}, \quad (7)$$

where $\boldsymbol{K}_{\boldsymbol{h}}$ is a kernel with determined bandwidth value equal to *h*. The bandwidth value is a smoothing parameter for the interpolation between two points, which has strong influence over the results. The bandwidth *h* kernel function can be defined as:

$$\boldsymbol{K}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{\boldsymbol{1}}{\boldsymbol{h}} \boldsymbol{K}\left(\frac{\boldsymbol{x}}{\boldsymbol{h}}\right) \quad (8)$$

The bandwidth value should be chosen in order to minimize MSE (minimum square error). Using a larger bandwidth value could hide the real data structure, joining many observations inside the same band. On the other hand, using a small bandwidth value causes an under smooth problem because it contains only a few observations and the trend model tends to reproduce the original series.

3.2.1 Choosing the ideal smoothing parameter value

The objective of the bandwidth choice is to ensure the estimator $\widehat{\boldsymbol{m}_{\boldsymbol{h}}}(\boldsymbol{x})$ (Equation 7) is the one which better describes the Y conditional expectation given X (Equation 6). A way to solve this problem, following Martin et al. (2012), is by minimizing the difference between the observed $\boldsymbol{y}_{\boldsymbol{t}}$ and its expected value $\boldsymbol{m}(\boldsymbol{x}_{\boldsymbol{t}})$ by SSE method (sum of squared errors) given by:

$$\boldsymbol{arg}_{\boldsymbol{h}}\boldsymbol{min}\left(\boldsymbol{S} = \sum_{\boldsymbol{t}=1}^{\boldsymbol{T}} (\boldsymbol{y}_{\boldsymbol{t}} - \widehat{\boldsymbol{m}_{\boldsymbol{h}}}(\boldsymbol{x}_{\boldsymbol{t}}))^{\boldsymbol{2}}\right) \quad (9)$$

The problem in choosing the estimator which minimizes the squared error is that the optimal solution occurs when bandwidth equals to zero (Equation 10). Yet, this is not ideal once the smoothing would make the trend model identical to the original series.

$$\lim_{h \to 0} S = 0 \quad (10)$$

Thus, a way to contour this problem, still following Martin et al., (2012), is through the *leave-one-out approach* algorithm. In this procedure, the j-th observation is removed for the estimator calculation of the jth realization $\boldsymbol{m}(\boldsymbol{x_j})$. The kernel estimator found using the *leave-one-out* approach is defined by:

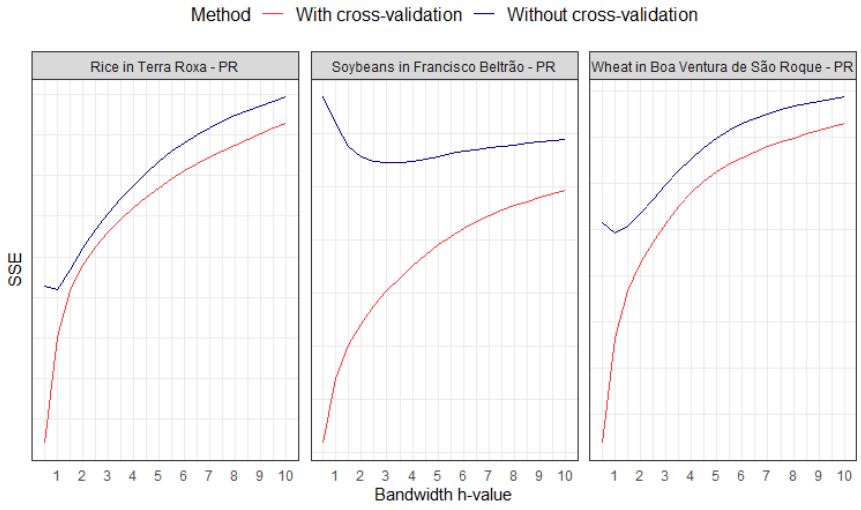
$$\tilde{m}_h(x_j) = \frac{\frac{1}{Th}\sum_{\substack{t=1 \\ t \neq j}}^{T} y_j K\left(\frac{x_j - x_t}{h}\right)}{\frac{1}{Th}\sum_{\substack{t=1 \\ t \neq j}}^{T} y_j \left(\frac{x_j - x_t}{h}\right)} \quad (11)$$

Therefore, the optimization problem to be solved is:

$$arg_h min \left(\tilde{S} = \sum_{t=1}^{T} (y_t - \tilde{m}(x_t))^2 \right) \quad (12)$$

Figure 4 presents the sum of squared errors (SSE) of the models in function of *h* bandwidth parameter for three crops in Parana State: (i) rice, in Terra Roxa county, (ii) soybeans in Francisco Beltrao county, and (iii) wheat in Boa Ventura de Sao Roque county. In every one, the cross-validation process was used. The blue lines show that using the *cross-validation* process there is a well-defined minimum point. On the other hand, the red line show the relation between bandwidth and SSE, without this procedure is monotonically growing, taking the minimum SSE value to h = 0, as expected.

Figure 4 – difference in squared error in function of bandwidth with (blue line) and without (red line) cross-validation process for Parana counties



Source: own elaboration.

3.3 Heteroskedasticity assumptions

From the detrended data, generated through the kernel trend model (specifically, Nadaraya-Watson estimator), we did the Harri et al. (2011) test to evaluate the residuals' behavior, whose methodology is presented as following. The heteroskedasticity modeling is given in function of the estimated yield. This will allow us to test several assumptions over heteroskedasticity presented in literature. These assumptions over the heteroskedasticity structure can be represented by the following relation:

$$Var(e_t) = \sigma^2[\mathbb{E}(y_t)]^{\beta} = \sigma^2 \hat{y}_t^{\beta} \quad (13)$$

where $Var(e_t)$ is the variance of the error term between the trend model and observed value. From the β exponent, each assumption presented in Table 1 can be tested.

Table 1 – Residuals behavior pattern in function of exponent $\boldsymbol{\beta}$.

$\boldsymbol{\beta}$	**Residuals behavior**
0	Residuals have no variance (homoskedasticity)
1	Yield variance is directly proportional to yield expected value
2	Standard deviation of yields in directly proportional to the expected value of yields (coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ is constant).

Source: elaborated by authors

In order to empirically estimate $\boldsymbol{\beta}$ value the following linear trend model is followed:

$$\boldsymbol{ln(\hat{e}_t^2) = \alpha + \beta ln(\hat{y}_t) + \varepsilon_{t,}} \quad (\mathbf{14})$$

where $\hat{\boldsymbol{e}}_{\boldsymbol{t}}$ is the difference between the found residual and the expected value modelled by Equation 11 in t, and $\hat{\boldsymbol{y}}$ is the estimated yield value. In order to detrend the series (required step for insurance rates calculation, once yields are growing), residuals are adjusted as the following:

$$\hat{\boldsymbol{y}}_{\boldsymbol{t}}^{*} = \hat{\boldsymbol{y}}_{\boldsymbol{T+1}} + \hat{\boldsymbol{\epsilon}}_{\boldsymbol{t}} \left(\frac{\hat{\boldsymbol{y}}_{\boldsymbol{T+1}}}{\hat{\boldsymbol{y}}_{\boldsymbol{t}}}\right)^{\frac{\hat{\boldsymbol{\beta}}}{\mathbf{2}}} \quad (\mathbf{15})$$

One could notice that Equation (15) is different from Equation (4) of Harri et al. (2011) in two points. The first, the forecasted value used is T+1 instead of T+2 used in Harri et al. (2011). The purpose is to keep conformity with RMA procedures. The second is about an inconsistency found by Ker & Tolhurst (2019). The exponent value should be $\frac{\boldsymbol{\beta}}{\mathbf{2}}$ for consistency instead of $\boldsymbol{\beta}$ presented by Harri et al. (2011).

For the insurance contract pricing, there would be made a comparison between the three obtained rates by the three-time series, the one estimated by homoskedastic residuals, the one estimated by proportionally heteroskedastic residuals, and the one where they are empirically estimated. Besides that, through a hypothesis test, we will verify which counties could accept the homoskedasticity ($\boldsymbol{\beta = 0}$) and proportional heteroskedasticity ($\boldsymbol{\beta = 2}$) assumptions. This examination will be implemented by a hypothesis test to obtain the confidence interval for the parameter $\boldsymbol{\beta}$, estimated by Equation (14).

With adjusted residuals (by Equation 15), we obtain the detrend time series. In order to check stationarity and autocorrelation, tests were conducted. From this corrected series, pricing procedures to calculate insurance rates will be performed through the *Loss-Cost* method.

3.4 *Loss-Cost* pricing

In the *Loss-Cost* method, the claims ($\boldsymbol{indem_t}$) paid and the insurance rated are obtained by the following:

$$\boldsymbol{indem_t = max\left[\frac{Y_c - y_t^*}{Y_c}(\hat{y}_{T+1}), 0\right]} \quad (\mathbf{16})$$

where

$$\boldsymbol{Y_c = \hat{Y}_{T+1} * cov} \quad (\mathbf{17})$$

$$\boldsymbol{Y_t^* = \hat{Y}_{T+1} + e\_adj_t} \quad (\mathbf{18})$$

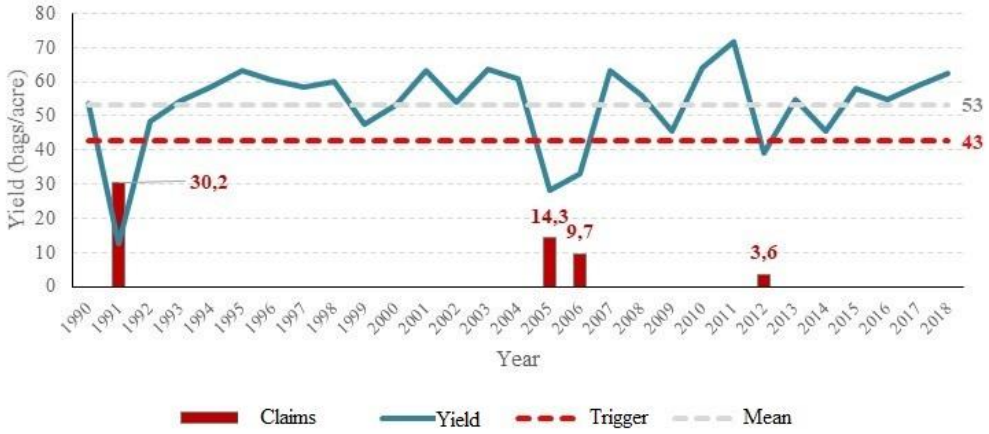
$\boldsymbol{cov}$ is the coverage level, $\boldsymbol{Y_c}$ is the payout trigger, $\hat{\boldsymbol{Y}}_{\boldsymbol{T+1}}$ expected yield for the next year ($\boldsymbol{T+1}$) increased by the factor derived from Equation (15).

The insurance rate obtained by the Loss-Cost method (used in GRP) is calculated by:

$$\boldsymbol{r = \frac{\mathbb{E}(indem_t)}{Y_c}} \quad (\mathbf{19})$$

Figure 5 illustrates the method by an example of historical soybean in Francisco Beltrao, Parana county, showing the historical shortfalls and how claims are estimated, impacting in the insurance rate $\boldsymbol{r}$. In the example, $\boldsymbol{r}$ is calculated as the mean of the four historical losses (in bags/acres), which happened in 1991, 2005, 2006 and 2012, throughout 29 years.

Figure 5 –Evolution of soybean yields in Francisco Beltrao (PR), trigger and claims



Source: own elaboration.

In this example, considering a trigger (in bags/acre) $\boldsymbol{Yc = 43}$ and the 25 years, where no shortfall is observed, the crop yield insurance rate (**r**) is calculated by:

$$\boldsymbol{r = \frac{\frac{30,2 + 14,3 + 9,7 + 3,6 + 25 \times 0}{29}}{43} = 0,04635}$$

4. Empirical results

4.1 Data source – Programa de Subvenção ao Prêmio do Seguro Rural (PSR)

In order to operationalize this methodology, we present the premium, subsidy and sum insured crop insurance in Brazil for the three selected crops: rice, soybeans and wheat. The data source for the study is PSR, once this data is available in an analytical level, which is enough for our analysis. Every issued policy is registered in PSR, which is after made available by MAPA. Another data source would be the SUSEP statistical system (SES). However, once the source of this information is accounting, there is no detailed geographic information, what would make our analysis impossible.

Thereby, Table 2 presents the premium amount paid by farmers and subsidized for the three interest crops in 2018. This year was chosen because while we were writing this paper, the 2019 information was not fully available in PSR.

Initially, our database had 1,876 counties. However, once some of the counties did not reached our two minimum requirements (planted area and data gaps), 1,056 counties which received resources from PSR are out of this study. In addition, there are 80 counties which reached our requirements but had not received any resources from MAPA in 2018.

Therefore, the analysis will be made over the intersection between the counties lists. In conclusion, the analysis has three and not two selection criteria, as shown in Table 2. Consequently, the premium amount analyzed is smaller than compared to the MAPA official information. This difference is evidenced by Figure 6.

Table 2 – Crop insurance premium (in BRL million) in 2018 for PSR crops and premium considered in analysis, by the counties which met the criteria.

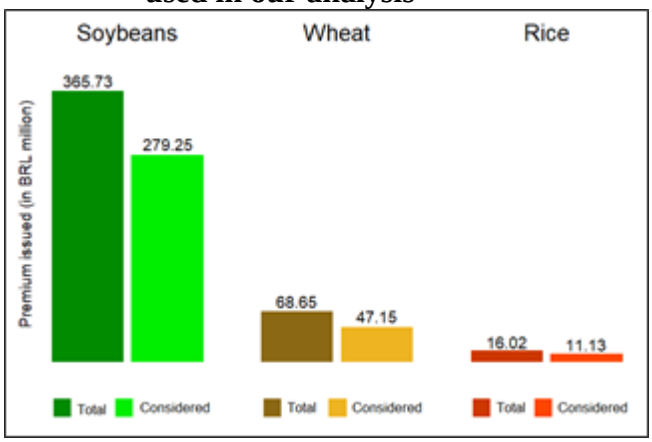
Crop	State	Premium PSR	Premium Analysis	Number of counties - PSR	Number of counties - analysis
Rice	MA	0,24	0,13	3	1
	RS	10,46	7,55	83	37
	SC	5,27	3,42	39	8
	SP	0,03	0,00	4	0
	TO	0,02	0,02	1	1
Soybeans	BA	9,29	7,98	9	7
	GO	42,62	34,33	136	51
	MA	4,26	3,47	22	8
	MG	7,16	5,06	83	31
	MS	49,70	40,76	58	36

	MT	32,29	25,43	88	46
	PA	0,27	0,15	5	1
	PI	2,61	1,62	13	5
	PR	116,40	93,99	348	182
	RO	0,39	0,00	3	0
	RS	57,76	46,77	230	119
	SC	4,74	3,15	63	16
	SP	30,77	13,94	191	21
	TO	7,47	2,60	49	9
Wheat	MG	0,16	0,00	10	0
	MS	0,22	0,00	3	0
	PR	33,31	25,33	221	76
	RS	22,26	14,64	154	42
	SC	0,71	0,00	24	0
	SP	12,00	7,18	36	3
Total		**450,40**	**337,52**	**1.876**	**700**

Source: elaborated by authors.

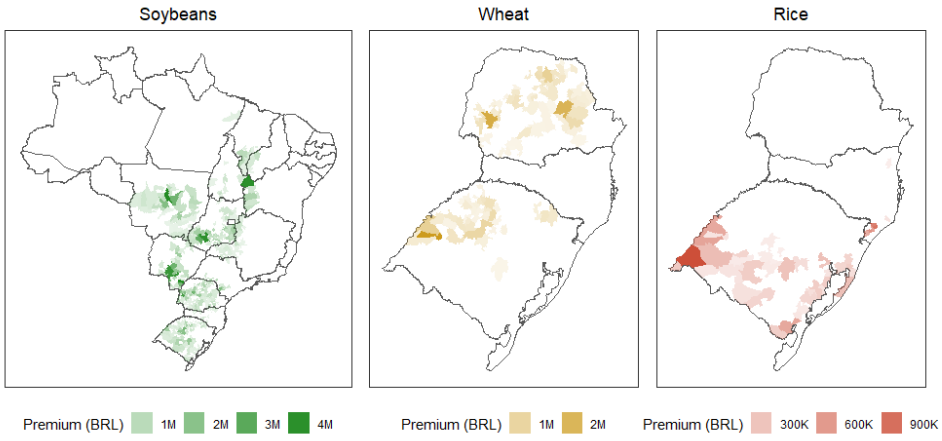
We can notice a relevant concentration for rice and wheat crops in South region. Thus, for these crops Figure 5 focuses on this region in order to evaluate the premium concentration. We also noticed relevant concentration for the soybean crop premiums issued in 2018.

Figure 6 – Comparison between premium issued in PSR for the crops and value used in our analysis



Source: own elaboration.

Figure 7 – Premiums issued by crop in PSR in 2018 on the analysis counties



Source: own elaboration.

4.2 Methodology effectivity

In order to reach the main objective of this study, we will evaluate the main assumptions regarding the variance pattern of the crop yield time series. The answer to the following question is ought: in front of the residuals conditioned to the observed yield, given a set confidence level of 95%, is it possible to accept the homoskedasticity and proportional

heteroskedasticity assumptions? Table 3 presents, by crop, the number of counties where is not possible to reject one or both assumptions.

Table 3 – Results for the hypothesis test for the residuals by crop

Crop	Assumption	Accepts	Rejects	Acceptance Rate
Rice	Homoskedasticity	42	5	11%
	Proportional Heteroskedasticity	37	10	21%
	Both	36	4	10%
Soybeans	Homoskedasticity	482	50	9%
	Proportional Heteroskedasticity	436	96	18%
	Both	426	40	9%
Wheat	Homoskedasticity	111	10	8%
	Proportional Heteroskedasticity	116	5	4%
	Both	109	3	3%

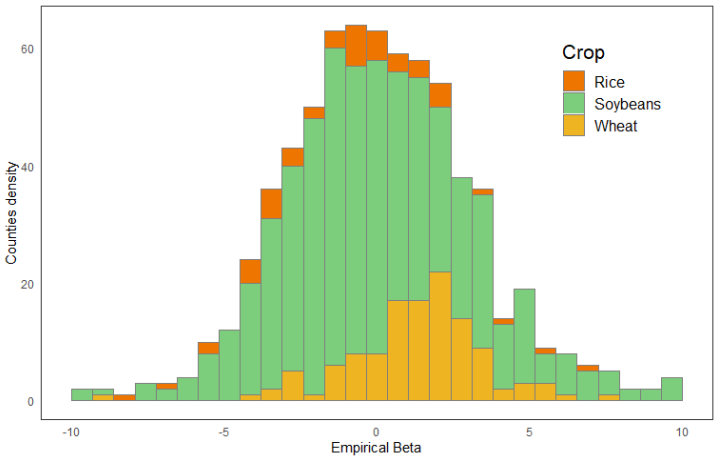
Source: own elaboration

These results are comparable to that found by Harri et al. (2011) in the common crop in our paper and theirs: soybeans. In Harri et al. (2011), the proportional heteroskedasticity assumption was rejected in 26% counties, while the homoskedasticity assumptions was widely rejected: approximately 60%. In this study, these numbers were respectively 18% and 9%.

For soybeans as well as the other two crops, the rejection rate for both assumptions can be considered low. Therefore, in most counties, one can adopt an assumption for the residuals pattern. Thus, there is evidence that using one or another assumption in the crop insurance pricing directly impacts the outcome. Later, we show that even for those counties, the final premium is quite sensitive to this choice, what could distort the pricing.

Plus, for both assumptions, three crops presented an acceptance rate between 80% and 90%. Most counties have its $\widehat{\boldsymbol{\beta}}$ parameter estimated between 0 and 2 (implicit values in homoskedasticity and proportional heteroskedasticity assumptions) as shown in Figure 8.

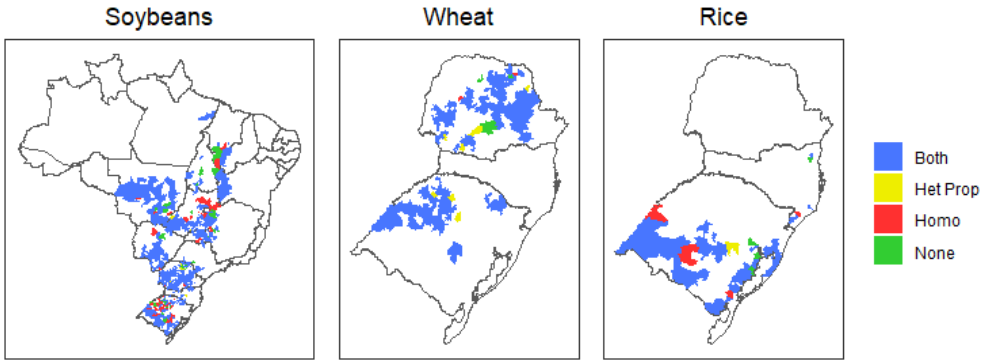
Figure 8 – $\widehat{\boldsymbol{\beta}}$ coefficient distribution for the crops



Source: own elaboration.

Duarte et al. (2018) empirically verified the homoskedasticity assumption for several Parana counties, aiming develop soybean yield forecasts using a probability distribution. One of the assumptions for this forecast model is that historical productivities are homoskedastic. In our paper, similar results were found for the analyzed counties: Cascavel ($\boldsymbol{IC(\beta) = [-9,11; 3,55]}$) Guarapuava ($\boldsymbol{IC(\beta) = [-8,69; 13,03]}$) and Castro ($\boldsymbol{IC(\beta) = [-5,06; 15,01]}$). We highlight, however, that Duarte et al. (2018) evaluated the homoskedasticity through the Breusch-Pagan test. Therefore, by distinct methodology we could verify the same assumptions, strengthening the results robustness. However, due to the confidence intervals range, is also possible to claim that the proportional heteroskedasticity assumption could be adopted once these confidence intervals contains the value 2.

Figure 9 – Geographical distribution of volatility assumptions acceptance



Source: own elaboration.

On the other hand, through Figure 9 we can verify the presence of a few counties where both assumptions are rejected (highlighted in green). Now these counties are well-defined, we will evaluate the potential premium difference in adopting the proportional heteroskedasticity assumption and empirically estimating these residuals. This comparison between the two procedures replicates the analysis of Harri et al. (2011), which was presented as counterproof to the actual RMA methodology (proportional heteroskedasticity).

4.3 PSR results

Based on the three methodologies, actuarially fair insurance rates were estimated for every county and both crops through the loss-cost method. Therefore, for a county which received resources for the three crops, there will be calculated nine insurance rates (three methodologies for three crops each). The insurance rate was estimated by the mean payment (Equation 19). This rate was calculated for every coverage level between 60% and 80% (coverage levels used in crop yield insurance) and, from these values, were obtained the mean rates, once again, for every crop, methodology and county.

With these mean rates and based on the issued premium for the crops in the counties, the potential under or overpayment will be estimated. In the analysis, the issued premiums obtained will be under the proportional heteroskedasticity assumptions and are compared to the premiums derived from the empirical estimation for the residuals' volatility. Some outlier counties where the results were outliers (empirical/proportional rate) were removed from the analysis. We considered as outliers every county where the mean rate calculated by one method is 10 times greater than the second one, i.e., the quotient between them is greater than 10. For these outliers, the differences between the two procedures were not considered.

Table 4 – Premium differences regarding different methodologies application

Crop	State	Number of Counties	Counties Premium (*) (A)	Counties Premium: empirical (*) (B)	Difference (*) (B)-(A)	Underpayment	Overpayment (*)
Rice	MA	1	131	19	-112	-112	0
	SC	4	1.685	3.401	1.716	-586	2.302
	RS	23	5.453	924	-4.530	-4.538	8
Soybeans	PA	1	148	4	-144	-144	0
	TO	6	1.321	205	-1.116	-1.116	0
	MA	7	3.335	1.865	-1.470	-1.804	335
	PI	5	1.622	243	-1.379	-1.379	0
	BA	7	7.979	17.171	9.192	-1.817	11.008
	MG	16	1.661	2.558	897	-732	1.630
	SP	15	10.784	15.435	4.652	-2.388	7.040
	PR	142	77.447	94.507	17.060	-33.067	50.127

	SC	10	2.400	854	-1.546	-1.546	0
	RS	79	28.586	7.607	-20.979	-21.706	727
	MS	25	27.129	16.109	-11.020	-18.469	7.449
	MT	28	17.469	4.217	-13.252	-15.683	2.431
	GO	23	14.675	1.180	-13.495	-13.495	0
	SP	3	7.182	417	-6.765	-6.765	0
Wheat	PR	72	24.347	19.096	-5.251	-13.816	8.565
	RS	41	14.142	6.708	-7.435	-7.837	402
	Total	**508**	**247.496**	**192.519**	**-54.978**	**-147.001**	**92.023**

(*) Financial amount in BRL thousand Source: own elaboration.

Table 4 presents the analysis results made in county-level, then aggregated to state-level for the crops. For example, taking the difference for wheat crop in Parana state, 72 counties were evaluated where premiums were issued with subsidies (column A) to an amount of BRL 24.35 million. In this amount, implicitly is assumed the proportional heteroskedasticity assumptions was used in the pricing. For the same counties, the insurance rate calculated by the empirical heteroskedasticity (column B) was realized, totaling BRL 19.1 million.

In aggregate, the empirical volatility evaluation developed a historical time series with less accentuated shortfalls and, consequently, less claims in state-level. This result is opposite to the found for soybean crop in Parana, showing the empirical heteroskedasticity effects can rise the insurance rates, whose results are presented in the last two right columns. Table 4 presents the decreases totaled BRL 33.1 million and the increases, BRL 50.1 million. Therefore, the net aggregate is result is an increase of BRL 17 million.

The results suggest the assumptions widely influence the results. The biggest difference between empirical and proportional rates occurred in soybean crop, in Bahia State (115% difference). The lower difference happened also in soybeans, in Parana State (22% difference). In the overall result, this variation was negative in 22%.

In conclusion, differently from the U.S. case, the analyzed Brazilian crops have a difficult in reject the stablished assumptions in literature, causing rates variation in function of the assumptions' choice. One of the reasons for this difference is the size of the historical series used. Harri et al. (2011) used the National Agriculture Statistics Service (NASS) series, which contain data from 1955 to 2019. The power of the hypothesis test is directly related to the size of sample used. For the evaluation based in a shorter historical series, is expected that the confidence interval range, for the same confidence level, would be wider, and consequently, the rejection rate for these hypotheses decreases. A way to contour this problem would be the empirical estimation of probability distributions for these residuals as an additional way to refine the pricing of the crop yield insurance.

5 Final Remarks

In this paper we analyzed the residuals' pattern in function of the historical crop yield series, bringing evidence about the series' volatility. Evaluations like these are important once not only agricultural public politics are economically relevant in Brazil, but also the private rural insurance market presented a tremendous expansion in last 20 years.

Based on the results, we could verify the assumptions adoption over residuals volatility for crop yield time series influence crop yield insurance premiums calculated through the loss-cost method. In general, one can stand that adopting one of the two discussed assumption tends to, in average, increase the empirical crop insurance rate. Yet, a case-to-case verification is recommended, once Parana, the biggest market in PSR, regarding the soybean crop presented empirical rates higher than proportional ones.

Wariness is advised on results reading by the following reasons: the issued premium values presented regarding the resources made available through the PSR program, not the whole market. This was defined in order to quantify these insurance premium rates variation

regarding different pricing strategies. Implicitly, one should take on the assumption that these contracts were priced under one or another assumption. In practice, every insurance company has a different market strategy and own underwriting process. In this question, this paper differs from Harri et al. (2011), once in the U.S., RMA insures and subsidies part of the crop yield insurance premium, making these pricings more comparable to the county-level. In addition, results are substantially influenced by the data quality: in the face of lack of data, the assumptions adoption is needed, which clearly influences the results.

Results bring a Brazilian rural insurance market outlook, that regardless its expansion, always had small area penetration in country-level. One of the main reasons is the appraisal that insurance premiums charged by insurers are not compatible to farmers risk (Duarte et al., 2018), generating the anti-selection spiral. In our paper, 80% of the evaluated counties presented lower premium rates for the empirical estimation against proportional estimation, suggesting the residuals analysis could reduce charged rates, making the product more accessible to farmers, broadening its coverages and contributing for adverse selection reduction.

Ultimately, our paper focused on residuals formation pattern estimation. Future studies could replicate this methodology for different crops and counties, increasing the range of evaluations and evaluating its own probability distribution. In addition, is relevant to analyze this distribution in order to capture any skews and extreme claims present in historical data.

References

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 175–185.

Atwood, J., Shaik, S., & Watts, M. (2003). Are Crop Yields Normally Distributed? A Reexamination. *American Journal of Agricultural Economics*, 888–901.

Botts, R. R., & Boles, J. N. (1958). Use of Normal-Curve Theory in Crop Insurance Ratemaking. *American Journal of Agricultural Economics*, 733–740.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., & Nesbitt, C. J. (1997). *Actuarial Mathematics* (D. Anderson (ed.); 2nd ed.). The Society of Actuaries.

Coble, K. H., Heifner, R. G., & Zuniga, M. (2000). Implications of Crop Yield and Revenue Insurance for Producer Hedging. *Journal of Agricultural and Resource Economics*, 21.

Day, R. H. (1965). Probability distributions of field crop yields. *Journal of Farm Economics*, 713–741.

de Medeiros, E. A. (2013). Avaliação da implementação do programa de subvenção do prêmio do seguro rural. *Revista de Economia e Sociologia Rural*. https://doi.org/10.1590/S0103-20032013000200005

Deng, X., Barnett, B. J., & Vedenov, D. V. (2007). Is There a Viable Market for Area-Based Crop Insurance? *American Journal of Agricultural Economics*, 508–519.

Duarte, G. V., Braga, A., Miquelluti, D. L., & Ozaki, V. A. (2018). Modeling of soybean yield using symmetric, asymmetric and bimodal distributions: implications for crop insurance. *Journal of Applied Statistics*, *45*(11), 1920–1937. https://doi.org/10.1080/02664763.2017.1406902

Gallagher, P. (1987). US Soybean yields: estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics*, 796–803.

Goodwin, B. K., & Ker, A. P. (1998). Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group-Risk Crop Insurance Contracts. *American Journal of Agricultural Economics*, 139–153.

Halcrow, H. G. (1949). Actuarial Structures for Crop Insurance. *Journal of Farm Economics*, *31*(3), 418. https://doi.org/10.2307/1232330

Harri, A., Coble, K. H., Erdem, C., & Knight, T. O. (2009). Crop Yield Distributions: A Reconciliation of Previous Research and Statistical Tests for Normality. *Review of Agricultural Economics*, 163–182.

Harri, A., Coble, K. H., Ker, A. P., & Goodwin, B. J. (2011). Relaxing heteroskedasticity assumptions in area-yield crop insurance rating. *American Journal of Agricultural Economics*, *93*(3), 703–713. https://doi.org/10.1093/ajae/aar009

Instituto Brasileiro de Geografia e Estatística. (2018). *Produção Agrícola Municipal - PAM*. Produção Agrícola Municipal - PAM.

Just, R. E., & Pope, Ru. D. (1978). Stochastic specification of production functions and economic implications. *Journal of Econometrics*, 67–86.

Just, R. E., & Weninger, Q. (1999). Are Crop Yields Normally Distributed? *American Journal of Agricultural Economics*. https://doi.org/10.2307/1244582

Ker, A. P., & Coble, K. H. (2003). Modeling Conditional Yield Densities. *American Journal of Agricultural Economics*, 291–304.

Ker, A. P., & Tolhurst, T. N. (2019). On the Treatment of Heteroskedasticity in Crop Yield Data. *American Journal of Agricultural Economics*, *101*(4), 1247–1261. https://doi.org/10.1093/ajae/aaz004

Martin, V., Hurn, S., & Harris, D. (2012). Econometric Modelling with Time Series. In *Econometric Modelling with Time Series*. https://doi.org/10.1017/cbo9781139043205

Miranda, M. J. (1991). Area-Yield Crop Insurance Reconsidered. *American Journal of Agricultural Economics*, *73*(2), 233–242. https://doi.org/10.2307/1242708

Ozaki, V. A., Goodwin, B. K., & Shirota, R. (2008). Parametric and nonparametric statistical modelling of crop yield: Implications for pricing crop insurance contracts. *Applied Economics*, *40*(9), 1151–1164. https://doi.org/10.1080/00036840600749680

Ozaki, V. A., & Silva, R. S. (2009). Bayesian ratemaking procedure of crop insurance contracts with skewed distribution. *Journal of Applied Statistics*. https://doi.org/10.1080/02664760802474256

Racine, J., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130. https://doi.org/10.1016/S0304-4076(03)00157-X

Shao, Q., Li, Z., & Xu, Zo. (2010). Trend detection in hydrological time series by segment regression with application to Shiyang River Basin. *Stochastic Environmental Research and Risk Assessment*, 221–233.

Skees, J. R., Black, J. R., & Barnett, B. J. (1997). Designing and Rating an Area Yield Crop Insurance Contract. *American Journal of Agricultural Economics*, *79*(2), 430–438. https://doi.org/10.2307/1244141

Skees, J. R., & Reed, M. R. (1986). Rate Making for Farm-Level Crop Insurance: Implications for Adverse Selection. *American Journal of Agricultural Economics*, *68*(3), 653–659. https://doi.org/10.2307/1241549

Superintendência de Seguros Privados. (2019). *Sistema de Estatísticas da SUSEP*.

Wedekin, I. (2019). *Política Agrícola no Brasil - O Agronegócio na Perspectiva Global*.

Xiao, Y., Wang, K., & Porth, L. (2017). A bootstrap approach for pricing crop yield insurance. *China Agricultural Economic Review*, *9*(2), 225–237. https://doi.org/10.1108/CAER-08-2015-0105

Ye, T., Nie, J., Wang, J., Shi, P., & Wang, Z. (2015). Performance of detrending models of crop yield risk assessment: evaluation on real and hypothetical yield data. *Stochastic Environmental Research and Risk Assessment*, *29*(1), 109–117. https://doi.org/10.1007/s00477-014-0871-x

Zeng, S., Xia, J., & Du, H. (2014). Separating the effects of climate change and human activities on runoff over different time scales in the Zhang River basin. *Stochastic Environmental Research and Risk Assessment*, *28*(2), 401–413. https://doi.org/10.1007/s00477-013-0760-8