An application of geographically weighted quantile LASSO to weather index insurance design

Daniel Lima Miquelluti, University of São Paulo, danielmiq@usp.br; Vitor Ozaki, University of São Paulo; David José Miquelluti, State University of Santa Catarina

Selected Paper prepared for presentation at the 2020 Agricultural & Applied Economics Association Annual Meeting, Kansas City, MO
July 26-28, 2020

Abstract

This article studies the efficiency of a novel regression approach, the geographically weighted quantile LASSO (GWQLASSO) in the modelling of yield-index relationship for weather index insurance products. GWQLASSO allows regression coefficients to vary spatially, while using the information from neighboring locations to derive robust estimates. The LASSO component of the model facilitates the selection of relevant explanatory variables. A weather index insurance (WII) product is developed based on 1-month SPI derived from a daily precipitation dataset for 41 weather stations in the State of Paraná (Brazil) for the period of 1979 through 2015. Soybean yield data are also used for the 41 municipalities from 1980 through 2015. The effectiveness of the GWQLASSO product is evaluated against a classic quantile regression approach and a traditional yield insurance product using the Spectral Risk Measure (SRM) and the Mean Semi-deviation. While GWQLASSO proved as effective as quantile regression it outperformed the yield insurance product, thus proving an alternative to the crop insurance market in Brazil and other locations with limited data.

Keywords: GWQLASSO; Index-insurance; Systemic risk

1. Introduction

The unpredictability of climatic variations is the principal risk factor in soybean cultivation on the south of Brazil. Reports on indemnities paid by government risk management programs, the Program for the Guarantee of Agricultural and Livestock Activity (Proagro)[1] and Rural Insurance Premium Subsidization Program (PSR) (MAPA, 2015; BACEN, 2018), shows that the occurrence of droughts are the main event of loss (85% of the insured sum), followed by

1 Created with the objective of exempting the rural producer from the fulfillment of financial obligations in rural credit operations in case of income losses motivated by climatic adversities.

excessive rain (7.6% of the insured sum) and hail (4.2% of the insured sum). In addition, losses due to strong wind, excessive temperature fluctuation and flood are also mentioned.

Crop insurance is recognized as one of the most efficient mechanisms of income protection in agriculture, transferring risk from agriculture to other agents and economic sectors. Insurance tends to stimulate the increase of cultivated area and the use of technology, especially as it acts as an additional guarantee for access to credit (Goodwin et al, 2004). In this sense, it not only contributes to the achievement of lower interest rates (Cai, 2016) by the rural producer, since the reduction of agricultural risk translates into lower credit risk, but also contributes to the development of financial, insurance and capital markets. As a result, it minimizes the pressure for subsidized credit and ex-post government financial bailout, reducing the recurring pressure for renegotiations of rural debts.

However, the degree of penetration of agricultural insurance, considering the size and relevance of Brazilian agribusiness, is still insignificant. One of the reasons for the restriction of the subsidized crop insurance program and the massification of rural insurance in the country is the limited availability of budgetary resources to fund the policies. Also, these budgetary resources depend on congress approval, thus preventing the long-term planning of investments by the private sector, imposing costs on the beneficiaries and generating dissatisfaction of the target public (MAPA, 2017).

The Proagro risk management program also faces difficulties, according to Oñate et al. (2016) there was no increase in welfare for participating farmers. Considering the fact that the pricing of Proagro does not take into account regional differences, only crop type and cultural management practices such as the use of irrigation (BACEN, 2018), we believe different approaches must be sought by the government.

A possible alternative to overcome these issues is parametric insurance, which has lower administrative and regulation 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 e Barrett, 2017). Another advantage of parametric insurance is the rapid payment of indemnities.

Parametric insurance first appeared in the pioneering written by Chakravarti (1920). After more than a decade studying the subject, the author developed an insurance product based on rainfall levels for Chitradurga in India. Indemnities were paid if total rainfall measures in the beginning of the agricultural year were 35% below normal. The payouts were divided in two periods, from January through July and from July through October, according to the production cycle. The author noted that the area should be as uniform as possible, in respect to rainfall, for the insurance to work properly. The premiums were calculated to be as close as possible to land tax value, with both premiums and indemnities depending on the land's quality. In order to keep the farmers enrolled in the insurance scheme, contracts would be ranging from 5 to 10 years, so that each farmer would receive at least one indemnity and thus perceive the value of crop insurance (Mishra, 1995; Rao, 2011).

Halcrow (1949) devised a different form of index insurance, based in the area-yields. The main idea was to develop an insurance product where indemnities would be due when the mean-yield of a uniform area fell below a pre-defined level (which could be defined as a proportion of the expected mean-yield). The size of the area could vary as long as the homogeneity of yields was maintained, and the insured farmer would select a percentage of the expected yield for the area.

The main advantage of this type of insurance over the traditional crop-insurance products is the reduction of moral hazard[2]. Since the farmer could not significantly alter the area-yield, risk

2 When the insured incur in risk increasing activities or stop taking risk-mitigating actions due to being covered by the insurance.

increasing measures are not economically viable. This would also lead to a reduction in deductibles and coverage levels limitation by the insurers (Miranda, 1991). This author also notes that adverse selection[3], which is caused by information asymmetry, is reduced in area-yield insurance as this information is available to the general public. Adding to the advantages of this type of insurance are the reduced administrative costs since an index-based insurance does not require individual assessment of yields, a major cost for traditional crop-insurance.

Two years after the work published by Miranda (1991) a yield-based index-insurance was developed by the United States Federal Crop Insurance Corporation (FCIC) in conjunction with Skees et al. (1997). The product named Group Risk Plan (GRP) was expanded in 1994 and reached 70% of market participation in 1997, considering the seven major crops and excluding forage. An additional feature of GRP was the possibility to scale the protection (the product of expected yield and expected price) up to 150%. This option was intended to increase protection since farm and county yields are not perfectly correlated. The difference between the county-yields, the index, and the value of individual yields, is called basis-risk, a problem that is always present in index-based insurance. In this way, GRP was designed to reduce basis-risk by using double exponential smoothing to forecast the central tendency of yields, scaling the protection and paying indemnities based on the percentage reduction of yields rather than the weight/volume reduction. Since yield data provided by the National Agricultural Statistics Service (NASS) are available only at the county level, it wasn't possible to change the area in order to increase homogeneity of yields.

The GRP insurance was later expanded in 1999 to cover price variations and the index turned into a revenue index, named the Group Risk Income Protection (GRIP). The expected price was

3 The inability to correctly measure farmer risk lead insurers to price the insurance incorrectly and in consequence to a greater proportion of high-risk farmers in their portfolio. This will ultimately lead to a market collapse.

calculated individually for each crop and region. Both GRP and GRIP were replaced by the Area Risk Protection Insurance Policy (ARPI) in 2013. This new policy is formed of three insurance plans, Area Revenue Protection (ARP), Area Revenue Protection with Harvest Price Exclusion (ARPwHPE) and Area Yield Protection (AYP). The ARP and ARPwHPE are similar to the GRIP and the AYP is similar to the GRP, with the harvest price exclusion option meaning the amount protected will not rise if harvest prices rise (Schnitkey, 2014).

Weather based index products were to be operated only from 2006 with the approval of flood insurance by the Peruvian government (Khalil et al., 2007). Following that, several studies and pilots were launched, mostly in developing countries (Skees et al., 2001, 2007; Giné et al., 2010; Leblois et al., 2014; Maestro et al., 2016).

Parametric insurance in Brazil is quite limited, with only one insurer offering tailored weather index insurance products (Swiss Re) as of 2018. Past initiatives include a yield index product, commercialized by AgroBrasil (Carter et al., 2015) in the state of Rio Grande do Sul and a hypothetical yield index insurance for the municipality of Castro in the state of Paraná (Ozaki, 2005).

Therefore, aiming to contribute for the expansion of parametric insurance in Brazil, we intend to assess if the Paraná state presents a suitable environment for this type of product. This study specifically targets soybean in Paraná, the second largest soybean 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). We develop a weather index product based on the Standardized Precipitation Index (SPI) and analyze its hedging effectiveness against a common yield insurance.

We also extend the work of Conradt (2015), who proposed the use of quantile regression to model the yield-index relationship, by applying the Geographically Weighted Quantile least absolute shrinkage and selection operator (GWQLASSO) (Wang, 2018) framework. Our

hypothesis is that the spatial component, captured by the latter, plays an important role in the determination of the yield-index relationship. Also, this methodology is less data intensive, as it borrows information from neighboring locations. The effectiveness of our model is compared to the traditional yield insurance and the quantile regression approach by means of two risk measures, the Spectral Risk Measure (SMR) and the Mean-semideviation model.

This article is organized as follows: in the empirical framework section, we present in detail the different methodologies utilized throughout the article, then in empirical application we give some context in our data base and the proposed index insurance product for Paraná. Our findings and discussion are found in results and discussion and we finish with conclusions.

2. Empirical Framework

This section outlines the conceptual framework strategy used in the article. We present an overview of the methods used to model the yield-index relationship and to evaluate the proposed index insurance contract.

2.1. Geographically Weighted Quantile LASSO

A natural extension to the Geographically Weighted Regression (GWR)[4] is the geographically weighted quantile regression (GWQR) model, which has the following form:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_\tau(u_i, v_i) + \epsilon_{\tau,i}$$

where $\epsilon_{\tau,i}$ is the random error term, τ is the quantile of interest, Y_i and $\boldsymbol{X}_i^T = [X_{i1}, \dots, X_{ip}]$ are respectively, the response variable Y and the explanatory variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_p$ at the geographical location $(u_i, v_i)(i = 1, \dots, n)$.

[4] A detailed description of the Quantile Regression, Geographically Weighted Regression and LASSO methods is found in Appendix A.

If $\rho_\tau(z) = z\big(\tau - I(z<0)\big)$ is the check loss function at quantile $\tau \in (0,1)$, with $I(\cdot)$ as the indicator function. For a location (u_t, v_t), let $d_{it} = \| (u_i, v_i) - (u_t, v_t) \|$, where $\|\cdot\|$ is the Euclidean norm. According to Chen et al. (2012), the local-linear GWQR estimates of the coefficients, and their partial derivatives, are the ones that minimize the local weighted quantile loss function:

$$\mathcal{L}_h(u_t, v_t) = \Sigma_{i=1}^{n} \rho_\tau \left\{ Y_i - \boldsymbol{X}_i^T \begin{bmatrix} \boldsymbol{\beta}_\tau(u_t, v_t) + \boldsymbol{\beta}_\tau^{(u)}(u_t, v_t)(u_t - u_i) + \dots \\ \dots + \boldsymbol{\beta}_\tau^{(v)}(u_t, v_t)(v_t - v_i) \end{bmatrix} \right\} K_h(d_{it})$$

with respect to $\beta_\tau(u_t, v_t)$, the partial derivatives of $\boldsymbol{\beta}_\tau(u, v)$; $\boldsymbol{\beta}_\tau^{(u)}(u_t, v_t)$ and $\boldsymbol{\beta}_\tau^{(v)}(u_t, v_t)$ for a specified kernel function $K_h(\cdot) = K(\cdot/h)/h^2$ and bandwidth h. The latter is chosen via a cross validation procedure that is identical to its GWR counterpart, while the only difference is in the check loss function replacing the quadratic loss function.

Applying the aforementioned LASSO method to the GWQR we have:

$$\mathcal{L}_{h,\lambda} = \sum_{t=1}^{n} \mathcal{L}_h(u_t, v_t) + \sum_{j=1}^{p} \begin{pmatrix} \lambda_{1j} \| \left[\beta_j(u_1, v_1), \dots, \beta_j(u_n, v_n)\right]^T \| + \cdots \\ \dots + \lambda_{2j} \| \begin{bmatrix} \beta^{(u)}(u_1, v_1), \dots, \beta^{(u)}(u_n, v_n), \\ \beta^{(v)}(u_1, v_1), \dots, \beta^{(v)}(u_n, v_n) \end{bmatrix}^T \| \end{pmatrix} \qquad (5)$$

where $\boldsymbol{\lambda}_1 = \left(\lambda_{11}, \dots, \lambda_{1p}\right)^T \in \mathbb{R}^p$ and $\boldsymbol{\lambda}_2 = \left(\lambda_{21}, \dots, \lambda_{2p}\right)^T \in \mathbb{R}^p$ are the tuning parameters. This combination of the GWQR technique and the lasso method, is named by Wang et al. (2018) the geographically weighted quantile lasso (GWQLASSO).

Given that both the local weighted quantile loss function and the penalty function in (5) are nondifferentiable at the origin, what results in the common derivative-based algorithm being unusable for obtaining the solution of $\mathcal{L}_{h,\lambda}$. Therefore, a quadratic approximation (Hunter and Lange, 2000) is used to approximate the local weighted quantile loss function, while the local quadratic approximation (Fan and Li, 2001) is used to approximate the penalty function and establish the iterative algorithm of the GWQLASSO.

2.2. Spectral Risk Measures

Traditional risk measures, such as the value at risk (VaR) and expected shortfall (ES) have some limitations. The two measures of risk do not explicitly consider the degree of risk aversion of the user of the method (Cotter & Dowd, 2010). It is implicit, when using VaR as a risk measure, that the agent has a negative risk aversion, whereas the choice of ES implies risk neutrality (Grootveld & Hallerbach, 2004). In the case of VaR, the negative risk aversion is explicit when it is verified that the agent does not weigh the losses that exceed the VaR. For ES, risk neutrality is illustrated by the fact that the agent weighs losses that exceeds the VaR uniformly. Therefore, Acerbi (2002), Dowd, Cotter and Sorwar (2008) and Cotter and Dowd (2010) argue that VaR and ES are not consistent risk measures when the agent using the technique has risk aversion.

To overcome this limitation, Acerbi (2002) proposed a measure of spectral risk that is consistent when applied to agents with risk aversion. Thus, consider the risk measure defined by:

$$M_{\varphi} = \int_0^1 q_p \, \varphi(p) dp$$

where q_p is the quantile p of the distribution of losses, $\varphi(p)$ is a weight function defined in p, and p is a cumulative probability interval such that $p \in [0,1]$.

The measure of risk M_{φ} satisfies the conditions of coherence if and only if $\varphi(p)$ satisfies the following properties:

• $\varphi(p) > 0$: the weights must always be non-negative.

• $\int_0^1 \varphi(p)dp = 1$: the sum of the weights must be equal to the unit.

• $\varphi'(p) \geq 0$: high losses are associated with weights greater than or equal to losses of smaller magnitude.

Now, one must select a suitable risk aversion function that satisfies the above properties. Here we use the exponential function of risk aversion:

$$\varphi(p) = \frac{ke^{-k(1-p)}}{1 - e^{-k}}$$

where $k > 0$ is the absolute risk aversion coefficient. This measure of spectral risk attributes greater weights to losses in the higher levels of cumulative probability distribution (the worst losses). In addition, for any dp, the weights vary more rapidly the more risk averse the agent is. The growth rate depends on the value of k, that is, the more risk averse the investor, the more the weights will grow.

2.3. Mean Semi-deviation

The standard deviation considers both the below and above average values to be equally undesirable, and this may not be consistent with the objectives of the farmers, as the concern is generally about losses, which become more serious in the case of distributions. Alternatively, we can use an indicator that considers only the dispersion of values on the left side of the distribution, that is, the semideviation, given by:

$$\sigma_{ssd} = \sqrt{\frac{\sum_{i=1}^{n}(min[0, w_i - \overline{w}_i])^2}{n}}$$

where σ_{ssd} is the default semideviation of the wealth stream, w_i are the wealth values generated by the Bayesian bootstrap procedure, and $\overline{w}_i$ is the critical point below which the farmer cares, and n is the number of observations. The value of $\overline{w}_i$ represents the minimum acceptable return, that is, the point at which the dispersion of the left distribution is measured.

The concept of semideviation is not new, and its applications in the area of finance have emerged with Markowitz (1959), who in his classic book notes that the choice between the two measures depends on the convenience, familiarity, and differences between the portfolios produced by different metrics, among other pertinent characteristics.

An important feature to be emphasized is that the numerical value of the standard deviation is at least equal to the semideviation. The immediate implication is that we cannot make a comparison between the standard deviation and the semi deviation, even though the two have equal units.

Thus, the mean semi-deviation method is expressed by σ_{ssd} and:

$$U_{it} = \begin{cases} W_{it} - E(\boldsymbol{W}_i), & \text{if } W_{it} < E(\boldsymbol{W}_i) \\ 0, & else \end{cases}$$

where U_{it} is the farmer utility and E is the expectation operator. The exposure to adverse weather conditions relative to the semideviation is then measured by:

$$V_i = E(\boldsymbol{W}_i) - \frac{1}{2} k \sigma_{ssd}$$

where V_i is the revenue risk. A higher value of V_i is conditioned to a lower level of semideviation, thus indicating less exposure to weather risk.

For both risk measures we chose a k value of 0.5 following Conradt et al (2015) and Dowd et al. (2008).

3. Empirical Application

3.1.Data cleaning and yield detrending

We utilize the National Water Agency (ANA) daily precipitation data set, focusing only on municipalities, in the state of Paraná, with an operational weather station. The time series spans from 01/10/1979 through 01/04/2015 for a total of 41 weather stations, one per municipality. We also use the series of annual soybean yields for these 41 municipalities, from 1980 through 2015, obtained from the National Institute of Geography and Statistics (IBGE).

Crop yields were detrended using the following equation (Duarte et al., 2018):

$$\tilde{y}_{t,i} = \hat{y}_{2015,i}\left(1 + \frac{\hat{e}_{t,i}}{\hat{y}_{t,i}}\right)$$

where $\tilde{y}_{t,i}$, $\hat{y}_{t,i}$, $\hat{y}_{2015,i}$ and $\hat{e}_{t,i}$ are, respectively, the corrected yield, the fitted yield, the fitted yield for 2015 and the residual for year t and municipality i.

3.2. Data pre-processing and clustering

In order to fill missing values we applied Multiple Imputation by Chained Equations (MICE) using the R software (Van Buuren, 2000) and then calculated the standardized precipitation index (SPI) with a three-month scale, thus capturing severe drought events during the crop season (Mckee et al, 1993). 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 majority vote of 30 indices, an algorithm implemented in Charrad et al (2014).

3.3.Weather Index-insurance

The state of Paraná is an important producer of soybean, being the second largest producer in Brazil. In spite of the evolution in crop technology and crop management, yields are highly susceptible to drought in some regions of the state, with as much of 50% of the final yields being dependent on water availability (Farias et al., 2001; Carmello & Sant'anna Neto, 2016).

Our WII hypothetical product is based on the standardized precipitation index (SPI) rather than cumulative rainfall. We chose this approach as there is a weak correlation between monthly precipitation and yields. This is because water availability depends on variables other than rainfall, such as water storage capacity in the soil and evapotranspiration potential, which is greatly influenced by air temperature (van Lier, 2014). The option for a rainfall-based index is also due to the better coverage of rainfall stations in Paraná.

The Standardized Precipitation Index (SPI) is based on the probabilities of overcoming a certain accumulated precipitate volume. Rainfall values are summed over several scales, for example 3, 6, 12 or 24 months, depending on the interest or need of the analyst. For a given month, for example, October, the 7-month SPI (SPI-7) is obtained from the sum of the precipitations over the seven months preceding the reference month.

The series of data, resulting from the sum of the precipitations over the months, is then adjusted to a probability distribution. In the original formulation, McKee (1993) used the Gamma distribution. From the adjustment of the probability distribution, each element of the adjusted series is assigned a probability of non-overrun. Each of these probabilities of non-overrun is finally associated with the corresponding quantile of the standard normal distribution. The quantile value of the $N(0,1)$ associated with the probability calculated in the period of interest is the SPI value for the month.

One of the advantages of using SPI, according to McKee (1993), is that SPI is only a function of probability. Thus, regardless of the probability distribution function to be used, the SPI can

be properly calculated. Other advantages are that SPI is able to characterize both dry and rainy periods, as well as the fact that it is suitable for any hydrological variable. However, the use of this index also has limitations. Mishra and Singh (2010) argue that the main one is the need for long historical records for its consistent calculation, which is not always possible (Weschenfelder et al., 2011).

The relationship between SPI and soybean yields is then modeled using the GWQLASSO framework. We follow Conradt et al. (2015) and use a method based on the inverse function of the estimated regression to determine the triggers and exits of the contract. This approach permits a precise definition of the coverage level and does not require individual tinkering of the product parameters for each location, thus facilitating and streamlining product development. For our study we chose a coverage level of 100% of the expected yield.

In preliminary assessments we found that the 1-month SPI has the highest correlation with soybean yields, thus we only present here the results for this index from October through March, the months that correspond to the planting and harvesting of soybean in most of Paraná[5].

3.4. Premium Estimation

The insurance premium is derived from the probability distribution function (pdf) of indemnities, or an approximation of this distribution. In our study, we use the Historical Burn Analysis (HBA) method to approximate the pdf of indemnities. This method is based in actual realizations of the proposed index which are then converted in payouts. The average value of these payouts represents the expected loss.

HBA is the simplest method to estimate an insurance premium, it also does not require assumptions on the pdf parameters, in contrast to other methods such as Historical Distribution

5 Planting and harvesting progress reports are available at the state level, with the months of October and March corresponding to more than 50% of the total crop area planted/harvested. These months are also assumed as planting and harvesting dates in Franchini et al. (2016).

Analysis and Monte Carlo based methods (Hess et al., 2005). We refrain from using these latter methods as our data is aggregated at the municipality level and thus it may misrepresent variability at the farm level.

In order to provide a representative data set for the premium estimation we use the first 30 years of data for this part of the analysis, with the remaining six years being used for the evaluation of the methods.

3.5. Product Evaluation

For the product evaluation we use the values of the final wealth realizations for a hypothetical farm with an area of 1 ha. The only assets present in such farm are the soybean yield and the proposed weather index insurance contract:

$$W_{it} = (1/60)vy_{it} + I_{it} - P_i,$$

where W_{it} is the final wealth, v is the price paid to the farmer for each 60kg of soybean[6], y_{it} is the corrected yield, I_{it} is the indemnity and P_i the premium, with i being the municipality and t is the year. Final wealth realizations are calculated for farmers without insurance, thus having only the first component of the right-hand side, and for farmers with the WII parameters estimated by the quantile regression and the GWQLASSO.

In order to measure the efficiency of the proposed index insurance to mitigate the risk faced by farmers we use two risk measures, namely the Spectral Risk Measure (SRM) and the Mean Semi-deviation, coupled with a Bayesian bootstrap procedure. The latter is necessary given that

6 Considering we corrected yields we utilized the 2015 average prices of soybean provided by the Department of Rural Economy (DERAL) of the State Secretariat for Agriculture and Food Supply (SEAB) in Paraná.

we only dispose of only six years of data for the evaluation. In this step a cross-validation (CV) method would be ideal but the computational requirements of GWQLASSO makes the use of CV not feasible in our case.

The Bayesian bootstrap (Rubin, 1981) is very similar to its classical counterpart (Efron, 1979) differing only in how probabilities are attached to each data value. While in the classical bootstrap a $1/n$, being n the sample size, is attributed to all n observations, in the Bayesian bootstrap the probabilities are given by a posterior distribution centered in $1/n$ but varying across replications. The main difference is in the interpretation of the results as the Bayesian bootstrap is a simulation of the posterior distribution of the parameter being estimated, whereas the classical bootstrap simulates the sampling distribution of an estimator for the parameter of interest.

The relative risk reduction (RR) is structured to compare the risk exposure of farmers in three situations: the first one being a farmer with a WII insurance designed with the GWQLASSO method against a farmer without insurance; the second situation is a farmer with WII insurance (designed with GWQLASSO or quantile regression); and a third situation for a WII insurance (designed with GWQLASSO) versus a yield insurance (YI). Thus, the general formula for the RR is:

$$RR_{case\ 1/case\ 2} = \frac{RM(W_{case\ 1}) - RM(W_{case\ 2})}{RM(W_{case\ 2})},$$

where RM stands for the risk measurements previously described and W for the final wealth realizations.

In our evaluations we consider 4000 Bayesian bootstrap replications to provide better estimates of the relative risk reduction. The latter is also tested against a hypothesis of null relative risk reduction by means of a non-parametric Wilcoxon test.

4. Results and Discussion

The optimal number of clusters from the precipitation data was two, these clusters managed to capture the different precipitation regimes identified by Keller Filho et al (2005), with cluster 1 representing areas with higher total precipitation in the year aggregate but greater variability among years and cluster 2 indicating areas with a lower total precipitation but with less variability. For the yield data, the optimal number of clusters was also two, with both clusters presenting 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. Cluster 1 contains municipalities in regions prone to drought, and thus presents lower yields and higher variability.[7]

4.1. Yield-index modelling

We observe for the cluster representing the western and northern portions of Paraná that the December SPI presents the greatest impact on yields (Appendix A - Figures 1 and 2). Given that we assume, based on state reports[8], soybean planting dates are beginning on October, the crop would be in the reproductive stage in December, thus, highly sensible to water shortage. Therefore, for the premium estimation in cluster 1 we select the December SPI as the index. For the central and eastern portions of the state, however, both December and February SPI are impacting yields, with the February SPI having a slightly higher impact and thus being the one selected as the index for cluster 2.

Note that these coefficients are selected based on their boxplot as GWQLASSO does not account for temporal structure of the data, thus assuming that all observations come from a unique point in time and resulting in 30 coefficients for each location. Albeit this represents a

[7] A thorough discussion in the results of the clustering analysis is found in Miquelluti (2019).

[8] Available in: http://www.agricultura.pr.gov.br/modules/conteudo/conteudo.php?conteudo=32

limitation to the modelling, the impact is reduced as we used time-detrended yields and are not interested in the temporal behavior of the yield/index series, just in their intrinsic relationship. Also, the incorporation of the LASSO method permits a better identification of relevant variables, as the ones with little importance to yields rapidly converge to zero.

The possibility to model yields at several locations simultaneously while considering the spatial structure of the relationship between yields and explanatory variables and also having a method to quickly dismiss unimportant variables present a great opportunity to WII scalability. One of the major issues of WII is its low ability to grow at scale as the models developed for one location may prove completely obsolete as you move away from it, however, with GWQLASSO one may inspect both the general significance of the explanatory variables and their coefficient for each location. This permits a faster screening of possible indices, along with their respective triggers when using the methodology detailed here and proposed in Conradt (2015).

Another hindrance to the spread of WII is the absence of long series of yield data. The GWQLASSO method is less affected by this issue, as it uses information from neighboring yields in the estimation process. This characteristic is especially important in developing countries, which in general does not have long series of yield and weather data.

The relationship between the December SPI and yields, for cluster 1, varies by more than five times when we compare municipalities on the west of Paraná to the ones in the center region of the state (Appendix A - Figure 3). This goes in line with the characteristics of these regions, soils to the west, mostly in the northwest, are sandy, and the climate is classified as Cfa, with higher temperatures in the summer, both unfavorable to soybean. These characteristics result in diminished yields and higher susceptibility to drought, which are translated in the coefficients

for December SPI. A higher variability in yields, when compared to other regions in the state, coupled with susceptibility to drought is also observed by Franchini et al (2016).

The center and east of the state are classified as Cfb in Koeppen's system. This means that these municipalities have lower temperatures in the summer, what benefits soybean plants. Also, the soil in these regions have more clay, what leads to a higher capacity to contain water and consequentially mitigate the effects of drought, resulting in a less pronounced relation to the index, for both cluster 1 and 2 (Appendix A - Figures 3 and 4).

4.2. Weather Index Insurance premium and performance

WII is generally associated with lower premiums, mainly due to the lack of in situ crop inspection after a claim is filed. Here we compare a traditional yield insurance product with a 65% coverage to our proposed WII product with a 100% coverage. While this comparison may not be fair to our product, as by definition a higher coverage means a higher premium, we found the 65% coverage level to be the most common for soybean in Paraná. Coverage levels above 90% are rare in Brazil and suffer from two problems, the higher premiums and the inferior percentage in subsidization by the government, thus they are not attractive to the farmer.

Our results show that the index insurance may vary from half to three times the price of the common yield insurance. There is a tendency of pricier index insurance, compared to the yield insurance, as we move to the western portion of Paraná (Appendix A - Figure 5), what is expected as this region is more susceptible to drought.

In our design we do not consider gains from scale and the spatial diversification of risk by the insurer, this would lead to a lesser difference between our product and the commercial product depicted here. Even so, the WII results in a net gain for the producer, as for both clusters and

risk measurements it performs better than the yield insurance. The results from both risk measures indicate that both GWQLASSO and quantile regression provide similar risk reduction, with both being more effective than a yield insurance product with a 65% coverage level (Appendix A - Tables 1 through 4). Borrowing from results in Conradt (2015) we can also derive that GWQLASSO is superior to ordinary least squares.

As for the public policy implications, WII has proved to be a superior alternative to basic crop insurance products as a yield insurance with a 65% coverage level. Considering that Oñate (2016) showed that PROAGRO, a risk management tool similar to a crop-credit insurance, has not increased farmers welfare and is not priced according to regional characteristics, we favor the expansion of government operated or funded parametric insurance products. A WII product could be implemented as a microinsurance policy to small farmers or as a macroinsurance directly to the government. The latter would also be further advantageous, as the efficiency of WII grows with scale (Miranda and Farrin, 2012). Several products of this type have been successfully implemented in developing countries such as the "Comité de Ayuda a Desastres y Emergencias Nacionales" (CADENA) program in Mexico (de Janvry et al, 2016) and the "Pradhan Mantri Fasal Bima Yojana" (PMFBY) index insurance scheme in India (Rathore et al, 2017).

5. Conclusion

Despite the efforts by the central government, crop insurance is yet to take off in Brazil. Inconsistent budget, information asymmetry and moral hazard are some of the issues that crippled the program and continue to impede its expansion. In this sense, parametric insurance may present an alternative to the local insurance market. Thus, aiming to foster the growth of parametric insurance in Brazil and contribute to the development of this type of insurance throughout the globe we design a WII product by using a novel approach to model the yield-index relationship, the GWQLASSO. This methodology compounds the flexible modelling and

robustness of quantile regression with the spatial component of geographically weighted regression and variable selection prowess of the LASSO method.

We test our assumptions using a crop insurance application in Paraná, Brazil. The 36 years long time series of precipitation and soybean yield data are split in design and evaluation sets, with the latter having only six years of data and thus requiring the use of Bayesian bootstrap to improve the reliability of results. To measure the ability of WII to reduce risk, when compared to yield insurance and between yield modelling approaches, we use two different risk measures, the Spectral Risk Measure and the Mean Semi-deviation.

Regarding the performance of WII in Paraná our findings indicate that index insurance is superior to a 65% coverage yield insurance in 41 municipalities of the state, despite being up to three times more expensive than this product. However, the GWQLASSO approach proved as effective as the regular quantile regression. The latter may seem as a discouragement to the use of a more complex model, nevertheless, some of the characteristics of GWQLASSO (less data intensive and simpler conjoint variable selection) argue in its favor.

Future studies are needed to confirm the viability of WII in other regions and crops throughout the country. Also, regardless of our efforts to mitigate the effect of the level of aggregation in the crop yield data and lack of precise planting dates, these may lead to a loss of accuracy in product design that is unacceptable in a commercial environment. Therefore, tighter cooperation between risk bearers and insurance researchers and product developers is needed.

Compliance with Ethical Standards

Funding

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

Conflict of Interest

Author Daniel Lima Miquelluti declares that he has no conflict of interest. Author Vitor Ozaki declares that he has no conflict of interest. Author David José Miquelluti declares that he has no conflict of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

Acerbi, Carlo. "Spectral measures of risk: A coherent representation of subjective risk aversion." *Journal of Banking & Finance* 26, no. 7 (2002): 1505-1518.

BANCO CENTRAL DO BRASIL - BACEN. Departamento De Regulação, Supervisão E Controle Das Operações Do Crédito Rural E Do Proagro – DEROP. *Programa De Garantia Da Atividade Agropecuária PROAGRO Relatório Circunstanciado 2015 a 2018*. 2018.

Brunsdon, Chris, Stewart Fotheringham, and Martin Charlton. "Geographically weighted regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47, no. 3 (1998): 431-443.

Cai, Jing. "The impact of insurance provision on household production and financial decisions." *American Economic Journal: Economic Policy* 8, no. 2 (2016): 44-88.

Carmello, Vinicius, and João Lima SANT'ANNA NETO. "Rainfall Variability and Soybean Yield in Paraná State, Southern Brazil." *International Journal of Environmental & Agriculture Research* 2, no. 1 (2016): 86-97.

Carter, Michael, Alain de Janvry, Elisabeth Sadoulet, and Alexander Sarris. "Index-based weather insurance for developing countries: A review of evidence and a set of propositions for up-scaling." *Development Policies working paper* 111 (2014).

Cleveland, William S. "Robust locally weighted regression and smoothing scatterplots." *Journal of the American statistical association* 74, no. 368 (1979): 829-836.

Collier, B., J. Skees, e B. Barnett, 2009 Weather index insurance and climate change: opportunities and challenges in lower income countries. The Geneva Papers on Risk and Insurance Issues and Practice **34**: 401–424.

Collier, B., B. Barnett, e J. Skees, 2010 State of knowledge report–data requirements for the design of weather index insurance.

Chakravarti, J. S. "Agricultural insurance: a practical scheme suited to Indian Conditions." (1920).

Charrad, Malika, et al. "Package 'NbClust'." *Journal of Statistical Software* 61 (2014): 1-36.

Conradt, Sarah, Robert Finger, and Raushan Bokusheva. "Tailored to the extremes: Quantile regression for index-based insurance contract design." *Agricultural economics* 46, no. 4 (2015): 537-547.

Cotter, John, and Kevin Dowd. "Estimating financial risk measures for futures positions: A nonparametric approach." *Journal of Futures Markets: Futures, Options, and Other Derivative Products* 30, no. 7 (2010): 689-703.

de Carvalho, José Ruy Porto, et al. "Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults." *Revista Brasileira de Meteorologia* 32.4 (2017): 575-583.

De Janvry, Alain, Elizabeth Ramirez Ritchie, and Elisabeth Sadoulet. *Weather index insurance and shock coping: evidence from Mexico's CADENA Program*. The World Bank, 2016.

Dowd, Kevin, John Cotter, and Ghulam Sorwar. "Spectral risk measures: properties and limitations." *Journal of Financial Services Research* 34, no. 1 (2008): 61-75.

Duarte, Gislaine V., Altemir Braga, Daniel L. Miquelluti, and Vitor A. Ozaki. "Modeling of soybean yield using symmetric, asymmetric and bimodal distributions: implications for crop insurance." *Journal of Applied Statistics* 45, no. 11 (2018): 1920-1937.

Efron, B. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7, no. 1 (1979): 1-26.

Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96, no. 456 (2001): 1348-1360.

Farias, José Renato Bouças, Eduardo Delgado Assad, IR de Almeida, Balbino Antônio EVANGELISTA, C. Lazzarotto, Norman Neumaier, and Alexandre Lima Nepomuceno. "Caracterização de risco de déficit hídrico nas regiões produtoras de soja no Brasil." *Revista Brasileira de Agrometeorologia* 9, no. 3 (2001): 415-421.

Fotheringham, Stewart, Chris Brundson, and Martin Charlton. *Geographically weighted regression & associated techniques*. Wiley, 2002.

Franchini, Julio Cezar, Alvadi Antonio Balbinot, Jr., Pablo Ricardo Nitsche, Henrique Debiasi, and Ivani De Oliveira Negrão Lopes. "Variabilidade Espacial E Temporal Da Produção De Soja No Paraná E Definição De Ambientes De Produção." (2016). Accessed September 22, 2018. https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1052786

Gallagher, Paul. "US soybean yields: Estimation and forecasting with nonsymmetric disturbances." *American Journal of Agricultural Economics* 69.4 (1987): 796-803.

Giné, Xavier, Lev Menand, Robert Townsend, and James Vickery. *Microinsurance: a case study of the Indian rainfall index insurance market*. The World Bank, 2010.

Goodwin, Barry K., Monte L. Vandeveer, and John L. Deal. "An empirical analysis of acreage effects of participation in the federal crop insurance program." *American Journal of Agricultural Economics* 86, no. 4 (2004): 1058-1077.

Grootveld, H., and W. G. Hallerbach. "Upgrading value-at-risk from diagnostic metric to decision variable: a wise thing to do? Pp, 33-50 in G, Szegö (Ed,) Risk Measures for the 21st Century." (2004).

Halcrow, Harold G. "Actuarial structures for crop insurance." *Journal of Farm Economics* 31, no. 3 (1949): 418-443.

Hess, Ulrich, J. R. Skees, Andrea Stoppa, B. J. Barnett, and John Nash. "Managing agricultural production risk: Innovations in developing countries." *Agriculture and Rural Development (ARD) Department Report* 32727-GLB (2005).

Hunter, David R., and Kenneth Lange. "Quantile regression via an MM algorithm." *Journal of Computational and Graphical Statistics* 9, no. 1 (2000): 60-77.

Jensen, Nathaniel, and Christopher Barrett. "Agricultural index insurance for development." *Applied Economic Perspectives and Policy* 39, no. 2 (2017): 199-219.

Khalil, Abedalrazq F., Hyun-Han Kwon, Upmanu Lall, Mario J. Miranda, and Jerry Skees. "El Niño–Southern Oscillation–based index insurance for floods: Statistical risk analyses and application to Peru." *Water Resources Research* 43, no. 10 (2007).

Koenker, Roger. *Quantile Regression*. Econometric Society Monographs. Cambridge: Cambridge University Press, 2005. doi:10.1017/CBO9780511754098.

Leblois, Antoine, Philippe Quirion, Agali Alhassane, and Seydou Traoré. "Weather index drought insurance: an ex ante evaluation for millet growers in Niger." *Environmental and Resource Economics* 57, no. 4 (2014): 527-551.

Maestro, Teresa, Barry J. Barnett, Keith H. Coble, Alberto Garrido, and María Bielza. "Drought Index Insurance for the Central Valley Project in California." *Applied Economic Perspectives and Policy* 38, no. 3 (2016): 521-545.

Markowitz, Harry. "Portfolio Selection, Cowles Foundation Monograph No. 16." *John Wiley, New York. S. Moss (1981). An Economic theory of Business Strategy, Halstead Press, New York. TH Naylor (1966). The theory of the firm: a comparison of marginal analysis and linear programming. Southern Economic Journal (January)* 32 (1959): 263-74.

McKee, Thomas B., Nolan J. Doesken, and John Kleist. "The relationship of drought frequency and duration to time scales." *Proceedings of the 8th Conference on Applied Climatology*. Vol. 17. No. 22. Boston, MA: American Meteorological Society, 1993.

McKee, Thomas B. "Drought monitoring with multiple time scales." *Proceedings of 9th Conference on Applied Climatology, Boston, 1995*. 1995.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO - MAPA. Departamento De Gestão De Riscos. *Relatório das indenizações pagas entre 2006 a 2015 – Programa de Subvenção ao Prêmio do Seguro Rural (PSR)*. 2015.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO - MAPA. Departamento De Gestão De Riscos. *Relatório Geral 2017 – Programa de Subvenção ao Prêmio do Seguro Rural (PSR)*. 2017.

Miquelluti, Daniel Lima. "Weather index insurance design: a novel approach for crop insurance in Brazil." PhD diss., Universidade de São Paulo. 2019.

Miranda, Mario J. "Area-yield crop insurance reconsidered." *American Journal of Agricultural Economics* 73, no. 2 (1991): 233-242.

Miranda, Mario J., and Katie Farrin. "Index insurance for developing countries." *Applied Economic Perspectives and Policy* 34, no. 3 (2012): 391-427.

Mishra, Ashok K., and Vijay P. Singh. "A review of drought concepts." *Journal of hydrology* 391, no. 1-2 (2010): 202-216.

Mishra, Pramod K. "Is rainfall insurance a new idea? Pioneering work revisited." *Economic and Political Weekly* (1995): A84-A88.

Ozaki, Vitor Augusto, and Ricardo Shirota. "Um estudo da viabilidade de um programa de seguro agrícola baseado em um índice de produtividade regional em Castro (PR)." *Revista de Economia e Sociologia Rural* 43, no. 3 (2005): 485-503.

Rao, Kolli N. "Weather Index Insurance: Is it the Right Model for Providing Insurance to Crops?" *ASCI Journal of Management* 41, no. 1 (2011): 86-101.

Rathore, Vandana, and M. J. Rao. "The performance of PMFBY and other crop insurance models in India." *Int. J. Adv. Res. Dev* 2 (2017): 2455-4030.

Rubin, Donald B. "The Bayesian bootstrap." *The annals of statistics* (1981): 130-134.

Schnitkey, Gary. "Area Risk Protection Insurance Policy: Comparison to Group Plans." (2014).

Skees, Jerry R., J. Roy Black, and Barry J. Barnett. "Designing and rating an area yield crop insurance contract." *American journal of agricultural economics* 79, no. 2 (1997): 430-438.

Skees, Jerry Robert. *Developing rainfall-based index insurance in Morocco*. Vol. 2577. World Bank Publications, 2001.

Skees, Jerry R., Jason Hartell, and Anne G. Murphy. "Using index-based risk transfer products to facilitate micro lending in Peru and Vietnam." *American Journal of Agricultural Economics* 89, no. 5 (2007): 1255-1261.

Staniswalis, Joan G. "The kernel estimate of a regression function in likelihood-based models." *Journal of the American Statistical Association* 84, no. 405 (1989): 276-283.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

van Lier, Quirijn de Jong. "Water Availability to Plants." In *Application of Soil Physics in Environmental Analyses*, pp. 435-452. Springer, Cham, 2014.

Wang, Wentao, Shoufang Xu, and Tianshun Yan. "Structure identification and model selection in geographically weighted quantile regression models." *Spatial Statistics* (2018).

Weschenfelder, Adriana Burin, Karine Pickbrenner, and Eber José de Andrade PINTO. "Análise da frequência de ocorrência e a classificação das precipitações diárias máximas anuais na região da Laguna dos Patos (sub-bacia 87)." (2011).

Appendix A

Table 1. Relative Risk Reduction per Municipality in Cluster 1 According to the Spectral Risk Measure (SRM)

Municipality	$RR_{GWQLASSO}$			$RR_{GWQLASSO/QR}$			$RR_{GWQLASSO/YI}$		
	2.5%	Median	97.5%	2.5%	Median	97.5%	2.5%	Median	97.5%
Alto Piquiri	-0,1269	-0,0194	0,1459	-0,1127	-0,0056*	0,1117	-0,0691	0,0478*	0,2450
Ampére	-0,1291	-0,0150	0,1711	-0,1113	-0,0012	0,1221	-0,0816	0,0401*	0,2398
Andirá	-0,1317	0,0841	0,4253	-0,1722	0,0191*	0,2521	-0,0758	0,1470*	0,4714
Cambará	-0,1907	0,0076	0,2491	-0,1984	0,0025*	0,2538	-0,1523	0,0732*	0,3356
Campo Mourão	-0,1007	-0,0090	0,0953	-0,0927	0,0041*	0,1055	-0,0546	0,0398*	0,1621
Céu Azul	-0,1610	0,0004	0,2551	-0,1595	-0,001*	0,2099	-0,1101	0,0543*	0,2525
Clevelândia	-0,1156	-0,0077	0,1239	-0,1149	-0,0089*	0,1219	-0,0639	0,0424*	0,1830

Coronel Vivida	-0,2051	-0,0012	0,2519	-0,1915	0,0027	0,2465	-0,1703	0,0452*	0,2845
Formosa do Oeste	-0,1223	0,0513	0,3600	-0,1938	-0,0255*	0,1583	-0,0884	0,0939*	0,3761
Foz do Iguaçu	-0,2061	-0,0061	0,2812	-0,1968	-0,0007	0,2641	-0,1477	0,0468*	0,2916
Guaraniaçu	-0,0838	0,0086	0,1902	-0,1242	-0,0096*	0,1387	-0,0047	0,1035*	0,3146
Guarapuava	-0,0654	-0,0146	0,0306	-0,0536	-0,0015*	0,0517	-0,0159	0,0365*	0,0877
Ivaiporã	-0,0754	0,0101	0,0962	-0,0729	0,006*	0,0817	-0,0289	0,0664*	0,1580
Janiópolis	-0,0819	0,0148	0,1739	-0,1032	-0,0142*	0,0665	-0,0403	0,0678*	0,2515
Mamborê	-0,0609	0,0161	0,1068	-0,0766	-0,0017*	0,0708	-0,0072	0,0737*	0,1718
Manoel Ribas	-0,0960	-0,0051	0,0966	-0,0940	-0,0049*	0,0866	-0,0569	0,0423*	0,1548
Mariluz	-0,1125	0,0021	0,1730	-0,1149	0,0011*	0,1650	-0,0496	0,0722*	0,2705
Mariópolis	-0,1355	-0,0196	0,1762	-0,1117	0,001	0,1146	-0,0928	0,0255*	0,2021

Matelândia	-0,2177	-0,0355	0,2408	-0,1952	-0,0178*	0,2228	-0,1713	0,0107*	0,2166
Nova Esperança	-0,1745	-0,0164	0,1832	-0,1541	0,0047*	0,1982	-0,1191	0,0520*	0,2792
Palmas	-0,0514	0,0006	0,0545	-0,0511	0,0003	0,0556	-0,0098	0,0432*	0,1019
Pato Branco	-0,1411	-0,0264	0,1552	-0,1354	-0,0198*	0,1392	-0,0949	0,0344*	0,2290
Pitanga	-0,1261	-0,0224	0,0809	-0,1127	-0,008*	0,0985	-0,0768	0,0360*	0,1489
Prudentópolis	-0,0671	-0,0100	0,0406	-0,0643	-0,0025	0,0627	-0,0207	0,0386*	0,0914
Quedas do Iguaçu	-0,1423	-0,0341	0,1200	-0,0852	0,0212*	0,1112	-0,0926	0,0204*	0,1876
Roncador	-0,0955	-0,0026	0,1026	-0,0945	-0,0007	0,1072	-0,0466	0,0548*	0,1758
Salto do Lontra	-0,1587	-0,0316	0,1312	-0,1331	-0,0128*	0,1211	-0,1310	0,0035*	0,1887
Santo Antônio da Platina	-0,1100	0,0181	0,1963	-0,1099	0,0089*	0,1605	-0,0549	0,0811*	0,2803
São Jorge d'Oeste	-0,1958	-0,0146	0,2451	-0,1800	-0,0077*	0,2292	-0,1669	0,0148*	0,2518

São Miguel do Iguaçu	-0,1811	0,0151	0,3174	-0,2162	-0,0129*	0,2007	-0,1401	0,0603*	0,3187
Terra Roxa	-0,2323	0,0040	0,3032	-0,2307	-0,0034	0,2747	-0,1869	0,0432*	0,2877
Tibagi	-0,0546	0,0040	0,0777	-0,0510	-0,0017*	0,0481	-0,0094	0,0529*	0,1341
Toledo	-0,2095	-0,0251	0,2577	-0,1832	0,0057*	0,2260	-0,1671	0,0121*	0,2199
Ubiratã	-0,1043	-0,0065	0,0989	-0,1068	-0,0033*	0,1055	-0,0677	0,0425*	0,1584

Note: *significant at the 5% significance level

Source: Authors

Table 2. Relative Risk Reduction per Municipality in Cluster 1 According to the Mean Semi-Deviation

Municipality	$RR_{GWQLASSO}$			$RR_{GWQLASSO/QR}$			$RR_{GWQLASSO/YI}$		
	2.5%	Median	97.5%	2.5%	Median	97.5%	2.5%	Median	97.5%
Alto Piquiri	-0,1264	-0,0063	0,1975	-0,1179	-0,0075*	0,1260	-0,0695	0,0611*	0,2836
Ampére	-0,1329	0,0027	0,2175	-0,1182	-0,0010	0,1383	-0,0803	0,0626*	0,2942
Andirá	-0,1192	0,1149	0,5029	-0,1770	0,0145*	0,2711	-0,0740	0,1663*	0,5406
Cambará	-0,2145	0,0042	0,2565	-0,2207	0,0039*	0,2835	-0,1647	0,0705*	0,3415
Campo Mourão	-0,1131	-0,0083	0,1159	-0,1062	0,0035*	0,1195	-0,0691	0,0454*	0,1772
Céu Azul	-0,1823	0,0135	0,3513	-0,1821	0,0034	0,2479	-0,1391	0,0501*	0,2829
Clevelândia	-0,1209	-0,0042	0,1430	-0,1177	-0,0021	0,1336	-0,0727	0,0500*	0,2040
Coronel Vivida	-0,2210	-0,0029	0,3084	-0,2217	-0,0002	0,2879	-0,1888	0,0433*	0,3327

Formosa do Oeste	-0,1360	0,0647	0,4381	-0,2105	-0,0324*	0,1685	-0,0956	0,1005*	0,4181
Foz do Iguaçu	-0,2108	0,0094	0,3584	-0,2213	-0,0010*	0,3078	-0,1740	0,0479*	0,3105
Guaraniaçu	-0,0795	0,0211	0,2260	-0,1127	-0,0073*	0,1362	0,0077	0,1257*	0,3703
Guarapuava	-0,0719	-0,0177	0,0360	-0,0585	-0,0012*	0,0587	-0,0217	0,0341*	0,0897
Ivaiporã	-0,0740	0,0154	0,1006	-0,0688	0,0062*	0,0830	-0,0230	0,0713*	0,1632
Janiópolis	-0,0900	0,0280	0,2251	-0,1225	-0,0184*	0,0775	-0,0409	0,0883*	0,3023
Mamborê	-0,0645	0,0197	0,1175	-0,0765	-0,0024	0,0782	-0,0064	0,0777*	0,1889
Manoel Ribas	-0,1054	-0,0015	0,1108	-0,1014	-0,0023	0,1090	-0,0597	0,0513*	0,1739
Mariluz	-0,1200	0,0115	0,2177	-0,1216	0,0068*	0,1904	-0,0566	0,0846*	0,3290
Mariópolis	-0,1422	-0,0030	0,2378	-0,1351	-0,0043*	0,1226	-0,1032	0,0394*	0,2540
Matelândia	-0,2383	-0,0270	0,2879	-0,2206	-0,0153*	0,2695	-0,2096	0,0042*	0,2616

Nova Esperança	-0,1800	-0,0154	0,2114	-0,1637	0,0031*	0,2192	-0,1216	0,0619*	0,3204
Palmas	-0,0509	0,0004	0,0562	-0,0529	0,0009	0,0563	-0,0098	0,0432*	0,1041
Pato Branco	-0,1490	-0,0101	0,1972	-0,1407	-0,0062*	0,1673	-0,0984	0,0503*	0,2741
Pitanga	-0,1350	-0,0257	0,0858	-0,1160	-0,0076*	0,1105	-0,0837	0,0336*	0,1576
Prudentópolis	-0,0706	-0,0164	0,0356	-0,0613	0,0001*	0,0714	-0,0274	0,0324*	0,0884
Quedas do Iguaçu	-0,1467	-0,0209	0,1694	-0,1007	0,0139*	0,1168	-0,1034	0,0335*	0,2417
Roncador	-0,1031	0,0001	0,1166	-0,1056	-0,0033	0,1189	-0,0498	0,0580*	0,1901
Salto do Lontra	-0,1636	-0,0222	0,1632	-0,1433	-0,0165*	0,1332	-0,1265	0,0203*	0,2314
Santo Antônio da Platina	-0,1060	0,0345	0,2447	-0,1200	0,0155*	0,1951	-0,0464	0,1032*	0,3303
São Jorge d'Oeste	-0,2115	-0,0019	0,3096	-0,2074	-0,0022*	0,2820	-0,1863	0,0238*	0,3015
São Miguel do Iguaçu	-0,1936	0,0278	0,4372	-0,2257	-0,0153*	0,2169	-0,1591	0,0621*	0,3582

Terra Roxa	-0,2673	0,0036	0,4015	-0,2827	-0,0060	0,3495	-0,2494	0,0257*	0,3068
Tibagi	-0,0538	0,0110	0,0876	-0,0578	-0,0040*	0,0493	-0,0107	0,0594*	0,1430
Toledo	-0,2288	-0,0173	0,3539	-0,2166	-0,0005	0,2440	-0,2068	0,0039	0,2347
Ubiratã	-0,1190	-0,0087	0,1096	-0,1136	-0,0044*	0,1130	-0,0768	0,0434*	0,1789

Note: *significant at the 5% significance level

Source: Authors

Table 3. Relative Risk Reduction per Municipality in Cluster 2 According to the Spectral Risk Measure (SRM)

Municipality	$RR_{GWQLASSO}$			$RR_{GWQLASSO/QR}$			$RR_{GWQLASSO/YI}$		
	2.5%	Median	97.5%	2.5%	Median	97.5%	2.5%	Median	97.5%
Palmeira	-0,0863	0,0083	0,0971	-0,0900	0,0039*	0,1028	-0,0470	0,0544*	0,1531
Piraí do Sul	-0,0359	-0,0019	0,0475	-0,0289	-0,0033*	0,0248	0,0055	0,0425*	0,0964
Ponta Grossa	-0,0427	-0,0058	0,0285	-0,0329	0,0032*	0,0444	0,0022	0,0411*	0,0782
Porto Amazonas	-0,1065	-0,0043	0,0976	-0,1010	0,0011*	0,1135	-0,0623	0,0428*	0,1542
Rio Negro	-0,0640	-0,0006	0,0623	-0,0598	0,0002	0,0665	-0,0096	0,0570*	0,1286
São Mateus do Sul	-0,0684	-0,0331	0,0048	-0,0318	-0,0045*	0,0240	-0,0160	0,0209*	0,0609
União da Vitória	-0,0705	-0,0017	0,0723	-0,0579	0,0056*	0,0690	-0,0279	0,0411*	0,1218

Note: *significant at the 5% significance level

Source: Authors

Table 4. Relative Risk Reduction per Municipality in Cluster 2 According to the Mean Semi-Deviation

Municipality	$RR_{GWQLASSO}$			$RR_{GWQLASSO/QR}$			$RR_{GWQLASSO/YI}$		
	2.5%	Median	97.5%	2.5%	Median	97.5%	2.5%	Median	97.5%
Palmeira	-0,0993	0,0032	0,0995	-0,0994	0,0014	0,1127	-0,0543	0,0502*	0,1535
Piraí do Sul	-0,0330	0,0061	0,0579	-0,0309	-0,0034*	0,0268	0,0111	0,0513*	0,1083
Ponta Grossa	-0,0435	-0,0056	0,0304	-0,0338	0,0060*	0,0504	-0,0003	0,0422*	0,0815
Porto Amazonas	-0,1155	-0,0058	0,1047	-0,1101	0,0013*	0,1335	-0,0773	0,0422*	0,1638
Rio Negro	-0,0657	-0,0009	0,0679	-0,0607	0,0007	0,0704	-0,0114	0,0576*	0,1330
São Mateus do Sul	-0,0620	-0,0289	0,0096	-0,0310	-0,0045*	0,0229	-0,0112	0,0261*	0,0685
União da Vitória	-0,0757	0,0013	0,0812	-0,0664	-0,0012	0,0696	-0,0323	0,0444*	0,1291

*Significant at the 5% significance level

Source: Authors

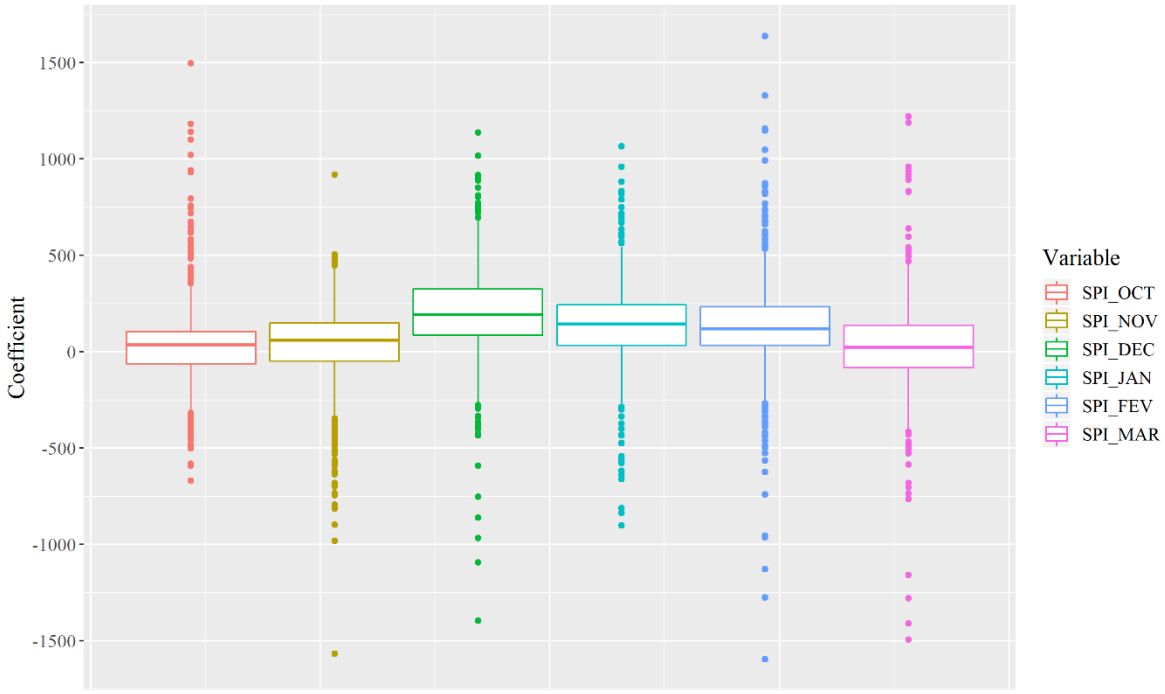

Figure 1. Boxplot of the GWQLASSO coefficients for cluster 1

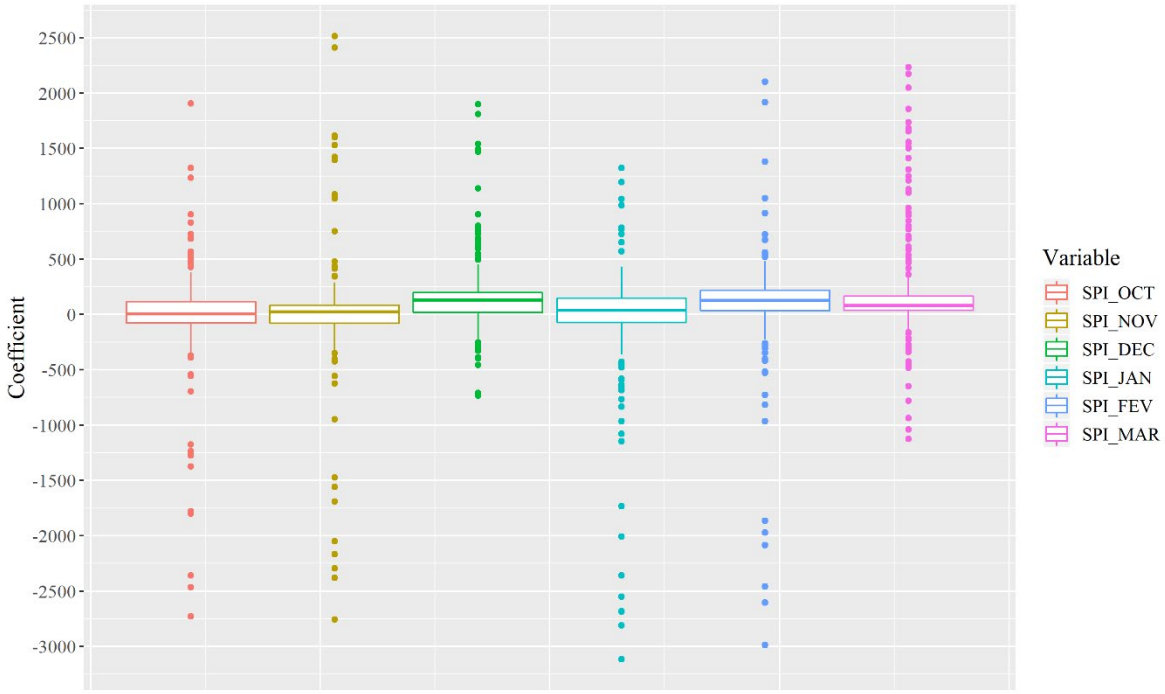

Figure 2. Boxplot of the GWQLASSO coefficients for cluster 2

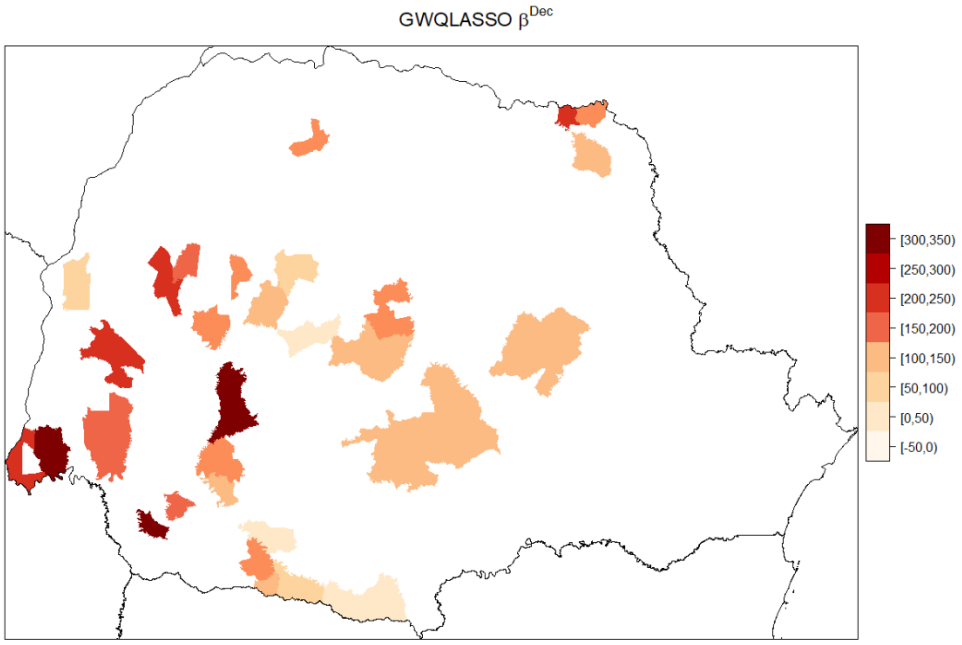

Figure 3. GWQLASSO β^{Dec} coefficients spatial distribution for cluster 1

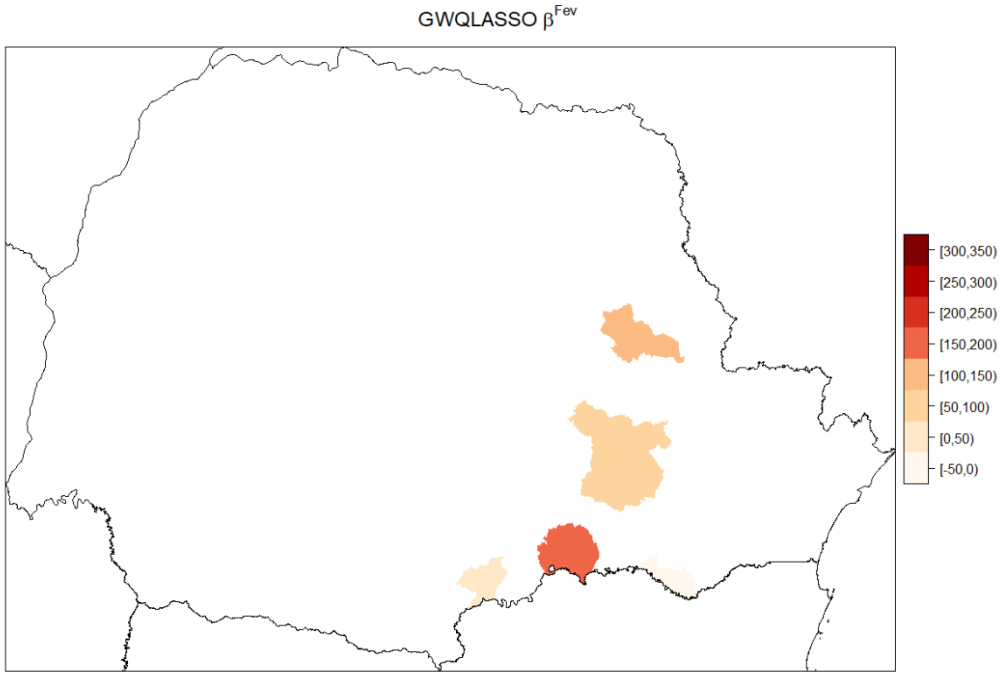

Figure 4. GWQLASSO β^{Feb} coefficients spatial distribution for cluster 2

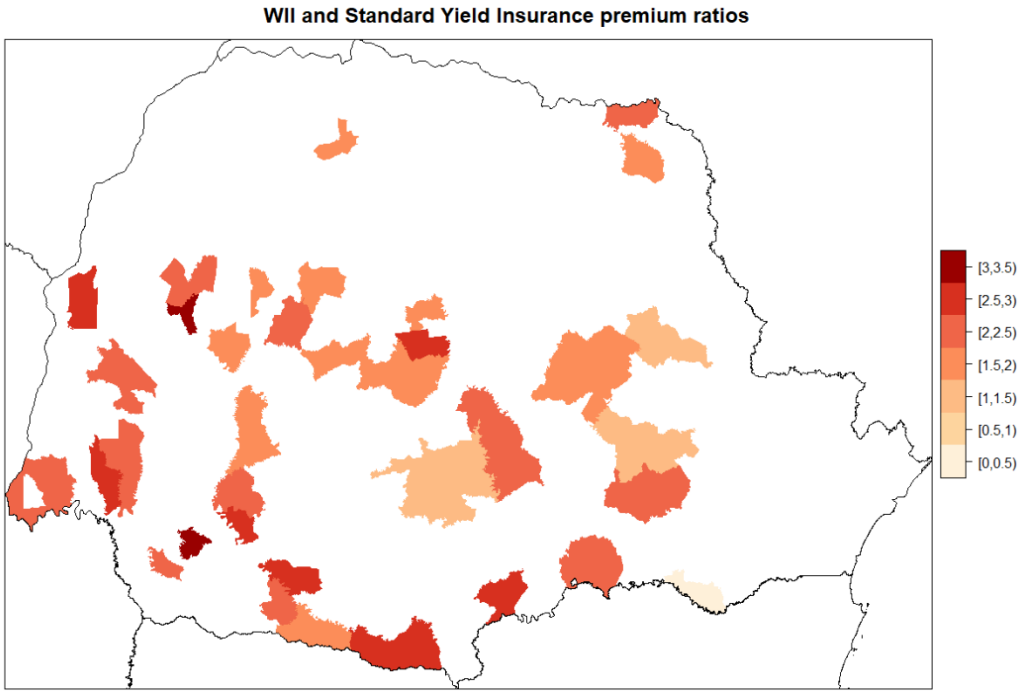

Figure 5. Ratio of GWQLASSO calculated premiums to standard yield insurance premiums

Appendix B

Quantile Regression

Quantile regression models the causal effects of covariates in different quantiles of the cumulative distribution function of the response variable, and therefore are an alternative approach to the usual linear regression methodology. That is, while the classical models are limited to the analysis of conditional mean, the quantile regression allows analysis throughout the conditional distribution of the response variable in the covariates.

The quantile regression models emerged as a generalization of the absolute residual minimization method developed in the early 19th century. The quantile regression has long stumbled on the difficulty of estimating the parameters that, unlike the usual linear regression models, have no analytical formula. However, with the advent of computers, as well as the development of linear programming techniques, the methodology has been gaining more space in empirical studies and academic research.

Let $Y_i, i = 1, \dots, n$, random variables and $\boldsymbol{x}_i \in \mathbb{R}^p$ the observed vector of covariates. Consider that the variables Y_i are conditionally independent given $\boldsymbol{x}_i$; $\forall i = 1, \dots, n$.

Whereas the usual regression is limited to describing the relationship of Y_i with the covariates of the study under the terms of conditional methods, quantile regression is a statistical modeling technique which allows analyzing this relationship in any quantile τ of interest, $\tau \in [0,1]$. In other words, it is a methodology capable of describing the function $f(\cdot, \tau)$ such that

$$Q_{Y_i|\boldsymbol{x}_i}(\tau) = f(\boldsymbol{x}_i, \tau) \quad (1)$$

for all $\tau \in [0,1]$. The function $f(\cdot, \tau)$ is the systematic part of the regression model. Note that $f(\cdot, \tau)$ may be different for each τ.

An intuitive way of understanding quantile regression, which is usually presented in the area literature, is an analogy to classical regression models (see, for example, Koenker (2005)).

In this case, each observed value of the response variable of the study is given by the sum of a systematic part, which is the quantile of order τ of Y_i, $f(\boldsymbol{x}_i, \tau)$ and a random error ϵ_i. This is:

$$y_i = f(\boldsymbol{x}_i, \tau) + \epsilon_i,$$

with independent and identically distributed ϵ_i. Assuming that the order τ of ϵ_i, conditional to $\boldsymbol{x}_i$, is equal to zero in expectation, note that the function to be modeled can be expressed as presented in (1).

As discussed in Koenker (2005), for example, the assumption of errors identically distributed is not a necessary condition for adjusting the quantile regression. Unlike the classical regression methodology, the quantile regression models can incorporate heteroscedasticity information from independent random errors.

Once the concept of quantile regression has been defined, it is necessary to understand how to interpret the model coefficients. Consider, for example, that $f(\boldsymbol{x}_i, \tau) = \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau)$ for τ fixed. In this case, the interpretation of the parameters $\beta(\tau)$ is essentially the same as the linear model, being the rate of change. That is, the coefficient $\beta_j(\tau), j = 1, \ldots, p$, can be interpreted as the rate of change in the τ quantile of the variable Y by varying in a unit the value of the jth covariate while maintaining the values of other variables fixed. This is, $\beta_j(\tau) = \partial Q_{Y|\boldsymbol{x}}(\tau) / \partial x_j$.

Geographically Weighted Regression

Geographically Weighted Regression (GWR) is an extension of the classical linear regression for the analysis of non-stationary spatial data. The model allows its parameters to vary spatially, without limiting the form of this variation. The idea of GWR is to make a local adjustment for each point in the study region based on the closest observations. Thus, a continuous function

$\beta_j(u_i, v_i)$ is created for each parameter, where (u_i, v_i) are the spatial coordinates of the *i*th point. The objective of GWR is to provide non-parametric estimates of these continuous surfaces using the kernel function.

The GWR model (Fotheringham et al., 2002) is presented below:

$$y_i = \beta_0(u_i, v_i) + \sum_j \beta_j(u_i, v_i)x_{ij} + \epsilon_i \quad (2)$$

$$\epsilon_i \sim N(0, \sigma^2).$$

Note that the assumptions of the classical regression model (Normal, homoscedastic and uncorrelated errors) remain. However, by allowing spatial variation for the parameters, the problems of autocorrelation and heteroscedasticity are reduced. The persistent limitation is normality, so this model is not yet the most suitable for treating spatial counting data, for example. It is interesting to note that classical regression is a special case of GWR. This simplification occurs when there is no spatial variation in the parameters.

Mathematically, $\beta_j(u_i, v_i)$ is estimated in matrix form by:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X}^T\boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(u_i, v_i)\boldsymbol{y} \quad (3)$$

where $\widehat{\boldsymbol{\beta}}$ represents an estimate for $\boldsymbol{\beta}$, and $\boldsymbol{W}(u_i, v_i)$ is an $n \times n$ matrix with elements outside the diagonal equal to zero and diagonal elements representing the geographical weight of each observation at point *i*. Briefly, and defining (u_i, v_i) by (i) the parameters in each row of the matrix of Equation (3) are estimated by:

$$\hat{\beta}(i) = (\boldsymbol{X^T}\boldsymbol{W}(i)\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{W}(i)\boldsymbol{y}$$

where i represents the matrix line of Equation (2) and $\boldsymbol{W}(i)$ is a diagonal matrix of spatial weights $n \times n$ of the form:

$$\boldsymbol{W}(i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix},$$

where w_{in} is the weight given to point n in the calibration of the model for point i.

The estimator of Equation (3) is a weighted least square estimator but does not use a constant weight matrix. The weights in GWR, the values of the weighting matrix **W**, are calculated for each location i. In this way, each locality receives a different weight in the estimation in i, that is, a calibration is made for each point of interest. In this sense, the idea is that the weights are a measure of proximity of the observation to the point of estimation i.

The key point of this technique is the definition of the "circle of inclusion" of observations around point i, or more generally, of the spatial structure. The specified circle has a radius of size h. If h is too large, then almost all data will be included in the estimation of $\beta_j(u_i, v_i)$, making estimates close to the standard linear regression. If h is too small, few observations will be included in the calibration, resulting in $\beta_j(u_i, v_i)$ estimates with large standard errors. Finding the best h size is therefore extremely important in finding the best GWR fit.

The weight characteristic is also relevant in the adjustment, since it can be done in a discrete or continuous way, as discussed by Brunsdon et al. (1998). In the discrete case, to perform the calibration some points are excluded according to some criterion, for example an inclusion circle with radius h, that is, for a given locality i, the weight w_{ik} given to locality k can be:

$$w_{ik} = \begin{cases} 1, & \text{if } d_{ik} < h. \\ 0, & else. \end{cases}$$

where d_{ik} is the distance between i and k. Or another possibility:

$$w_{ik} = \begin{cases} 1, if\ k\ is\ one\ of\ the\ N\ closest\ neighbours\ of\ i, \\ 0, \qquad\qquad else. \end{cases}$$

The continuous case considers that the *k* localities closest to the locality *i* have more weight in the estimation than more distant localities, in addition the continuous form can follow diverse distributions. In the Gaussian case, the w_{ik} weight can be represented by:

$$w_{ik} = exp\left(\frac{-d_{ik}^2}{2h^2}\right).$$

In this situation, the weight value gradually decreases with distance and can be written:

$$w_{ik} = \begin{cases} [1 - (d_{ik}/h)^2]^2, \text{if } d_{ik} < h. \\ 0, \qquad\qquad else. \end{cases}$$

These functions are known as "Kernel functions" or Kernels and are denoted by the letter *K* such as: $w_{ik} = K(d_{ik})$. Note that *h* also defines the degree of influence of each observation. The problem, then, is to estimate the constant *h*, sometimes referred to as Kernel bandwidth or smoothing parameter, which also functions as a variability factor of the weight curve.

It is known that the results of GWR are relatively indifferent to the choice of Kernel function but are highly sensitive to the smoothing parameter of the Kernel function used (Fotheringham et al., 2002). In the more general case, a constant smoothing parameter for all points is efficient if the points are equally spaced. However, where data are not equally spaced (spatially dispersed or when areas have different sizes), a constant smoothing parameter might prove suitable for some, but not all, locations. This is because the estimated parameters may have large standard errors due to the few points used in the calibration, or in extreme cases, the estimation would not be possible due to the lack of variability. Thus, to reduce these problems, it is possible to use a variable smoothing parameter, which allows a large smoothing parameter, where the data is scattered, and a small smoothing parameter, where the data is more abundant.

A solution to determine the smoothing parameter is cross-validation (CV), which was suggested by Cleveland (1979), for the local regression of the form:

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(h)]^2 , \quad (4)$$

where $\hat{y}_{\neq i}(h)$ is the adjusted value for point y_i, omitting the observation *i*. When h becomes the smallest possible, the model is calibrated only in samples near *i* and not *i*. The value that minimizes Equation (4) is the optimal smoothing parameter of the cross-validation method.

It is important to note that the weighted least squares method for the GWR produces biased estimates for the parameters. The bias arises because the model adjusts local regressions assuming that the surface of the parameters is approximately flat in the vicinity of the analyzed regression point, when in fact the parameters probably vary continuously in the space. On the other hand, considering that there is no spatial stationarity, the estimates of the global regression model will be even more biased, since it assumes that the parameter is constant in every study region.

The bias of the GWR estimates, as well as the variance, will depend on the smoothing method. The choice of a very large smoothing parameter gives us an accurate estimate (with less variance) for the parameter, however, when considering more distant points in the calibration of the model, we are introducing bias in this estimation. The other extreme produces opposing results, that is, a small smoothing parameter produces an unbiased estimate, but with more variance, since it is based on a smaller sample size.

However, Staniswalis (1989) shows that under certain conditions (such as limited log-likelihood functions with first, second, and third derivatives also limited, and $b \to 0$ when $n \to \infty$), estimators that maximize the local likelihood, in this case $\widehat{\beta_J}(u_i, v_i)$, are asymptotically normal, non-biased and consistent.

Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO model, originally proposed by Tibshirani (1996), aims to shrink the parameters of a regression model allowing some of them to assume null value. Thus, the technique simultaneously produces the selection of the relevant variables in the model and the estimation of their respective coefficients. The estimates are obtained by a model of minimization of the error in the data subject to a penalty in the norm $L1$ of the coefficients:

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_p}{arg\ min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 \ subject\ to \sum_{j=1}^{p} |\beta_j| \leq a \qquad (4)$$

where a is the adjustment parameter that determines the intensity of the shrinkage. Written in the form of a Lagrangian, equation (4) takes the following form:

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_p}{arg\ min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where parameter $\lambda \geq 0$ becomes the intensity of the shrinkage. Put in matrix format, the model is such that:

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \underset{\widehat{\boldsymbol{\beta}}}{arg\ min} \parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \parallel_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\boldsymbol{\beta}$ is the vector of parameters $p \times 1, \boldsymbol{Y} = (y_i, \ldots, y_n)'$ is the data vector for the dependent variable, $\boldsymbol{X}$ is the matrix $p \times n$ of data from the series of predictors and $\lambda \geq 0$ is the shrinkage parameter.

Thus, the shrinkage parameter λ plays a fundamental role in the model. As λ is reduced to zero or close to zero, $\boldsymbol{\beta}$ reaches a value $\boldsymbol{\beta}^{OLS}$ such that the regularization term becomes insignificant and the parameters estimated by the LASSO method will be equivalent to those obtained by an OLS model. Otherwise, taking $\widehat{\boldsymbol{\beta}}^{LASSO}$ as the vector $p \times 1$ of estimated parameters obtained through the LASSO and $\widehat{\boldsymbol{\beta}}^{OLS}$ as the vector $p \times 1$ of estimated

parameters obtained through the OLS regression, when $\lambda = 0, \widehat{\boldsymbol{\beta}}^{LASSO} = \widehat{\boldsymbol{\beta}}^{OLS}$. As the value of λ increases, the regression parameters are shrunk to the case where only the intercept remains in the model, i.e. all other parameters are shrunk to zero.