Padilla, W.R.; García, J.; Molina, J.M. Knowledge Extraction and Improved Data Fusion for Sales Prediction in Local Agricultural Markets. ***Sensors*** **2019**, 19, 286.

DOI: 10.3390/s19020286

sensors

MDPI

Article

Knowledge Extraction and Improved Data Fusion for Sales Prediction in Local Agricultural Markets †

Washington R. Padilla [1], Jesús García [2,*] and José M. Molina [2]

[1] Research Group Ideia Geoca Quito, Salesian Polytechnic University Engineer Systems, Quito 170131, Ecuador; wpadillaa@ups.edu.ec

[2] Applied Artificial Intelligence Group, Carlos III University, 28270 Madrid , Spain; molina@ia.uc3m.es

* Correspondence: jgherrer@inf.uc3m.es; Tel.: +34-918-561-315

† This paper is an extended version of our paper published in: Padilla, W.R.; Jesús, G.H.; Molina, J.M. Model learning and spatial data fusion for predicting sales in local agricultural markets. In Proceedings of the 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018.

Received: 31 October 2018; Accepted: 8 January 2019; Published: 12 January 2019

Abstract: In this paper, a monitoring system of agricultural production is modeled as a Data Fusion System (data from local fairs and meteorological data). The proposal considers the particular information of sales in agricultural markets for knowledge extraction about the associations among them. This association knowledge is employed to improve predictions of sales using a spatial prediction technique, as shown with data collected from local markets of the Andean region of Ecuador. The commercial activity in these markets uses Alternative Marketing Circuits (CIALCO). This market platform establishes a direct relationship between producer and consumer prices and promotes direct commercial interaction among family groups. The problem is presented first as a general fusion problem with a network of spatially distributed heterogeneous data sources, and is then applied to the prediction of products sales based on association rules mined in available sales data. First, transactional data is used as the base to extract the best association rules between products sold in different local markets, knowledge that allows the system to gain a significant improvement in prediction accuracy in the spatial region considered.

Keywords: predictive analysis; data mining; alternative circuits of commercialization; association rules; time series; spatial prediction; kriging and co-kriging

1. Introduction

Fusion systems allow the integration of heterogeneous sensor data in databases, including knowledge rules, contextual description, external knowledge bases, etc., in order to obtain a general description of real situations. The goal of this process of information fusion is to take the best decisions based on a global view of situations.

The idea behind this work is to find specific patterns in the behavior of consumption of agricultural products: the relationship among the quantities of each product. Information Fusion and Artificial Intelligence (IF/AI) techniques are used to extract association rules for improving the prediction of the consumption of agricultural products. This prediction allows the establishment of better strategies to improve local operations in the network of markets in Ecuador named "CIALCO" (the Spanish acronym for Alternative Circuits of Marketing). This country is crossed by the equator and its territory extends both north and south of latitude zero. This work is centered in the provinces of Tungurahua and Chimborazo. These regions are located in the south and central region of Ecuador, and data were collected on sales of agricultural products in these marketing circuits, created to establish a direct relationship between farmers and consumers.

The final goal is to increase incomes of the people in the Andean region of Ecuador that works on small farmers, to prevent the migration from this area to larger population centers. There are several markets in the CIALCO circuit; the analysis carried out in this work is related to information of people involved in markets of agricultural fairs. Fairs are places where farmers meet periodically to sell products to consumers and conduct their business [1].

The first source of data was provided by the Ministry of Agriculture and Livestock of Ecuador. The dataset contains weekly performance of sales of products that were sold by small farmers located in the Ecuadorian provinces of Tungurahua and Chimborazo in 2014. The data for each fair is defined by the date, the place, and the volume of sales of products. An average of 300 items per week were sold. This research is based on previous work on information fusion and data mining techniques [2] and the ability to extract knowledge from the fusion of information [3]. Some preliminary results were presented by the same authors at the 2018 International Conference of Information Fusion, Cambridge, UK, on 10–13 July 2018 [4]. In this work, a theoretical approach of fusion system is presented to integrate information from different sources (a net of soft sensors deployed in a certain region) able to improve the predictions of sales by exploiting the association relations mined in the available data and compare their impact with respect to other available magnitudes to fuse such as close population and climatic variables. Associative relations mined from available datasets have proven to be a powerful means to discover causal relations useful for understanding the information coming from heterogeneous sources in different domains such as education [5], smart cities [6], or pervasive computing [7].

In this work, which is focused on the agricultural domain, the composition of several geographically dispersed local markets represents a global market. In each local market, the reported sales are dealt as a soft sensor measuring information about sales, giving information for each product and its associative relationships with the rest. Each market generates the following observational data:

- Local Market Information (LMI): geographical position, name of products, number of sellers, quantity of population that uses the local market, etc.
- For each sale and for each product, the quantity of product acquired in each sale.

The second source of data was provided by the National Institute of Meteorology and Hydrology of Ecuador, where climatic information is recorded for each geographic region based on meteorological sensor sources (temperature, precipitation, humidity, etc.). A third source of data considered in this study was provided by the National Institute of Statistics and Census of Ecuador, where information about the estimated population in each region is stored. At the end, the system is composed of a net of geographically dispersed soft sensors able to obtain data from local markets, fused with hard sensors providing climatic data about each market and also the population of cities close to the market.

The rest of this paper is organized as follows: Section 2 presents an overview of works related to agricultural production. Section 3 proposes a new view of the general problem as a hard and soft information fusion and the algorithm to predict agricultural prediction using special information. Section 4 presents the methodology to integrate data mining to generate relevant knowledge to improve the information fusion process. Section 5 presents the analysis of the target scenario, and Section 6 concludes the paper.

2. Related Works

A great effort in sustainable global agricultural production, Sustainable Development Goals from [8], has been carried out in recent years. However, many factors interact, such as extreme weather patterns, rising levels of population and wealth, water scarcity, increases in energy costs, and civil conflicts. Two complementary perspectives have been used to better prepare for disruptions in food supply and global crop market price fluctuations, helping to reduce global food insecurity:

(1) Monitoring system will be able to guarantee timely and accurate information on current food production;

(2) More accurate forecasting techniques for better understanding the key risks in food supply

There are many agricultural monitoring systems on a regional and national level. The main systems analyzed in [9] are:

- The Global Information and Early Warning System (GIEWS) from the Food and Agriculture Organization (FAO) of the United Nations.
- The Famine Early Warning Systems Network (FEWSNET) from the United States Agency for International Development (USAID).
- The Monitoring Agriculture with Remote Sensing (MARS) system from the European Commission (EC).
- CropWatch in China.
- The Crop Explorer service, which provides remote-sensing-based information used by agricultural economists and researchers to predict global crop production.
- GEOGLAM6, a flagship initiative from GEO (Group on Earth Observations), which was endorsed by the G20 in 2011.
- Seasonal Monitor System of the World Food Programme (WFP).
- Anomaly Hot Spots of Agricultural Production (ASAP)9, launched by the Joint Research Centre (JRC) of the European Commission (EC) in June 2017.

Several models and algorithms have been developed to predict the yield of agricultural productions. Prediction uses soil properties and environmental conditions as data to find correlations using several techniques:

- Linear correlation of yield with soil properties and environmental conditions [10];
- Linear methods, especially multiple linear regressions, to predict yields using soil properties [11];
- Nonlinear methods (Artificial Neural Networks- and fuzzy logics) for yield prediction [12,13].

In general, in these kinds of studies, authors work with different factors, mainly soil properties and farm inputs. Several uncontrolled factors could affect agricultural production; therefore, even complex and mathematical models cannot give accurate results. One of the main factors is related to climate variability [14].

In this work, a regional monitoring system of agricultural productions has been developed as a decentralized data fusion process. The main differences with respect to these previous cited works are:

(1) In previous works, authors have tried to forecast the production of some specific products and tried to correlate the amount of production with external conditions. In this paper, we work with several products that could be correlated.
(2) In this work, the monitoring system is modeled as a Data Fusion System, where information from several farms is aligned in time and space and then fused to obtain a global view of production.
(3) Besides an analysis of contextual information (meteorological and close population), a measure of the improvement of using this information is included.
(4) In this work, a "Monitoring Agriculture Fairs System" is proposed which systematically analyzes the relationships among variables based on sales (transactions) in order to improve the accuracy of prediction models for agricultural production.

3. Spatial Predictions Based on Data Fusion

In general, sensors are defined as sources of information from the environment. In real environments, sensors are composed of specific hardware and software dedicated to taking measurements of physical variables (such as location, size, temperature, pressure, etc.). In virtual environments, such as social networks, webs, etc., sensors are specific software tools that extract information from the digital world. These two different types of sensors can be labeled as hardware

sensors (h) and software sensors (s). Both types of sources generate information about the environment, and this definition allows to distinguish between real and digital environments [15]. Hard/soft data fusion has been shown to be an effective approach to improve models and understand situations in different domains. For instance, the work in [6] deals with the texts posted from persons as a distributed social sensor system which can be correlated with physical sensor data (audio, temperature, pollution, etc.) to improve information available for services provided in smart cities.

Sensors **2019**, *19*, x FOR PEER REVIEW 4 of 22

situations in different domains. For instance, the work in [6] deals with the texts posted from persons as a distributed social sensor system which can be correlated with physical sensor data (audio, temperature, pollution, etc.) to improve information available for services provided in smart cities.

Therefore, a sensor can be defined as a general data source in the following way:

$$S_i^t: \text{ Sensor i of type } t, \text{ with } t \in \{h, s\} \quad (1)$$

$$S_i^t: \text{ Sensor i of type } t, \text{ with } t \in \{h, s\} \quad (1)$$

Besides, several sensors of diverse types could be placed together in a platform of sensors to take advantage of measuring several variables in a coordinated way. For example, a meteorological station is composed of several hardware sensors (temperature, atmospheric pressure, altitude, etc.) and may include software sensors (human observations and forecast reports), or a surveillance system could be composed of a set of hard sensors (radar, camera and infrared camera aligned to detect any object in the search field) and may include human inputs from human operators.

Besides, several sensors of diverse types could be placed together in a platform of sensors to take advantage of measuring several variables in a coordinated way. For example, a meteorological station is composed of several hardware sensors (temperature, atmospheric pressure, altitude, etc.) and may include software sensors (human observations and forecast reports), or a surveillance system could be composed of a set of hard sensors (radar, camera and infrared camera aligned to detect any object in the search field) and may include human inputs from human operators.

A sensor platform could be defined in the following way:

$$P_j = \left\{ \left\{ S_{ij}^t \right\}_{i=1}^{m} \right\}_{t=1}^{n} : \text{Platform } j \text{ composed of } n \text{ sensors of different types.} \quad (2)$$

A sensor platform could be defined in the following way:

$$P_j = \left\{ \left\{ S_{ij}^t \right\}_{i=1}^{m} \right\}_{t=1}^{n} : \text{Platform } j \text{ composed of } n \text{ sensors of different types.} \quad (2)$$

In general, platforms should be distributed in the environment to cover a large area, and, in many cases, coverage areas are overlapped to improve detections in the boundaries, as shown in Figure 1. The information captured by platforms is sent to a fusion centre where it is integrated using spatial and temporal information.

In general, platforms should be distributed in the environment to cover a large area, and, in many cases, coverage areas are overlapped to improve detections in the boundaries, as shown in Figure 1. The information captured by platforms is sent to a fusion centre where it is integrated using spatial and temporal information.

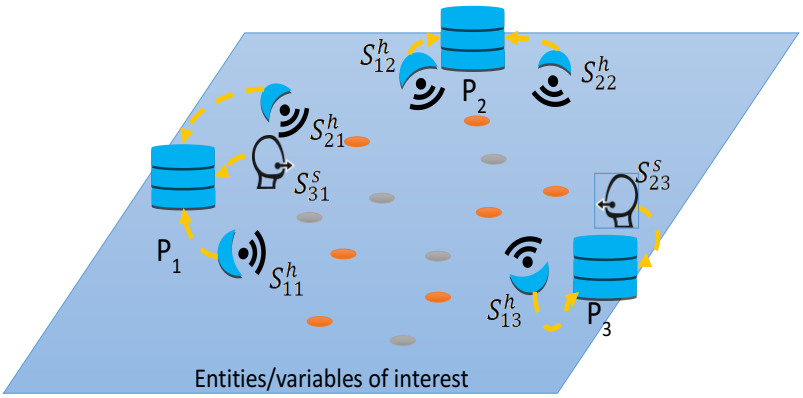

Figure 1. A deployed network including hard and soft sensors in platforms P_1, P_2, and P_3.

Figure 1. A deployed network including hard and soft sensors in platforms P_1, P_2, and P_3.

The fused information is used to predict future situations. This prediction depends on a model that could be based on measurements from a single source, for example current and past positions of a target to predict future positions based on measures and the movement model. In some situations, the information from the set of sensors of the platform could be integrated, and the complementary information of each one allows the final prediction to be improved. For example, the radar position, the visual position, and the thermic position could be combined to generate a "best" position to improve the prediction of future position, or the predicted future position based on radar could be improved considering the actual video position.

The fused information is used to predict future situations. This prediction depends on a model that could be based on measurements from a single source, for example current and past positions of a target to predict future positions based on measures and the movement model. In some situations, the information from the set of sensors of the platform could be integrated, and the complementary information of each one allows the final prediction to be improved. For example, the radar position, the visual position, and the thermic position could be combined to generate a "best" position to improve the prediction of future position, or the predicted future position based on radar could be improved considering the actual video position.

In the fusion centre, information of the network is received, and measures of each sensor are stored to improve the final prediction:

$$I_{ij}^t: \text{ Information from Sensor } i \text{ of platform } j \text{ of type } t. \quad (3)$$

In the fusion centre, information of the network is received, and measures of each sensor are stored to improve the final prediction:

$$I_{ij}^t: \text{ Information from Sensor } i \text{ of platform } j \text{ of type } t. \quad (3)$$

That could be seen as a sequence of values registered in a database, defining each register as a vector with the corresponding values (time stamp, platform j, sensor i, type of sensor t, value). These vectors would be stored in a database, as shown in Table 1.

That could be seen as a sequence of values registered in a database, defining each register as a vector with the corresponding values (time stamp, platform j, sensor i, type of sensor t, value). These vectors would be stored in a database, as shown in Table 1.

Table 1. Observations registered in database.

Time Stamp	Platform	Sensor	Type of Sensor	Value
0.1 s	P_1	S_{11}	Hard	(x,y)
0.2 s	P_2	S_{12}	Hard	(x,y)
0.4 s	P_1	S_{31}	Soft	ID
0.6 s	P_3	S_{31}	Hard	(x,y)
0.7 s	P_3	S_{23}	Soft	ID
0.9 s	P_1	S_{11}	Hard	(x,y)
...	...	...	...	...

The final prediction could be based on the information of only one sensor of the platform, for example, tracking a target using the radar of one platform. In this problem, we have a primary source, that is composed of a deployment with platforms of soft sensors located in each local market. Each platform is defined by the market data: position, number of products, number of sellers, quantity of population that uses the local market, etc., and has a set of soft sensors able to measure the level of sales of each product:

$$I_{ij}^{t}: \text{ Information of level of sales of product } i \text{ from market } j \text{ of type } t. \tag{4}$$

Each platform is composed of two other soft sensors:

- A soft sensor able to measure the climatic floors for each geographic region where the market is located. This is a software process applied over data coming from hard sensors (pressure, humidity, luminosity, etc.):

$$I_{ij}^{t}: \text{ Information of climatic characteristic} \tag{5}$$

  i-th source from market j of type t (e.g., for temperature, humidity, etc.)
- A soft sensor able to measure the population in each one of the cantons of the study.

I_{ij}^{t}: Information of population characteristic i from market j of type t (percentage, absolute, etc.)

Considering this spatially distributed data fusion problem, future predictions could be computed exploiting the geo-statistical properties of variables. The prediction of a variable spatially distributed is a specific problem considering local information. Citing [16], "in the geographical space everything is related to everything, but the closest spaces are more related to each other". This process usually starts by defining the function taking values on a certain spatial region D. This function provides a set of random variables for x taking values in the domain D, being defined as $Z = \{Z(x), x \in D\}$. The values are also random, being the expectation and variance, first and second order moments, defined as usual:

$$m(x) = E[Z(x)] \tag{6}$$

$$\sigma^2(\mathrm{x}) \;=\; \mathrm{var}[Z(x)] \;=\; \mathrm{E}\{[Z(x) - \mathrm{m}(x)]^2\} \;=\; \mathrm{E}[Z(x)^2] - \mathrm{m}(x)^2 \tag{7}$$

This last value assesses the scatteredness of $Z(x)$ around the expectation, while the covariance between different points in D is defined as:

$$C(x_1, x_2) = E[Z(x_1)Z(x_2)] - m(x_1)m(x_2) \tag{8}$$

This value characterizes the interaction between $Z(x_1)$ and $Z(x_2)$, usually integrated in the semi variogram function, defined as:

$$\gamma(x_1, x_2) = \frac{1}{2} var[Z(x_1) - Z(x_2)] \tag{9}$$

which reflects the way in which a point has influence on another point at different distances. The relation between variogram and covariance is given by:

$$\gamma(h) = C(0) - C(h) \tag{10}$$

As defined above, given a set of Z values provided by sensors deployed in n certain sites, $\{x_1, \ldots, x_n\}$, the variogram can be estimated considering the separation vector h as:

$$\gamma(h) = \frac{1}{2} E\left\{ [Z(x+h) - Z(x)]^2 \right\} \tag{11}$$

In the case that $\gamma(h)$ is isotropic (identical in all directions of the space), it does not depend on the angle of vector h, but only on its magnitude, $|h|$.

In the general case, it is difficult to obtain the experimental variogram from data, given the scarce distances and directions, and usually a theoretical model must be adjusted with the available dataset. A quite generic model for variogram considers growing from the origin until a distance for stabilization, around a plateau, so that the random variables *Z(x)* and *Z(x + h)* are correlated when the length of the separation vector h is lower than a certain distance, the zone of influence, and beyond $|h|$ = a the variogram keeps constant (the plateau). For instance, a spherical variogram of reach a and plateau C is defined as [17]:

$$\gamma(h) = \begin{cases} C\left\{ \frac{3}{2}\frac{|h|}{a} - \frac{1}{2}\left(\frac{|h|}{a}\right)^3 \right\} \; if \; |h| \leq a \\ C \;\; otherwise \end{cases} \tag{12}$$

In the isotropic case, assuming independent direction of the semi variance, the vector h is replaced with its magnitude, $\|h\|$. In this case, the variogram is computed taking pairs of data $Z(s_i)$, $Z(s_i + h)$ along the available distances in interval $\overline{h_j}$, defined as:

$$\hat{\gamma}\left(\overline{h_j}\right) = \frac{1}{2N(h)} \sum_{i=1}^{N_h} (Z(s_i) - Z(s_i + h))^2, \; \forall h \in \overline{h_j} \tag{13}$$

In some cases, the experimental variogram shows changes of slope in certain distances. In these cases, the variogram model can be a sum of simple models (nested structures):

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) + \ldots + \gamma_s(h) \tag{14}$$

In these cases, the adjustment is not based only on experimental data, but it must consider also contextual information about the region. More details appear in [18].

Kriging is a linear prediction model corresponding to the unbiased linear estimator. There are different types of kriging depending on the average of population: ordinary and simple.

In ordinary kriging, a stationary Z random function is obtained:

$$\begin{cases} \forall x \in V, \; E[Z(x)] = m \;\; unknown \\ \forall x, x + h \in V, \; cov[Z(x+h), Z(x)] = C(h) \end{cases} \tag{15}$$

where V is the neighborhood considered in the process. The model is defined as:

$$Z^*(x_0) = a + \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha) \tag{16}$$

where x_0 is the place to predict the variable, $\{x_\alpha,\ \alpha = 1, \ldots, n\}$ are the available sites with training data, and $\{\lambda_\alpha,\ \alpha = 1, \ldots, n\}$ are the weights to be computed, together with constant a. A first constraint of estimator is to be unbiased, (mean value of error must be zero):

$$E[Z^*(x_0) - Z(x_0)] = 0 = a + \sum_{\alpha=1}^{n} \lambda_\alpha E[Z(x_\alpha)] - E[Z(x_0)\ ] \ = a + m\left(\sum_{\alpha=1}^{n} \lambda_\alpha - 1\right) \tag{17}$$

With this constraint, the weights to minimize the variance of estimator can be computed as:

$$minimize\ (var[Z^*(x_0) - Z(x_0)]) = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(x_\alpha - x_\beta) + C(0) - 2\sum_{\alpha=1}^{n} \lambda_\alpha C(x_\alpha - x_\beta) \tag{18}$$

Replacing the variogram by the covariance through the relationship $\gamma(h) = C(0) - C(h)$, the kriging prediction is obtained as follows:

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta \gamma(x_\alpha - x_\beta) - \mu = \gamma(x_\alpha - x_0)\ \ \forall \alpha = 1 \ldots n \\ \sum_{\alpha=1}^{n} \lambda_\alpha = 1 \end{cases} \tag{19}$$

The core of the fusion process is carried out with co-kriging, a technique using multiple spatial variables to build an extended prediction model. This multivariable kriging process takes as input a set of m available spatial variables, $\{Z_1, \ldots, Z_m\}$, previously aligned with the variable to predict and collocated in the same coordinates. The prediction equation is extended for this case as follows:

$$Z_i{}^*(x_0) = \mathrm{a} + \sum_{\alpha=1}^{n} \lambda_\alpha Z_i(x_\alpha) + \sum_{j=1, j\neq i}^{m} \sum_{\alpha=1}^{n} \lambda_{\alpha j} Z_j(x_\alpha) \tag{20}$$

It is an estimation method that minimizes the error variance by exploiting the cross-correlation between the available variables. In an analogous way, the error covariance of prediction $Z_i{}^*(x_0)$ can be expressed as a function of the own coefficients, λ_α, and the coefficients for the collocated variables, $\{\lambda_{\alpha j}\}$, $j = 1, \ldots n$, $j \neq i$. Consequently, the expression to minimize, analogous to Equation (17), will depend both on the own variogram, $\gamma_i(h)$, and on the crossed variograms among the variables, $\gamma_{ij}(h)$.

Therefore, the first step of co-kriging also consists in computing the multivariable variogram in order to estimate the semi variances among all variables:

$$\gamma_{ij}(h) = \frac{1}{2} E\{(Z_i(s) - Z_i(s+h))(Z_j(s) - Z_j(s+h))\} \tag{21}$$

This crossed variogram between variables Z_i, Z_j can be estimated with the available training data:

$$\hat{\gamma}_{ij}(h) = \frac{1}{2|\mathrm{N(h)}|} \sum_{N(h)} [Z_i(x_\alpha) - Z_i(x_\beta)]\,[Z_j(x_\alpha) - Z_j(x_\beta)] \tag{22}$$

where the set *N(h)* is defined as $\{\alpha,\ \beta,$ such that $x_\alpha - x_\beta = h\}$, with variables Z_i *and* Z_j taking values in the corresponding locations x_α and x_β.

In this way, a methodology for data integration suggested by Doligez et al. [19] is stepwise, progressively integrating data at different scales to improve the interpolation, illustrated in fusing seismic and well data. This stepwise approach can be used to integrate many data types. Once correlations among variables are analyzed, regression or geostatistical methods are used to integrate the data with the co-kriging approach. On the other hand, hybrid models mixing machine learning and geostatistics, such as Neural Network Residual Kriging/Co-kriging (NNRK/NNRCK) [20], have proven their efficiency in real-world mapping problems.

Building extended models combining individual factors to fit the structure of the data faces the complexity of dimensionality and the risk of oversimplifying the unknown causal relationships of

the multiple factors. Usually, it is best to separate the spatial processes whenever possible and try to understand the causal relationships among the available variables. Therefore, a refined fusion strategy is needed to overcome problems of dimensionality or stationarity assumed in statistical methods, or the interpretability problem of machine learning approaches. In this work, the underlying hypothesis is that the most relevant associative patterns can be mined in transactional data and then exploited to improve the efficiency of multivariable models integrating correlated variables. In general, results must be cross-validated to demonstrate the contribution of this approach. Cross validation in geospatial data implies systematically removing points from the data set and re-estimating the predictions based on the model assessed. This will be the means to validate and assess the contribution of the secondary information resulting from the fusion of available data.

4. Methodology Based on Data Mining to Improve the Fusion Process

The methodology follows three steps: Fusion, Machine Learning, and Prediction. The Fusion System integrates data available from different platforms (local markets) that contain several data sources (market sales, climatic variables, population data, etc.). As indicated, the system can learn relationships between variables that may be useful for the prediction of future sales based on information considering spatial distributions, meteorological information, etc. An overview of the general process is detailed in Figure 2 and explained below:

(1) Clean and Transform Information

1.1. Records with no values are deleted
1.2. Similar values on records are standardized
1.3. Units are defined for all measurements
1.4. The set of products are selected for the study
1.5. Generate spatial structures:

1.5.1. Create common spatial structures (grids for extrapolation)
1.5.2. Transform data to a common coordinate system (the same reference for all sources)

1.6. Final database is generated and prepared for pattern search

(2) Extract best association rules

2.1. The database of sold products is discretized from transactional data
2.2. Association algorithms are applied to mine the best association rules
2.3. Establish the set of products with strongest associations

(3) Estimate future predictions applying geostatistical fusion techniques and validate the hypothesis of strongest conditional dependencies found by data association mining

3.1. Comparison of the fusion results using the set of most associated variables
3.2. Comparison of the results using climatic floors
3.3. Comparison of the results using population
3.4. Comparison of the results using the rest of the variables

(4) Finally, the improvement of the proposal is analyzed by comparing predictions with a single product vs the data fusion results (using residuals as evaluation metrics with Leave-one-out cross-validation, LOOCV)

3.4. Comparison of the results using the rest of the variables

4) Finally, the improvement of the proposal is analyzed by comparing predictions with a single product vs the data fusion results (using residuals as evaluation metrics with Leave-one-out cross-validation, LOOCV)

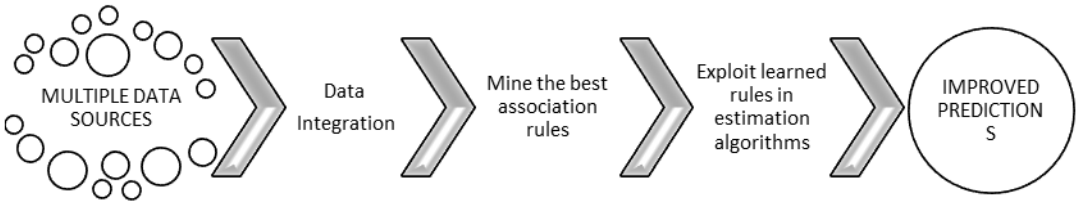

Figure 2. Global process description.

Figure 2. Global process description.

In Figure 2, global process for improving predictions is schematized. The first step is the integration of information received from knowledge sources (every local market) generating global information with every available source. From this information, system obtains the best association rules that should be used in the next step to improve predictions of sales (this is the learned sales model that represent the relation among products). The result of the first step is the fusion of local information received from local markets to generate a global view of sales in a region. In the second step, knowledge is extracted from this global view obtained relevant relationships among products. The third step, the final step, applies this knowledge to improve the future prediction of sales.

In Figure 2, global process for improving predictions is schematized. The first step is the integration of information received from knowledge sources (every local market) generating global information with every available source. From this information, system obtains the best association rules that should be used in the next step to improve predictions of sales (this is the learned sales model that represent the relation among products). The result of the first step is the fusion of local information received from local markets to generate a global view of sales in a region. In the second step, knowledge is extracted from this global view obtained relevant relationships among products. The third step, the final step, applies this knowledge to improve the future prediction of sales.

The global fusion process system needs knowledge extracted from local markets. In this sense, at the beginning of the process, it gathers representative information after applying the machine learning procedure to extract sales knowledge, and then this learned model is used to generate improved predictions by fusing observations and learned model.

The global fusion process system needs knowledge extracted from local markets. In this sense, at the beginning of the process, it gathers representative information after applying the machine

Mining association rules is the way to find causal relations among variables (in the simple way is a rule that find relations between two variables). These rules allow the prediction of changes in the value of a variable based on knowledge of another one. The form of an association rule is $\{A\} \Rightarrow \{B\}$—this rule means: "if A appears in the register then B also should appear in the same register". This kind of rule is useful to identify relationships between categorical attributes that are not explicit. As in any rule system, set A is named antecedent of the rule and B is named consequent. Each association rule should be evaluated to assess each quality, and evaluation is based on three common metrics: support, confidence, and lift [21]:

- *Support* (of a rule) is evaluated as the number of instances (register in the data set) the rule covers related to the whole set (of registers in the dataset).

$$\sup_a(x) = |x|,\ \sup_r(x) = \frac{|x|}{|D|}, \tag{23}$$

where D is the total set of transactions.

If antecedent A and consequent B are considered, the support is the intersection set: tecedent A and consequent B are considered, the support is the intersection set:

$$\sup_a(A \Rightarrow B) = \sup_a(A \bigcap B) \tag{24}$$

- *Confidence* (of a rule) is evaluated as the number (percentage) of times that consequent B appears among the instances that are selected by the antecedent A. The meaning of this concept is the accuracy of its prediction; it is defined as:

$$\text{conf}(A \Rightarrow B) = \frac{sup_a(A \Rightarrow B)}{sup_a(A)} = \frac{|A \bigcap B|}{|A|}. \tag{25}$$

- *Lift* (of a rule) is evaluated as the ratio of observed support considering that A and B were independent:

$$\text{lift} = \frac{\sup(A \Rightarrow B)}{\sup(A)\ \text{x}\ \sup(B)}, \tag{26}$$

There are several algorithms that extract association rule from a database. The most representative algorithm for this task is the Apriori algorithm [22]. Explanations about this algorithm in [23,24] clarify that Apriori algorithm finds trends using performance parameters (support, confidence, and lift) evaluated on "a priori" frequent sets (prior knowledge). The algorithm is composed of the following steps:

(1) Generate all item sets L with a single element; this set is used to form a new set with two, three, or more elements. All possible pairs are taken so that their *support* equals *minsup*
(2) For every frequent item set L' found:

For each subset J, of L'

Determine all association rules of the form:
If L'-$J \rightarrow J$
Select those rules whose *confidence* is greater than or equal to *minconf*

Repeat **1**, including next element into L

As explained, all item sets that satisfy a threshold of minimum support are searched. However, looking for all subsets would not be possible for the exponential size of search space of potential item sets to analyze. The Apriori algorithm prunes candidates with an infrequent subset before counting their supports. This is a Bread-First Search (BFS) process; it ensures that the support values of all subsets of a candidate are known in advance. All candidates of a cardinality k are counted in each scan in order to prune the branches below the support threshold, and then the search descends along the rest in the tree.

A possible alternative approach could be using Depth-First Search (DFS), expanding the candidate sets from the item sets of one of the nodes of the tree. Obviously, scanning the database for every node would result in tremendous overhead, so counting occurrences in a DFS mechanism is not practical. A more recent approach, called FP-growth, has been introduced in [25], and was shown to be more efficient that Apriori in representative situations [26]. In a preprocessing step, FP-growth builds a condensed representation of the transaction data, called FP-tree. FP-growth does not explore all the nodes of the tree, but directly descends to some part of the item sets in the search space and, in a second step, uses the FP-tree to derive the support values of frequent item set.

Besides, other recent extensions of Apriori go in the direction of temporal patterns. This algorithm cannot be used in many applications where patterns vary with time. In this case, entities follow periodic patterns such as transportation on time, load with time constraints, some trajectories, etc., and this kind of problem is not considered by Apriori basic algorithm. This kind of problem should be formulated as discovering patterns from dataset considering temporal attributes and try to model how they vary with time. Many algorithms for finding temporal patterns in sequence databases are listed in the bibliography; these algorithms are usually based on sequence mining techniques (or frequent patterns search) and, at the same time, temporal data association.

There are some extensions of the Apriori algorithm that consider lists of ordered objects using time as items. Then, searched result is the associations of items in the form of sequences of items. Some examples of these algorithms derived from Apriori are Generalized Sequential Pattern (GSP) for spatiotemporal associations [27], Sequential PAttern Discovery using Equivalence classes (SPADE) [28], and Sequential PAttern Mining (SPAM) [29].

Other techniques extract meta-rules describing how relationships vary in time [24,30]. These techniques are also based on APriori mining schema extended to consider time meta-relationships. Some studies conducted in several domains of science have tried to use rules extracted with the Apriori algorithm as a criterion to generate future estimation using associations. Typical works are [31], where the authors analyze the stock of a supermarket, and [32], where authors predict admission decisions by students. Additionally, works such as [33] have searched relationships between extracted rules (association rules) and other techniques such as fuzzy classification.

5. Case Study

As mentioned in the introduction, the global market is composed of several geographically dispersed local markets. Each local market provides sales information for each product in the market. Next, we describe first the preparation of data as the first step before fusion and knowledge extraction for sales forecasting.

5.1. Data Preparation

The dataset of products that are considered in the analysis is listed in Table 2. In this table, names in Spanish (the original dataset) appear in the first column, their corresponding scientific names appear in the second column, and their corresponding translations into English appear in the third column. The process starts validating the original data provided by the governmental office. Non-significant information is deleted, and product names and units of measure are standardized. The dataset provided contains data stored weekly in local fairs of CIALCO (related to the provinces of Chimborazo and Tungurahua, which belong to the central area of the Andes in Ecuador). The initial dataset is subjected to processing for homogenization data (cleansing data). This cleaning mainly affects the names of products, values for unit sales, and the deletion of products that are not relevant for this work.

Table 2. Set of agricultural products.

Spanish Name	Scientific Name	English Name
Acelga	*Beta vulgaris* var. cicla	Chard
Ajo	*Allium sativum*	Garlic
Arveja	*Pisum sativum*	Vetch
Babaco	*Carica pentagona*	Babaco
Brócoli	*Brassica oleracea italica*	Broccoli
Cebolla blanca	*Allium fistulosum*	White onion
Cebolla paiteña	*Allium fistulosum*	Onion
Choclo	*Zea mays*	Corn
Col	*Brassica oleracea*	Cabbage
Col verde	*Brassica oleracea* var. Sabellica	Green cabbage
Coliflor	*Brassica oleracea* var. Botrytis	Cauliflower
Espinaca	*Spinacia oleracea*	Spinach
Frejol	*Phaseolus vulgaris*	Frejol
Frutilla	*Fragaria*	Strawberry
Habas	*Vicia faba*	Broad beans
Hierbas	*Coriandrum sativum, Petroselinum crispum*	Weeds
Lechuga	*Lactuca sativa*	Lettuce
Melloco	*Ullucus tuberosus*	Melloco
Nabo	*Brassica rapa*	Turnip
Papas	*Solanum tuberosum*	Potatoes
Pepinillo	*Cucumis sativus*	Pickle
Pepino	*Cucumis sativus*	Cucumber
Pimiento	*Capsicum annuum*	Pepper
Rábano	*Raphanus sativus*	Radish
Remolacha	*Beta vulgaris*	Beet
Tomate de árbol	*Solanum betaceum*	Tamarillo
Tomate Riñón	*Solanum lycopersicum*	Tomato
Vainita	*Phaseolus vulgaris L*	Vainita
Zanahoria	*Daucus carota*	Carrot
Zapallo	*Cucurbita maxima*	Squash

The available data has been prepared at two levels in order to apply the methodology described. On the one hand, the sales values reported for each individual product have been aggregated in weeks and locations in order to build the prediction models, and also have been aligned in space and time for

fusion in a multivariable model. The tables prepared for each product contain the weekly variation of sales in each spatial location in the 48 weeks of available data.

On the other hand, the original data sheet contains the recorded transactions organized in packages named "canastas" (baskets), each one representing a sale in the local market containing a subset of products in the fair. These are the products contained in each purchase, used to prepare a base table for the search of association rules performed to discover the variables with strongest association and use in the fusion model.

5.1.1. Spatial Data

Table 2 is complemented in Table 3, with seventh and eighth columns that integrate the X and Y coordinates of each transaction. X-Y coordination is related to the commercialization area that corresponds to the provinces of Tungurahua and Chimborazo in the month of July 2014. X and Y represent the latitude and length of local fairs (a position for each local fair) that participate in this work, respectively. Additionally, a fourth column is integrated corresponding to sales values that reflect the behavior of commercialization of kidney tomato.

Table 3. Sales data and spatial location.

FAIR	BROCCOLI	WHITE ONION	TOMATO	TAMARILLO	CARROT	X	Y
Colta	6	60	26	30	30	−78.7238	−1.888
Tisaleo	37.9	36	88.25	185	47.5	−78.6923	−1.40951
Riobamba1	8	9.6	18	0	0	−78.6737	−1.66153
Riobamba2	26	42.5	60	15	17.5	−78.673	−1.66089
Riobamba3	0	10	8	8	5	−78.6687	−1.66025
Riobamba4	2.5	74	60	35	29.5	−78.6588	−1.67626
Cevallos	9	9.6	51	45	20	−78.6566	−1.25714
Riobamba5	7.2	50	30	35	20	−78.6558	−1.6871
Riobamba6	6	65	42	27	29	−78.6538	−1.67599
Riobamba7	2	0	10	0	5	−78.65	−1.67803
Riobamba8	6	26.5	33	0	36	−78.6497	−1.67468
Riobamba9	2.5	0	6.9	0	3	−78.6463	−1.68942
Chambo	21	50	30	20	20	−78.6077	−1.7303
Pillaro	20.1	141.5	92	78	40	−78.551	−1.32836

Sensors 2019, 19, x FOR PEER REVIEW 12 of 22
Cevallos 9 9.6 51 45 20 −78.6566 −1.25714
Riobamba5 7.2 50 30 35 20 −78.6558 −1.6871
Riobamba6 6 65 42 27 29 −78.6538 −1.67599
Riobamba7 2 0 10 0 5 −78.65 −1.67803
Riobamba8 6 26.5 33 0 36 −78.6497 −1.67468
Riobamba9 2.5 0 6.9 0 3 −78.6463 −1.68942
Chambo 21 50 30 20 20 −78.6077 −1.7303
Pillaro 20.1 141.5 92 78 40 −78.551 −1.32836

After this process, the dataset is transformed converting sales data into spatial type structures using latitude and longitude, as shown in Figure 3.

```
> str(FJespacial)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
  ..@ data       :'data.frame': 14 obs. of  8 variables:
  .. ..$ Feria          : chr [1:14] "Colta" "Tisaleo" "RIOBAMBA1" "RIOBAMBA2" ...
  .. ..$ BROCOLI        : num [1:14] 6 37.9 8 26 0 2.5 9 7.2 6 2 ...
  .. ..$ CEBOLLA.BLANCA : num [1:14] 60 36 9.6 42.5 10 74 9.6 50 65 0 ...
  .. ..$ TOMATE         : num [1:14] 26 88.2 18 60 8 ...
  .. ..$ TOMATE.DE.ARBOL: int [1:14] 30 185 0 15 8 35 45 35 27 0 ...
  .. ..$ ZANAHORIA      : num [1:14] 30 47.5 0 17.5 5 29.5 20 20 29 5 ...
  .. ..$ Poblacion      : num [1:14] 44.7 10.6 193.3 193.3 193.3 ...
  .. ..$ pclimatico     : int [1:14] 192 192 86 86 86 86 70 86 86 86 ...
  ..@ coords.nrs : int [1:2] 7 8
  ..@ coords     : num [1:14, 1:2] -78.7 -78.7 -78.7 -78.7 -78.7 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
  .. .. ..$ : chr [1:2] "x" "y"
  ..@ bbox       : num [1:2, 1:2] -78.72 -1.89 -78.55 -1.26
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:2] "x" "y"
  .. .. ..$ : chr [1:2] "min" "max"
  ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
  .. .. ..@ projargs: chr NA
```

Figure 3. Data structure spatial type.

5.1.2. Climatic Zones

The National Institute of Meteorology and Hydrology of Ecuador [34] provides public information related to the climatic floors for several zones in Ecuador (Figure 4). The information provided is a value that corresponds to a different climate category. Each category is associated to a particular interval of variables sensed in the ground network of meteorological stations: humidity,

```
.. ..- attr(*, "dimnames")=List of 2
.. .. ..$ : chr [1:2] "x" "y"
.. .. ..$ : chr [1:2] "min" "max"
..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
.. .. ..@ projargs: chr NA
```


Figure 3. Data structure spatial type.

3.1.2. Climatic Zones

The National Institute of Meteorology and Hydrology of Ecuador [34] provides public information related to the climatic floors for several zones in Ecuador (Figure 4). The information provided is a value that corresponds to a different climate category. Each category is associated to a particular interval of variables sensed in the ground network of meteorological stations: humidity, temperature, precipitation rate, etc., so that predefined intervals define each climate category. In the area that corresponds to the 14 fairs, there is a type of mesothermic climate with an average temperature of 18 °C throughout the year, average precipitation of 650 mm, humidity index that is between −16.5 and 10, and potential evaporation with values that vary between 64 and 106 cm.

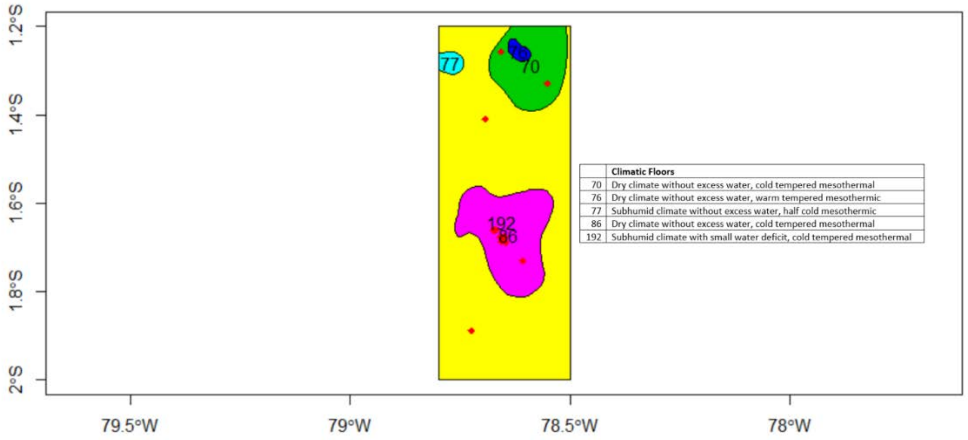

Figure 4. Climatic zones.

The variation of the climatic variables (Figure 5) does not present a representative slope; the one that contributes for this study is the humidity, with a correlation with the tomato variable.

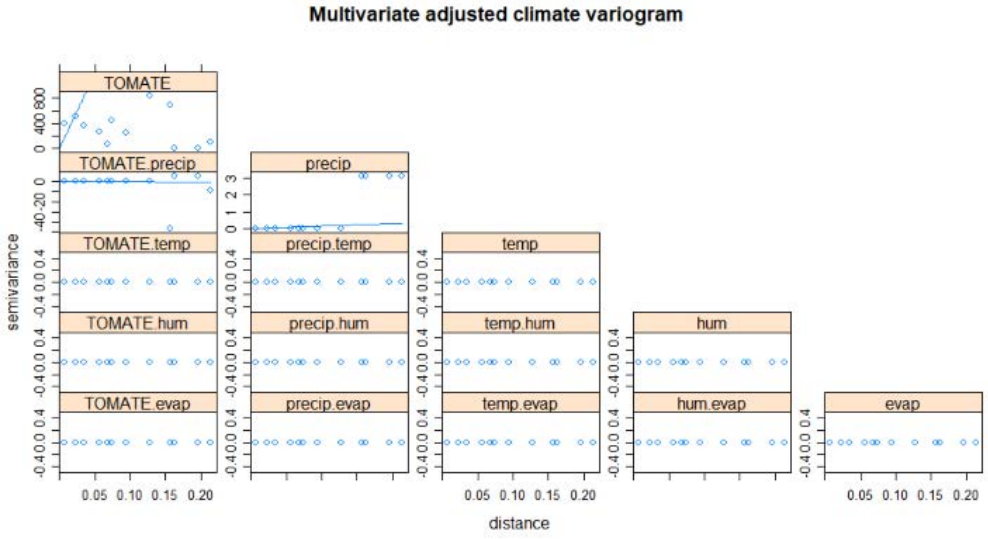

Figure 5. Variograms of climatic variables.

Another important source of information is population. Population dataset is public information that is provided by the Ecuadorian Institute of Statistics and Censuses. In particular, population information related to the cantons considered in this work can be found in [35].

Finally, Table 4 is generated using the information of the presented datasets where coordinates are transformed to common references. Table 4 is a unified file where sales information, population, and climate zones are integrated to build the multivariable prediction model.

Table 4. Merged file with sales and locations per product.

FAIR	BROC COLI	WHITE ONION	TOMAT O	TAMAR ILLO	CARROT	X	Y	POPULA TION	PCLIMAT ICO
Colta	6	60	26	30	30	–78.7238	–1.888	44.701	192

Table 4. Merged file with sales and locations per product.

FAIR	BROCCOLI	WHITE ONION	TOMATO	TAMARILLO	CARROT	X	Y	POPULATION	PCLIMATICO
Colta	6	60	26	30	30	–78.7238	–1.888	44.701	192
Tisaleo	37.9	36	88.25	185	47.5	–78.6923	–1.40951	10.565	192
RIOBAMBA1	8	9.6	18	0	0	–78.6737	–1.66153	193.315	86
RIOBAMBA2	26	42.5	60	15	17.5	–78.673	–1.66089	193.315	86
RIOBAMBA3	0	10	8	8	5	–78.6687	–1.66025	193.315	86
RIOBAMBA4	2.5	74	60	35	29.5	–78.6588	–1.67626	193.315	86
Cevallos	9	9.6	51	45	20	–78.6566	–1.25714	6.8730	70
RIOBAMBA5	7.2	50	30	35	20	–78.6558	–1.68710	193.315	86
RIOBAMBA6	6	65	42	27	29	–78.6538	–1.67599	193.315	86
RIOBAMBA7	2	0	10	0	5	–78.6500	–1.67803	193.315	86
RIOBAMBA8	6	26.5	33	0	36	–78.6497	–1.67468	193.315	86
RIOBAMBA9	2.5	0	6.9	0	3	–78.6463	–1.68942	193.315	86
CHAMBO	21	50	30	20	20	–78.6077	–1.7303	10.541	86
Pillaro	20.1	141.5	92	78	40	–78.551	–1.32836	34.925	70

5.2. Association Rules Mining

As mentioned, the initial file with transactions was analyzed in order to search significant association rules. For each transaction in the available dataset, each agricultural product was checked to determine whether it took part in the sale or not ("t" denotes True and "f" denotes False). However, the negative cases were removed from training set in order to concentrate only on "positive" rules, i.e., relations among products sold together, and avoid rules including absent products in the relationships.

After that, association rules were searched with algorithms available in Weka platform [36]. With a preprocessed training set containing 549 transactions and subsets of 31 items purchased was analyzed to search association rules, setting as parameters a support value greater than or equal to 0.4 (220 transactions in the training set).

The Apriori algorithm was applied to the dataset containing more than 500 transactions on a weekly basis. It was configured with parameters of minimum support set to 0.4 (220 occurrences) and a confidence value set to 0.8. The main findings were:

- The strongest association was found between white onion and tomato products, with a confidence of 87%. The next strongest were the association rules for tamarillo (86%), carrot (83%), and broccoli (82%). These products were selected as the multivariable set, as indicated in Figure 6.
- The product with the highest commercial ratio of the study sample is tomato.

As shown in Figure 7, these results were corroborated by applying the FPGrowth algorithm, which found the same five most relevant association rules among agricultural products in the transactions data sheet.

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.4 -S –1.0 -c –1
Relation:     eneroDiciembre-weka.filters.unsupervised.attribute.Remove-R1
Instances:    549
Attributes:   31
              Acelga
              Ajo
              Arveja
              Babacos
              Brocoli
              Cebolla Blanca
              Cebolla Paiteña
              Choclo
              Col
              Col Verde
              Coliflor
              Espinaca
              Frejol
              Frutilla
              Habas
              Hierbas
              Lechuga
              Melloco
              Nabo
              Papas
              Paquetes de Legumbres
              Pepinillo
              Pepino
              Pimiento
              Rabano
              Remolacha
              Tomate de arbol
              Tomate Riñon
              Vainita
              Zanahoria
              Zapallo
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.4 (220 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 12
Generated sets of large item sets:
Size of set of large item sets L(1): 14
Size of set of large item sets L(2): 15
Best rules found:
 1. Cebolla Blanca=t 338 ==> Tomate Riñon=t 293      <conf:(0.87)> lift:(1.1) lev:(0.05) [27] conv:(1.58)
 2. Tomate de arbol=t 312 ==>Tomate Riñon=t 268   <conf:(0.86)> lift:(1.09) lev:(0.04) [23] conv:(1.49)
 3. Zanahoria=t 374 ==> Tomate Riñon=t 311      <conf:(0.83)> lift:(1.06) lev:(0.03) [17] conv:(1.26)
 4. Brocoli=t 327 ==> Tomate Riñon=t 269      <conf:(0.82)> lift:(1.05) lev:(0.02) [12] conv:(1.19)
 5. Col=t 282 ==> Tomate Riñon=t 230      <conf:(0.82)> lift:(1.04) lev:(0.02) [8] conv:(1.14)
```

Figure 6. Main association rules found among agricultural products with the Apriori algorithm.

Figure 6. Main association rules found among agricultural products with the Apriori algorithm.

```
=== Run information ===
Scheme:       weka.associations.FPGrowth -P 2 -I –1 -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.4
Relation:     eneroDiciembre-weka.filters.unsupervised.attribute.Remove-R1
Instances:    549
Attributes:   31
              Acelga
              Ajo
              Arveja
              Babacos
              Brocoli
              Cebolla Blanca
              Cebolla Paiteña
              Choclo
              Col
              Col Verde
              Coliflor
              Espinaca
              Frejol
              Frutilla
              Habas
              Hierbas
              Lechuga
              Melloco
              Nabo
              Papas
              Paquetes de Legumbres
              Pepinillo
              Pepino
              Pimiento
              Rabano
              Remolacha
              Tomate de arbol
              Tomate Riñon
              Vainita
              Zanahoria
              Zapallo
=== Associator model (full training set) ===
FPGrowth found five rules (displaying top five):
1. [Cebolla Blanca=t]: 338 ==> [Tomate Riñon=t]: 293    <conf:(0.87)> lift:(1.1) lev:(0.05) conv:(1.58)
2. [Tomate de arbol=t]: 312 ==> [Tomate Riñon=t]: 268    <conf:(0.86)> lift:(1.09) lev:(0.04) conv:(1.49)
3. [Zanahoria=t]: 374 ==> [Tomate Riñon=t]: 311    <conf:(0.83)> lift:(1.06) lev:(0.03) conv:(1.26)
4. [Brocoli=t]: 327 ==> [Tomate Riñon=t]: 269    <conf:(0.82)> lift:(1.05) lev:(0.02) conv:(1.19)
5. [Col=t]: 282 ==> [Tomate Riñon=t]: 230    <conf:(0.82)> lift:(1.04) lev:(0.02) conv:(1.14)
```

Figure 7. Main association rules found among agricultural products with the FPGrowth algorithm.

5.3. Spatial Prediction

Spatial prediction was performed first using only the tomato variable as a benchmark, carried out by means of the kriging method with an individual variable. Then, several subsets of the available variables were used in composite structures for a fusion process, based on multivariable co-kriging, in order to perform the predictions and compare with the single-variable situation in order to analyze the relative gain.

order to perform the predictions and compare with the single-variable situation in order to analyze the relative gain.

Sensors **2019**, *19*, x FOR PEER REVIEW 17 of 22

5.3.1. Kriging Prediction

5.3.1. Kriging Prediction

The available libraries of R studio [37] were applied to process the available spatial data. The first step required is building a grid (or mesh) in order to fix the prediction area, setting certain parameters to describe the structure: cell size set to 0.05, offset = (−79.1085, y = −2.531218), dim x = 21; y = 32. In the central sector of the region, one degree of length is equivalent to 111.32 km, and the distance occupied by the two provinces, in the horizontal interval of 116 km length (−79.133499, −78.0834991) corresponds to 1.049 degrees. Red points in Figure 8 indicate the locations of fairs, and the grid was adjusted to cover the limits of the provinces, with the geographical coordinates (longitude, latitude) indicated in the axes.

The available libraries of R studio [37] were applied to process the available spatial data. The first step required is building a grid (or mesh) in order to fix the prediction area, setting certain parameters to describe the structure: cell size set to 0.05, offset = (−79.1085, y = −2.531218), dim x = 21; y = 32. In the central sector of the region, one degree of length is equivalent to 111.32 km, and the distance occupied by the two provinces, in the horizontal interval of 116 km length (−79.133499, −78.0834991) corresponds to 1.049 degrees. Red points in Figure 8 indicate the locations of fairs, and the grid was adjusted to cover the limits of the provinces, with the geographical coordinates (longitude, latitude) indicated in the axes.

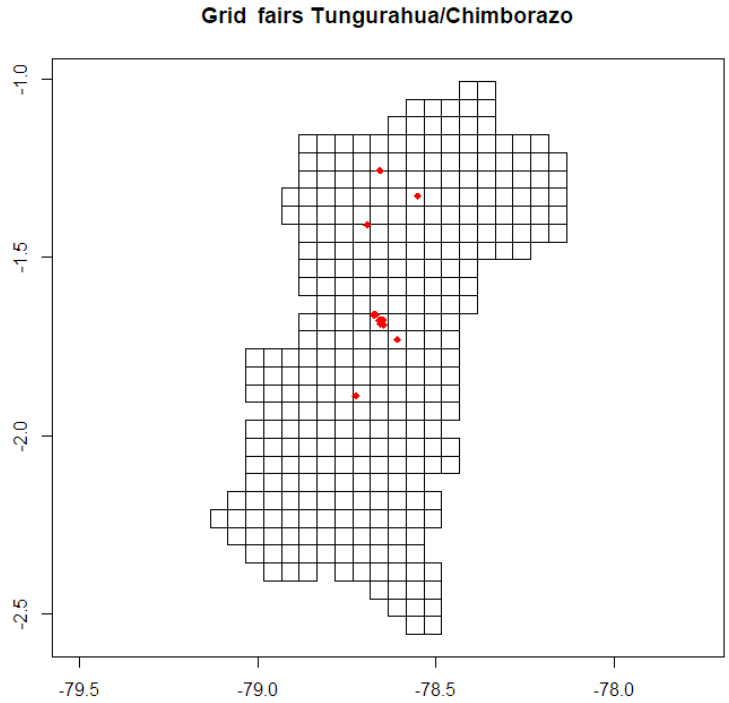

Figure 8. Spatial grid for Tungurahua and Chimborazo provinces.

Figure 8. Spatial grid for Tungurahua and Chimborazo provinces.

The second part consisted of generating the prediction function for the study domain, generated with the function based on the adjustment to the empirical variogram with a model of spherical type. The sale estimate of tomato for the month of July 2014 across the region containing the two provinces considered is shown in Figure 9.

The second part consisted of generating the prediction function for the study domain, generated with the function based on the adjustment to the empirical variogram with a model of spherical type. The sale estimate of tomato for the month of July 2014 across the region containing the two provinces considered is shown in Figure 9.

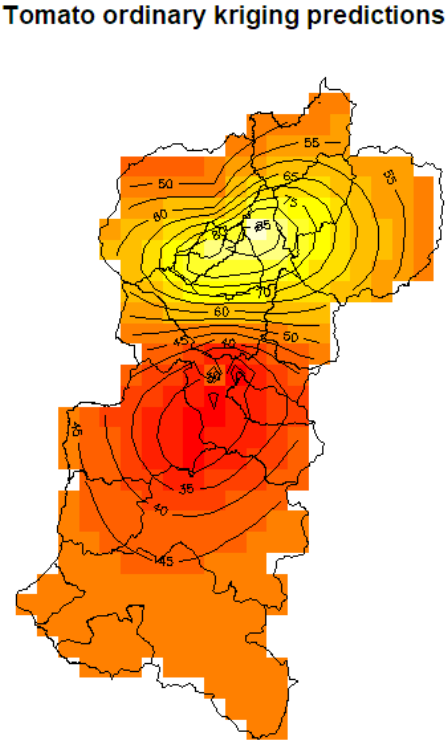

Figure 9. Tomato prediction with ordinary kriging.

5.3.2. Co-kriging Prediction

The prediction of product sales resulting from information fusion was done using the co-kriging method (multivariable kriging) described above, using the available variables aligned in the spatial grid. The correlations and associated variograms are shown in Figure 10, including the variogram models fitted with available data and used in the prediction process. As can be seen, the available data is scarce, so some deviations appear from theoretical model. A full model would require much data to represent all spatial properties in different directions, so the usual solution is a trade-off to fit standard robust models and avoid numerical problems in the spatial regression model.

The analysis of information fusion impact on the results was done by comparing the effects on prediction accuracy, that is, if the errors decrease in predictions. The predictions for tomato sales were computed with a fusion model considering the other variables available in the study, such as population, meteorological variables, as well as the set of products present in the strongest association rules.

LOOCV was applied to compare the different predictions (using the residuals). LOOCV has minimum bias because the training set contains almost the whole dataset. As discussed in [38], LOOCV can achieve the minimum variability both in bias and variance with stable learning schemes as linear regression, and it is usually the selected choice for comparative analysis when the available dataset has limited size.

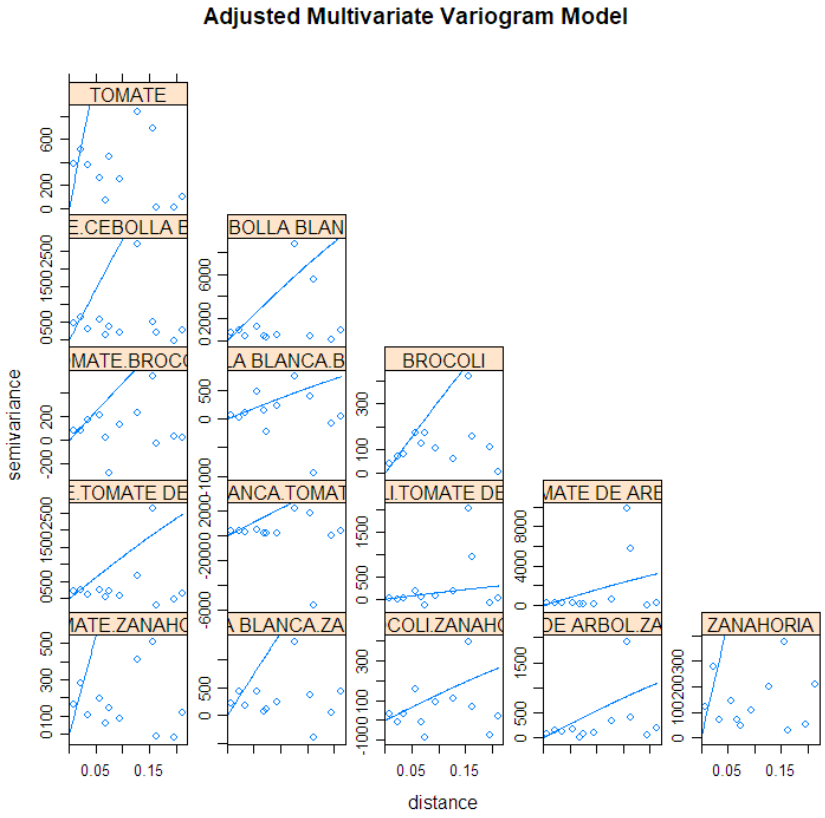

Figure 10. Fitted multivariate variograms of products with strong association.

5.4. Discussion of Results

Six sets (Figure 11 and Table 5) of data were defined because of future prediction (residuals); these are: inverse distance weighted (IDW) interpolation, which allows us to know a basic value; tomato only (OT), which sets the prediction to future (residual) using kriging techniques; TPop, which is tomato and population with co-kriging; Tomato and Precipitation (Prec), a cokriging technique; TAR, which means tomato association rules; and TAll, which means tomato and all the previous variables. The values of the cross validation for each of the data groups behave similarly within the range of the expected results. We can notice that there is an improvement (residual) between using IDW techniques, calculation with a single variable and multivariate calculation (co-kriging). When presenting the consolidated information, in all cases the fusion of additional information has an impact with respect to the case with a single variable. Including the population and precipitation variables to perform the multivariable future estimate does not decrease the value of the estimated error (residual). Therefore, the fusion of information with the association rules allows a more significant improvement in this prediction process. The fusion of information between the tomato variable and the precipitation (Prec) even presents an increase in the maximum and minimum interval of the residuals. Once the direct relationship is established in the improvement of the future estimation in the marketing of the tomato based on the products resulting from the best association rules (TAR), the data of precipitation and population are added to the information fusion (TAll), establishing an improvement in the value of the average of the residuals.

The decrease in the residual values allows us to infer that the fusion of data from the products with strongest association rules reduces the white noise and improves the predictions of the marketing of the tomato.

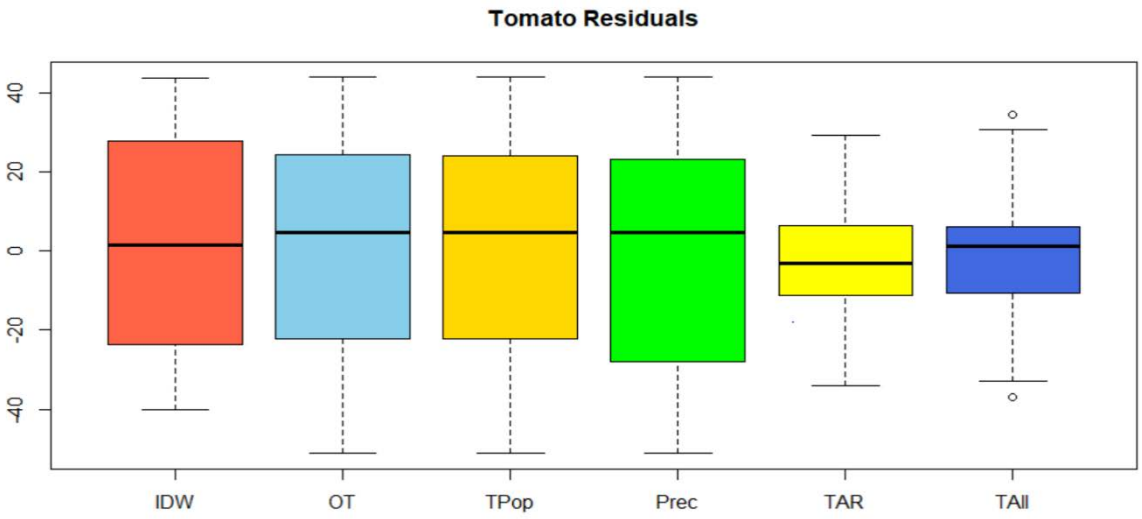

Figure 11. Cross validation of residuals.

Table 5. Root mean square error (RMSE).

Process	Min	1Q	Median	Mean	3Q	Max	RMSE
IDW	40.1610	21.955	1.669	2.259	23.613	41	27.804
Tom (Kriging)	−51.2331	−21.1409	4.8034	−0.0194	23.5698	43.9355	29.440
Tom/Population (cK)	−51.2289	−11.1405	−3.151	−0.9209	6.2992	29.2215	16.550
Tom/Prec (cK)	−51.2170	−26.465	4.87	−0.731	23.097	43.944	29.320
Tom/AR (cK)	−34.1243	−11.1405	−3.151	−0.9209	6.2992	29.2215	16.550
Tom/All Var (cK)	−36.8303	−9.4598	1.3216	0.1151	5.5997	34.6041	19.420

6. Conclusions

The focus of this paper was centred on the use of transactional information about market sales of small farmers to analyze dependencies among products in order to improve the accuracy of predictions. Dependency was learned from individual sales and recorded and stored in a database to mine association rules that measure the dependency among products. These rules show which products should be considered jointly to predict future values in an extended model. Additionally, contextual information was added and fused with the previous information to analyze the improvement of predictions. The research was focused on a target area for analysis located in two provinces of Ecuador, with available data from alternative marketing circuits.

Besides, a general fusion methodology based on variable selected from association rules mining was applied to improve the predictions. With this improvement in the sales prediction process it would be possible to establish scenarios for consumption maps in order to take strategic decisions aimed at improving the economic income of farmer families. Some limitations have been identified, namely the moderate size of available data and low quality of meteorological variables to integrate, pointing to future directions of research and resources needed to increase the size and quality of data to continue this work.

Author Contributions: W.R.P.: research methodology, data analysis and implementation; J.G.: supervision, design of experimental analysis, analysis of results; J.M.M.: supervision, system architecture and analysis of techniques.

Funding: This work was supported in part by Project MINECO TEC2017-88048-C2–2-R, Salesian Polytechnic University of Quito-Ecuador and by Commercial Coordination Network, Ministry of Agriculture and Livestock, Ecuador.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Padilla, W.R.; García, H.J. CIALCO: Alternative marketing channels. *Commun. Comput. Inf. Sci.* **2016**, *616*, 313–321.
2. Padilla, W.R.; García, J.; Molina, J.M. Improving Forecasting Using Information Fusion in Local Agricultural Markets. In *International Conference on Hybrid Artificial Intelligence Systems*; Springer: Cham, Switzerland, 2018; pp. 479–489.
3. Padilla, W.R.; García, J.; Molina, J.M. Information Fusion and Machine Learning in Spatial Prediction for Local Agricultural Markets. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*; Springer: Cham, Switzerland, 2018; pp. 235–246.
4. Padilla, W.R.; Garcia, J.; Molina, J.M. Model Learning and Spatial Data Fusion for Predicting Sales in Local Agricultural Markets. In Proceedings of the 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018.
5. Gomede, E.; Gaffo, F.H.; Briganó, G.U.; Miranda, R.; de Souza, L. Application of Computational Intelligence to Improve Education in Smart Cities. *Sensors* **2018**, *18*, 267. [CrossRef] [PubMed]
6. Costa, D.G.; Duran-Faundez, C.; Andrade, D.C.; Rocha-Junior, J.B.; Just, J.P. TwitterSensing: An Event-Based Approach for Wireless Sensor Networks Optimization Exploiting Social Media in Smart City Applications. *Sensors* **2018**, *18*, 1080. [CrossRef] [PubMed]
7. Miori, V.; Russo, D.; Concordia, C. Meeting People's Needs in a Fully Interoperable Domotic Environment. *Sensors* **2012**, *12*, 6802–6824. [CrossRef] [PubMed]
8. United Nations. Transforming our World: The 2030 Agenda for Sustainable Development. Available online: https://sustainabledevelopment.un.org/post2015/transformingourworld/publication (accessed on 15 July 2015).
9. Fritza, S.; Seea, L.; Bayasa, J.C.L.; Waldner, F.; Jacquesc, D.; Becker-Reshefd, I.; Whitcraftd, A.; Baruthe, B.; Bonifaciof, R.; Crutchfieldg, J.; et al. A comparison of global agricultural monitoring systems and current gaps. *Agric. Syst.* **2019**, *168*, 258–272. [CrossRef]
10. Gemtos, T.A.; Markinos, A.; Nassiou, T. Cotton Lint Quality Spatial Variability and Correlation with Soil Properties and Yield. In Proceedings of the 5th European Conference on Precision Agiculture, Uppsala, Sweden, 9–12 June 2005.
11. Wendroth, O.; Jurschik, P.; Giebel, A.; Nielsen, D.R. Spatial Statistical Analysis of On-site Crop Yield and Soil Observations for Site-specific Management. In Proceedings of the Fourth International Conference on Precision Agriculture, St. Paul, MN, USA, 19–22 July 1998.
12. Papageorgiou, E.I.; Aggelopoulou, K.D.; Gemtos, T.A.; Nanos, G.D. Yield Prediction in Apples Using Fuzzy Cognitive Map Learning Approach. *Comput. Electron. Agric.* **2013**, *91*, 19–29. [CrossRef]
13. Papageorgiou, E.I.; Markinos, A.T.; Gemtos, T.A. Fuzzy Cognitive Map Based Approach for Predicting Yield in Cotton Crop Production as a Basis for Decision Support System in Precision Agriculture Application. *Appl. Soft Comput.* **2011**, *11*, 3643–3657. [CrossRef]
14. Browna, J.N.; Hochmanb, Z.; Holzworthc, D.; Horanb, H. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agric. For. Meteorol.* **2018**, *260*, 247–254. [CrossRef]
15. Llinas, J.; Bowman, C.; Rogova, G.; Steinberg, A.; Waltz, E.; White, F. *Revisiting the JDL Data Fusion Model II*; Space and Naval Warfare Systems Command: San Diego, CA, USA, 2004.
16. Celemín, J.P. Autocorrelación espacial e indicadores locales de asociación espacial: Importancia, estructura y aplicación. *Rev. Univ. Geogr.* **2009**, *18*, 11–31.
17. Tonye, E.; Fotsing, J.; Zobo, B.E.; Tankam, N.T.; Kanaa, T.F.N.; Rudant, J.P. Contribution of Variogram and Feature Vector of Texture for the Classification of Big Size SAR Images. In Proceedings of the Seventh International Conference on Signal Image Technology Internet-Based Systems, Dijon, France, 28 November–1 December 2011.
18. Bivand, R.S.; Pebesma, E.; Gómez-Rubio, V. *Applied Spatial Data Analysis with R*; Springer: New York, NY, USA, 2013.
19. Doligez, B.; Beucher, H.; Lerat, O.; Souza, O. Use of a Seismic Derived Constraint: Different Steps and Joined Uncertainties in the Construction of a Realistic Geological Model. *Oil Gas Sci. Technol. Rev. IFP* **2007**, *62*, 237–248. [CrossRef]

20. Kanevski, M.; Maignan, M. *Analysis and Modelling of Spatial Environmental Data*; EPFL Press: Lausanne, Switzerland, 2004.
21. Agrawal, R.; Imielinski, T.; Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 25–28 May 1993.
22. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 12–15 September 1994.
23. Zulfikar, W.B.; Wahana, A.; Uriawan, W.; Lukman, N. Implementation of association rules with apriori algorithm for increasing the quality of promotion. In Proceedings of the 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 26–27 April 2016.
24. Chang, C.-C.; Li, Y.-C.; Lee, J.-S. An efficient algorithm for incremental mining of association rules. In Proceedings of the 15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications, Tokyo, Japan, 3–4 April 2005.
25. Han, J.; Pei, J.; Yiwen, Y.; Fraser, S. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Min. Knowl. Discov.* **2004**, *8*, 53–87. [CrossRef]
26. Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. Algorithms for Association Rule Mining—A General Survey and Comparison. *ACM Sigkdd Explor. Newsl.* **2000**, *2*, 58–64. [CrossRef]
27. Agrawal, R.; Srikant, R. Mining sequential patterns. In Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan, 6–10 March 1996.
28. Zaki, M.J. SPADE: An efficient algorithm for mining frequent sequences. *Mach. Learn.* **2001**, *42*, 31–60. [CrossRef]
29. Ayres, J.; Flannick, J.; Gehrke, J.; Yiu, T. Sequential pattern mining using a bitmap representation. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002.
30. Spiliopoulou, M.; Roddick, J.F. Higher Order Mining: Modelling and Mining the Results of Knowledge Discovery. In *WIT Transactions on Information and Communication Technologies*; Ebecken, N.F., Brebbia, C.A., Eds.; WIT Press: Ashurst, UK, 2000.
31. Asadifar, S.; Kahani, M. Semantic association rule mining: A new approach for stock market prediction. In Proceedings of the 2nd Conference on Swarm Intelligence and Evolutionary Computation, Kerman, Iran, 7–9 March 2017.
32. Mane, R.V.; Ghorpade, V.R. Predicting student admission decisions by association rule mining with pattern growth approach. In Proceedings of the International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, Mysuru, India, 9–10 December 2016.
33. Kumar, P.S.V.V.S.R.; Maddireddi, L.R.D.P.; Lakshmi, V.A.; Dirisala, J.N.K. Novel fuzzy classification approaches based on optimisation of association rules. In Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology, Bangalore, India, 21–23 July 2016.
34. Instituto Nacional de Meteorología e Hidrología. Geoinformación Hidrometeorológica. Quito, Ecuador, 2018. Available online: http://www.serviciometeorologico.gob.ec/geoinformacion-hidrometeorologica/ (accessed on 14 March 2018).
35. Instituto Nacional de Estadística y Censos. Población y Demografía; Quito, Ecuador, 2018. Available online: http://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/ (accessed on 14 March 2018).
36. Weka 3—Data Mining with Open Source Machine Learning Software in Java. Available online: http://www.cs.waikato.ac.nz/ml/weka/ (accessed on 28 September 2017).
37. RStudio—Open Source and Enterprise-Ready Professional Software for R. Available online: https://www.rstudio.com/ (accessed on 8 December 2017).
38. Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. *J. Econ.* **2015**, *187*, 95–112. [CrossRef]