agriculture

MDPI

Article

An Approach Based on Web Scraping and Denoising Encoders to Curate Food Security Datasets

Fabián Santos [1,*] and Nicole Acosta [2]

[1] Centro de Investigación Para el Territorio y el Hábitat Sostenible (CITEHS), Universidad Indoamérica, Quito 170301, Ecuador
[2] Research Unit Sustainability and Climate Risks, Universität Hamburg, 20144 Hamburg, Germany; nicole.acosta@studium.uni-hamburg.de
* Correspondence: ernestosantos@uti.edu.ec

Abstract: Ensuring food security requires the publication of data in a timely manner, but often this information is not properly documented and evaluated. Therefore, the combination of databases from multiple sources is a common practice to curate the data and corroborate the results; however, this also results in incomplete cases. These tasks are often labor-intensive since they require a case-wise review to obtain the requested and completed information. To address these problems, an approach based on Selenium web-scraping software and the multiple imputation denoising autoencoders (MIDAS) algorithm is presented for a case study in Ecuador. The objective was to produce a multidimensional database, free of data gaps, with 72 species of food crops based on the data from 3 different open data web databases. This methodology resulted in an analysis-ready dataset with 43 parameters describing plant traits, nutritional composition, and planted areas of food crops, whose imputed data obtained an R-square of 0.84 for a control numerical parameter selected for validation. This enriched dataset was later clustered with K-means to report unprecedented insights into food crops cultivated in Ecuador. The methodology is useful for users who need to collect and curate data from different sources in a semi-automatic fashion.

Keywords: web scraping; denoising autoencoders; plant traits; food security; Ecuador

Citation: Santos, F.; Acosta, N. An Approach Based on Web Scraping and Denoising Encoders to Curate Food Security Datasets. *Agriculture* **2023**, *13*, 1015. https://doi.org/10.3390/agriculture13051015

Academic Editors: Ebrahim Jahanshiri and Sayed Azam-Ali

Received: 22 March 2023
Revised: 29 April 2023
Accepted: 30 April 2023
Published: 6 May 2023

1. Introduction

Ensuring food security is a task that requires the integration of the will of the agricultural and political sectors [1]. Informed decision making depends on access to relevant data in a timely and open manner. Unfortunately, in countries such as Ecuador, policies regarding open data are ambiguous and access to government datasets is commonly restrictive to users. Bureaucratic processes required to obtain permission for the use of data can make the data unsuitable and unreliable for scientific purposes. In addition, the documentation of the methodologies and validation procedures of public data is occasionally incomplete. Therefore, merging these datasets with other non-government data sources is frequently practiced to fill data gaps and validate the original content [2]. As this information is highly dependent on web resources, such a task is often labor-intensive. A case-wise review is required to search, retrieve, and organize the data [3].

Combining different sources is a further challenge. Often, the use of incomplete databases is inevitable, and procedures to impute them can be counterproductive if they are based on mean substitution, constant values, or moving windows [4]. This makes it difficult for researchers and decision makers to obtain reliable information to support food security in the country, which also depends on information from other resources such as water availability [5].

Advances in data-enrichment techniques today offer different approaches to enhance raw data with additional information and make it more valuable and useful for analysis or decision making. Artificial Intelligence (AI) and machine learning are useful, for example,

to identify patterns and relationships in large datasets. On the other hand, natural language processing can extract information from textual sources [6], while geocoding can add geographical location data to enable location-based analysis [7]. Finally, access to a wide range of third-party data sources helps users from a variety of disciplines to create more complete datasets [8–10]. Of all these approaches, it is the latter that is of interest in this research, as it is often assumed that third-party data sources appear ready for download and analysis. However, the reality is that it is sometimes difficult to take advantage of them, due to challenges related to data extraction, data quality, data integration, and data security, among others [11]. Although there are studies documenting procedures for automating web data collection (or web scraping) and data analysis, these focus more on computer science, data science, and business intelligence [12] than on issues such as food safety. For example, a Google academic search gives us 9000 documents when the terms "web scraping" and "business" are combined, but less than 500 documents when the latter is replaced by "food security". Previous studies recognize the importance of automating the process of combining different datasets to obtain relevant information for food security [13]. Nonetheless, most of them automatize the process only partly or focus only on market data [14,15]. Hence, there is a need to investigate and test these new approaches and specialized methods in areas; this is perhaps of less interest to the scientific community but is still undoubtedly very important for decision making in the face of challenges such as climate change.

To cope with this research gap, an approach based on web scraping and denoising autoencoders to curate food security datasets is presented in this work. This allows for automating the extraction of data to process it in a database [16] and to impute its missing data cases with a deep learning-based technique called multiple imputation denoising autoencoders (MIDAS) [17] to estimate specific values to replace the missing ones [18]. This approach can reduce the time needed to enrich the data but also curate it, outperforming other human-aided procedures in data collection [19] and other non-deep learning data imputation techniques [20]. This approach can be of particular use in the context of food security when the aim is to study the relationship between plant traits and the nutritional composition of several crops compiled from different sources, but, when collated, missing data usually impede the analysis. To illustrate how this problem can be addressed with the proposed approach, in this study, the Survey of Surface and Continuous Agricultural Production (ESPAC) of Ecuador [21] was used to combine it with information from two other open data sources. The first one was the ECOCROP web database, a website that provides information on plant characteristics and crop environmental requirements for more than 2000 plant species [22]. The second, called the International Food Composition Database (FUNIBER), is also a website that describes the main nutrients in foods based on different nutritional composition tables for eight countries, including Ecuador [23]. This research aims to produce a multidimensional database, free of data gaps, that helps to cluster food crops based on their nutritional contents and describe their plant traits and planted areas in Ecuador.

To meet the objectives of this work, we first describe a brief theory of the methods applied in the proposed approach and then explain its implementation. The methods' results are then presented using graphs and tables and their advantages and limitations are discussed, along with the issues raised by open data policies. We conclude with some recommendations for users who need to collect and preserve data from different sources in a semi-automatic way.

2. Theoretical Framework

2.1. Web Scraping: Concepts, Methods, and Limitations

Web scraping is defined as the process of extracting data from specific websites to create or enrich databases to enable data analysis. In this way, a wide range of structured, semi-structured, and non-structured data types from web sites can be collected, making it a valuable tool for research and analysis [24]. Although web scraping can help to collect

large amounts of data, one of its weaknesses is that it relies heavily on the selection of data sources (whose selection can also be biased) to ensure that the quality and reliability of the data are acceptable [25]. There are multiple methods for its application, but the most popular are:

- HTML and XML parsing, which uses a software program that reads in the HTML or XML code and extracts specific data elements from a website, such as tags, elements, attributes, and text content [26]. Of the two, XML is more flexible as it is not tied to specific formats as is HTML.
- Application programming interface (API) scraping, which involves accessing the APIs to extract data, usually in a structured format and without the need to develop code to parse HTML [27]. As the number of APIs available is vast, it is recommended to use platforms, such as Postman [28], for designing, testing, and managing APIs.
- Automated browsing is another method, useful for collecting dynamically generated data from a website. It consists of software to navigate a website and extract data [29].

The choice of any of these methods will depend on the goals and requirements of the web-scraping project. However, these tools, as mentioned above, have limitations in terms of data reliability, as well as legal and ethical aspects [30], since copyright can be infringed if the databases used in web scraping are not open or used for malicious purposes. Moreover, technical challenges cannot be ignored as web scraping can be complex, requiring specialized knowledge of programming languages and web technologies.

2.2. Data Imputation and the MIDAS Algorithm

Data imputation is a technique used to replace missing or incomplete data values with estimated values to preserve data integrity. This is a common procedure in data analysis workflows, which assumes that the omission is related to and can be explained by the observed data. Although there are different imputation techniques that replace data gaps with measures of central tendency (i.e., mean, median, and mode), as well as others based on statistical models (e.g., regression models, K-means, and PCA) [31], here we will focus on one of the most novel ones called MIDAS, as it uses deep learning and has been well received by the scientific community, especially for data imputation in multivariate datasets [32–34]. This algorithm is based on denoising autoencoders (DAs) [35,36], which consist of a deep neural network of interconnected nodes organized in layers that are commonly used to remove noise from data by reconstructing them from noisy inputs. In this sense, MIDAS uses autoencoders to impute missing values in a dataset by training a neural network to reconstruct the complete data matrix from the incomplete dataset D. It does this by incorporating D into the neural network through an input layer, processing it through nodes in one or more hidden layers, and returning it through nodes as output layers as M complete datasets.

To initiate this process, the algorithm prepares D for the training phase, "one-hot" encoding categorical variables, and rescaling numerical variables between 0 and 1 to improve the network convergence. Then, a missingness indicator matrix R is built for D to distinguish the missing Dmiss from the observed Dobs data. During the training phase of the algorithm, the next steps are performed:

1. D and R are shuffled and divided into mini-batches;
2. Additional missingness to R is introduced;
3. Half of the nodes in hidden layers are corrupted following the standard dropout implementations [37];
4. A forward pass is performed to initiate the computation of the output values, deriving the reconstruction error on the predictions of the originally observed corrupted values Xobs by employing the root mean square error and cross-entropy loss functions for continuous and categorical variables;
5. The loss values are aggregated into a single term and backpropagate DA, using the resulting error gradients to adjust the weights for the next epoch.

Once the training is complete, D is passed into DA to reconstruct the corrupted values and derive M by replacing Dmiss with predictions of the originally missing values. According to the literature, MIDAS has demonstrated increased accuracy over other algorithms; flexibility for imputing categorical, logical, or numerical data types; reduced bias with noisy data; and scalability [34]. It innovates with respect to other implementations of DA by considering missing values as part of the overall data corruption, minimizing the reconstruction error on the original observed portion. Moreover, it has a reduced risk of overfitting as it introduces the dropout technique during the training phase [17].

However, MIDAS depends on certain circumstances to function effectively, and these depend mainly on the assumptions related to multiple imputation. For example, missing data not at random (MNAR) is a condition in which the missing data mechanism is non-random and it is not predictable from other variables in the dataset [38]. Then, if the data are not MNAR, it is not guaranteed that MIDAS can be biased. Moreover, unconventional data structures such as non-exchangeable data, multilevel data, and spatially lagged data are also potentially not well suited to MIDAS. However, it is said that it may perform better than other algorithms even with these data structures; in addition, given its autoencoder-based architecture, it has been [17] shown to be able to properly impute datasets with missingness levels of 20–40% [39]. Finally, the drawbacks associated with neural network methods, such as complex hyperparameter fitting, the propensity for overfitting (especially in sparse datasets), the dependence on large datasets, limited interpretability of the model, and demanding computational resources, are potentially unavoidable and should be taken into account before using MIDAS [40,41].

3. Materials and Methods

3.1. Agricultural Survey Processing and Construction of Food Crops Database

Since 2014, the National Institute of Statistics and Census of Ecuador (INEC) has published a survey to gather agricultural statistics. Referred to as ESPAC, this dataset consists of 20 individual databases describing land use, cultivation areas, livestock and poultry numbers, and land tenure and work in the production units. Among these, the land use database is the most informative, as it reports the planted areas of 107 food crops, along with other crops related to the agroindustry. For this study, the latest survey, which was conducted in 2021, was used. To process it, the food crop data were filtered out and exported to the R language [42], which was used to implement the entire methodology and to generate plots to represent the data using the *ggplot* and *reshape* R libraries [43,44]. All datasets collected and processed, as well as the scripts developed in this work, are available in a GitHub repository for readers interested in reproducing this approach (see data availability statement).

The ESPAC structure was designed using a sampling frame recommended by the Food and Agriculture Organization (FAO) of the United Nations. Because this survey structure is complex, the *survey* R library was used to conduct the analysis [45]. This tool is designed to compute survey estimates in complex survey designs using two functions. The first one, called *svydesign*, helps to define the survey design by identifying its strata, sampling weights, and finite population correction. In the case of ESPAC, only the sampling weights were available, so we applied them to the *svydesign* function for each available food crop. The second function, called *svyby*, calculates statistics based on subsets and factors. As the ESPAC reference manual warns users that estimates are only valid for the provincial level, the planted areas for each food crop were calculated for each province. To check that the above procedure was correct, first the food crop tables published by INEC as official figures were reproduced. They use data from different databases focusing on 15 permanent crops and 17 non-permanent crops of commercial interest. After obtaining similar values, we proceeded to process the 107 crops included in the land use database. To observe their differences from the INEC official figures, tables of these permanent and non-permanent food crops were extracted according to their codes to match with those derived from the land use database. The two figures were differentiated by first adding their provincial

planted areas to the national totals, assessing their similarities, and noting discrepancies. This was necessary to validate the results because, in the ESPAC documentation, it is not specified whether the land use database is valid for deriving planted areas of food crops.

3.2. Web Scraping the Characteristics of Food Crops, Their Environmental Requirements, and Nutritional Composition

To automate data collection from the web databases, the names of the food crops derived from ESPAC were synonymized with their scientific names. This allowed for searches in the ECOCROP database, as retrieving crop information required this nomination. Since some crops are listed as different species but, according to taxonomy, belong to only one species (e.g., *Colocasia esculenta* was linked to two different crops, i.e., "papa china" and "malanga"), they were merged to follow the taxonomical convention. Similarly, some crops referring to different parts of the plant were also merged to avoid duplicates. After this step, a routine with Selenium software [46] was developed. Selenium is open-source software for automated browsing and web scraping, and has an implementation in the R language, called the *Rselenium* R library [47]. This software requires specifying a browser to start the Selenium server. In our case, Google Chrome version 105.0.5195.19 was used. By specifying the scientific name of the crop and the code used by ECOCROP to index it, it was possible to navigate to the web page where the plant data sheet was stored. Thus, by pasting the root name of the web page, i.e., "https://gaez.fao.org/pages/ecocrop" (accessed on 22 March 2023) and its crop code, a list of links to each crop was obtained. These links were browsed and parsed automatically to extract their XML/HTML content using the *XML* library [48]. XML software includes a function called *readHTMLTable*, which allows for the detection of data components, such as header labels, table names, and columns, to store the information as a database. As the ECOCROP data sheets included 25 different plant parameters, the numerical data (i.e., 9 parameters) were preferred to the nominal data, because the nominal categories were not clearly differentiated in the records and rather merged with other information, which complicated their analysis. Thus, 142 tables were collected for later integration with the ESPAC dataset.

After this processing, the FUNIBER database was examined. Following the procedure mentioned above, a list of links was obtained by pasting the web page root name: "https://www.composicionnutricional.com/foods/view/EC-" (accesed on 23 September 2022) to each of the food item codes described in the Ecuadorian database. These links were browsed and parsed to extract their nutritional content tables, with a total sum of 510 tables. Each of the collected tables was structured in 26 numerical parameters related to nutritional content, of which 17 were valid since the rest had zero or almost zero variance. After compiling and integrating the tables into a single database, the food crops were collated. For averaging the multiple nutrient content tables, their values were averaged using the median, which was chosen as a conservative measure [49] since some nutrient content tables included food items that were cooked or processed, and their values varied.

3.3. Missing Data Imputation with MIDAS of the Data-Enriched Food Crops Database

With the data-enriched food crops database, the next step was to fill in the missing data with the MIDAS algorithm, which has an implementation in the R language called the *rMIDAS* library [50]. To apply it, the enriched food crop database was first formatted, ensuring that the numerical columns were not characters. Other columns referring to duplicate variables and non–informative parameters, such as data identifiers and unstructured labels, were then filtered out. Since it was deemed necessary to ensure the performance of the algorithm, a test parameter was created by corrupting the carbohydrate contents by introducing 43 additional null values from the 8 that occurred in the original dataset. This parameter was selected because of its numerical format and this made it possible to compare the results between the original and corrupted values after the data imputation and to evaluate the algorithm independently of its validation procedure. Following this, the function *convert* was applied. This function allows for the reshaping of the database and

prepares it for the training phase, showing which parameters were "one-hot" encoded or scaled. To build and run the MIDAS algorithm on the supplied missing data, the function *train* allows us to control this process. The calibration parameters that were not defined by default were:

- Training epochs, which defines the number of forward passes. It was set to 100 and had a processing time of 10 to 2111 s, depending on the complexity of the network architecture.
- Layer structure, which specifies the number of nodes in each layer of the network. A total of 3 network architectures were tested: (1) the default mode, which consists of 3 layers, each with 256 nodes; (2) following Mac et al. [51], a deep autoencoder network of 5 layers with nodes of sizes: 128, 40, 8, 40 and 128; and (3) an empirical approach using 2 layers, each with 72 nodes.

Since elevating the number of parameters may result in an improved generalization [52], the three network architectures were multiplied by the factor results of the logarithmic sequence: 1, 8, 27, 64, and 125. Therefore, MIDAS was run 5 times per network architecture, except in default mode, where factor 125 was omitted, as it exceeded the available computing resources (i.e., Intel core i7-8700 CPU, 40 GB RAM). After training, the *complete* function was applied to impute the missing values and retrieve a complete dataset. By default, 10 datasets were built, so this value was maintained. To identify the one that was more accurate, the test parameter (i.e., predicted carbohydrate contents) was compared with the one in the original dataset to derive the R-square and select the maximum among the 10 datasets predicted by MIDAS. The last step was to compare the network architectures and their multiplied factors to decide which achieved the highest R-square. Its complete dataset was then used in food crop clustering and reporting.

3.4. Food Crops Clustering and Report

The completed dataset was applied to cluster food crops based on their nutritional composition. For this, the K-means algorithm [53], implemented in the *stats* R base library, was used as it performed better than the other methods tested [53], but also because the nutritional composition parameters were all numeric. The K-means allows for partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean. To define k, the elbow method [54], implemented in the *factoextra* R library [55], was applied. It is based on a graph showing the sum of the squared distances between the observations and clusters on the *Y*-axis, followed by the sum of the squared distances for all clusters (a metric called the total within the sum of the square), and on the *X*-axis, the k clusters to identify the "elbow point". The elbow point is the point at which the value starts to decrease slowly and corresponds to the best k number. After observing the plot (Figure 1) and identifying that after 3 clusters, it became flat, the food crops were clustered. These clusters had sizes of 5, 53, and 13, which explain 76.4% of the total variance of the database.

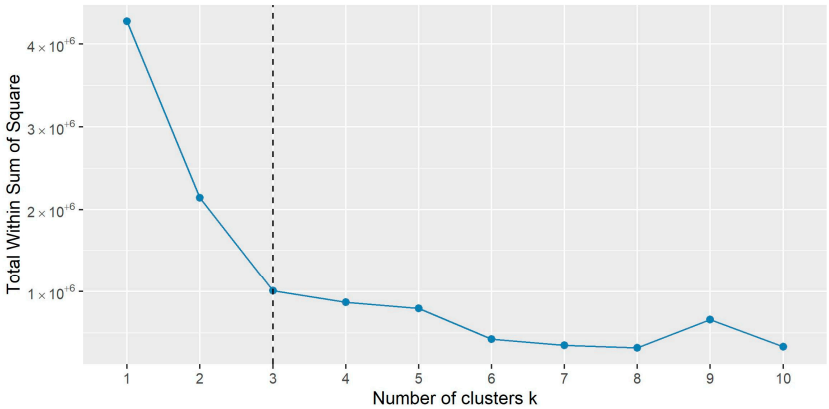

Figure 1. Elbow point observed before clustering with K-means.

4. Results

4.1. Database Integration after Web Scraping

The simplified ESPAC database resulted in 948 observations and 6 features, describing the common names, food crop identification codes, and their planted areas, disaggregated by provinces. After merging crop varieties to have them listed by plant species, the number of food crops was reduced from 107 to 89. As a result, from the web scraping from the ECOCROP database, 26 plant parameters were obtained for 72 food crops and collected in ~216 s, as each link was web scraped at 3 s delay intervals to avoid congesting the server. The categorical parameters obtained included information about the plant physiology, life form, habit, vegetable category, life span, planting attributes, plant uses, light requirements, soil characteristics (e.g., depth, texture, fertility, salinity, and drainage), and Koppen–Geiger climate types. The numerical parameters described the optimum crop requirements in nine columns, including temperature, rainfall, soil pH, early-growth killing temperature, and crop cycle duration. Since the FUNIBER server was slower to retrieve links, the interval applied between link web scraping was 5 s, resulting in a data collection time of ~2250 s. The information collected describes the nutritional content per 100 g of food, specifically: carbohydrates, monounsaturated and polyunsaturated fats, saturated fats, fiber, energy, protein, fat, calcium, iron, phosphorus, thiamine, vitamins A and B, pyridoxine, ascorbic acid, and tocopherol content. The integration of the information that was obtained for the same crop ranged from 1 to 18 tables, but in other cases, it resulted in no tables. The latter occurred for 8 out of 72 available food crops, and thus, data were obtained for 64 crop species. As a result of the integration of all databases, the food crops, their plant characteristics and environmental requirements, and their average nutritional contents were summarized in 49 parameters. However, of the information on the nutritional contents, 11.1% had missing values, while for the rest of the parameters, this reached 4.9%.

4.2. MIDAS Imputation and Accuracy

As mentioned in Section 3.3, duplicate variables and other non–informative parameters were filtered out before MIDAS imputation. The duplicated parameters were mostly categorical versions of quantitative variables such as soil fertility, salinity, and texture, which do not improve results. This resulted in 39 target parameters to be used with the *rMIDAS* functions, where 14 corresponded to nutrient contents and the rest to plant characteristics and environmental requirements. After the data imputation, network architectures were compared. In this regard, the architecture based on 3 layers and 256 nodes, multiplied by a factor of 64 (i.e., 1024 nodes for each layer), resulted in the best accuracy (Figure 2, green line on the left graph), which achieved an R-square of 0.84. In addition, the difference with the carbohydrate content parameter, which was used as a control, shows that this network architecture was the least variable (SD = 8.38) and the closest to the mean value (Mean = −0.78) (Figure 2, right graph). However, this model seems to underestimate carbohydrate content more often than the other models.

4.3. Differences in Planted Areas between the Land Use Database and INEC Tables

The enriched food crop database was constructed using the land use database, which is also used by INEC to report areas of permanent and transitory crops, pasture, fallow, and natural cover. The difference between this database and those reported for specific crops in the INEC tables is presented here. The INEC tables account for 29 of the 72 food crops processed with the land use database and enriched with the other databases. Figure 3 shows a bar chart to facilitate comparison of the tables.

Here, the areas reported in the INEC tables are shown in red, and those obtained from the land use database are shown in blue. The bar chart on the left shows the crop names (for Spanish and scientific crop names, please see) with planted areas greater than 40,000 ha, while the one on the right shows those below this area threshold. The figures in the INEC tables are higher for most food crops (i.e., 27 of 29 cases), which represent, in total, 1,116,669 ha (or 22%) more than those estimated from the land use database. Some

important differences were observed for food crops such as coffee, corn, and soy, which represent less than 80% of the area reported in the INEC tables. Only 10 of the 29 food crops matched were below 20%, indicating comparable figures. Because of these discrepancies, the planted areas of food crops shown later in this section correspond to the data in the INEC tables.

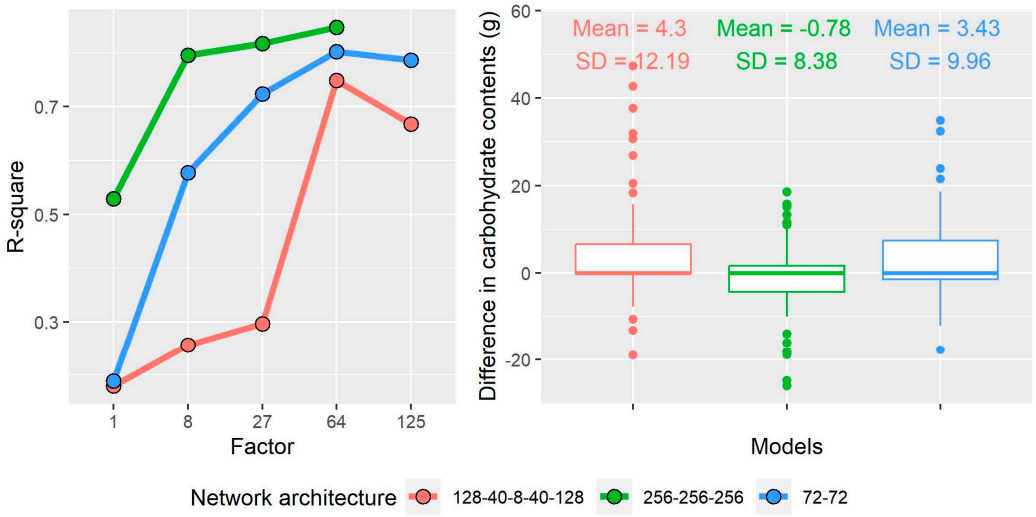

Figure 2. R-square and difference in carbohydrate content of the different network architecture models.

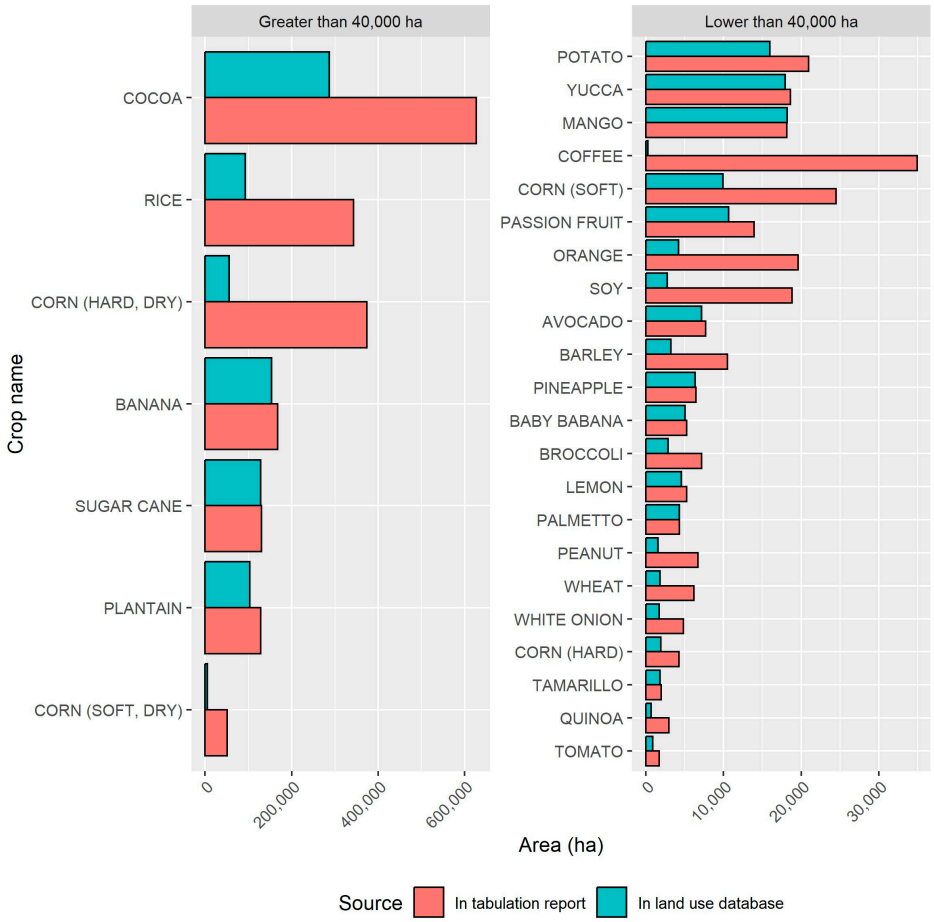

Figure 3. Comparison of planted areas obtained from table and land use.

4.4. Cluster Center Descriptions Based on Their Nutritional Content

The resulting clusters are represented in the form of bar graphs in Figure 4, in which the nutritional contents are shown.

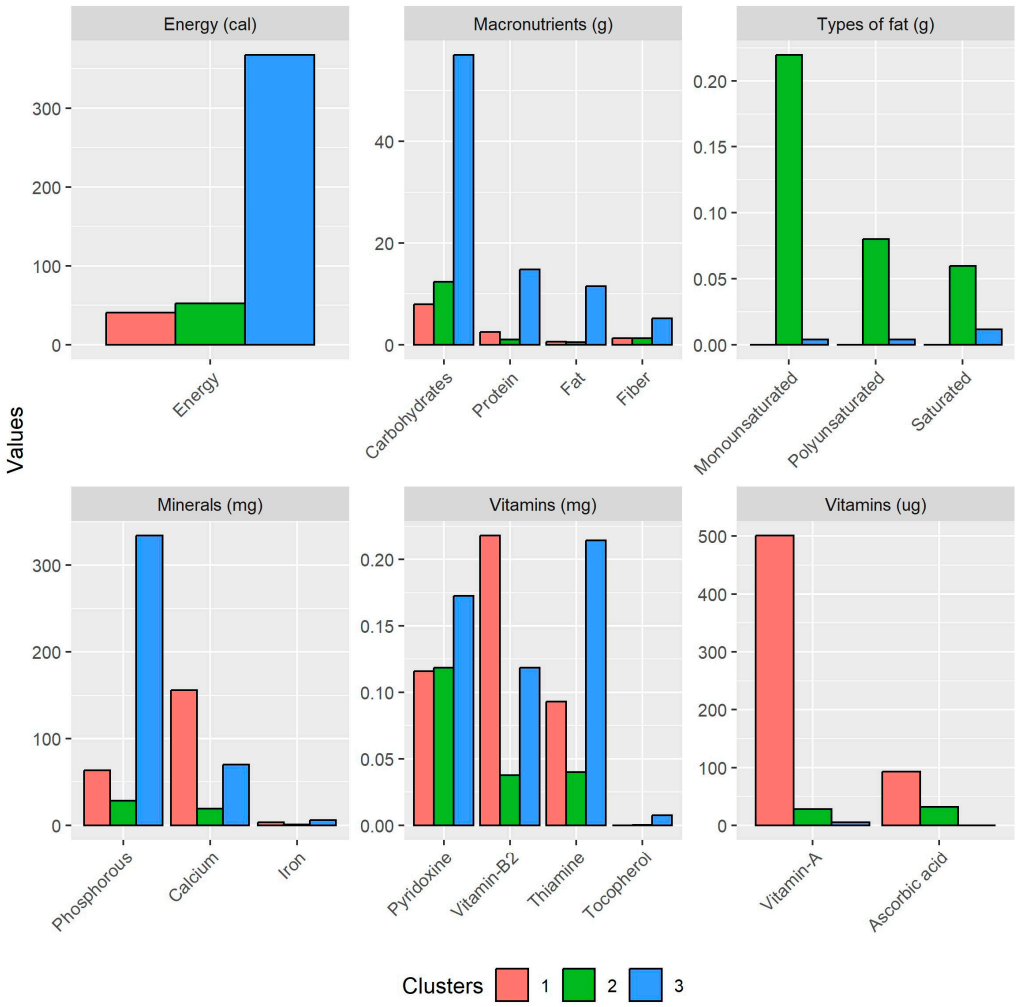

Figure 4. Nutritional composition observed in the clusters.

These were classified according to the micronutrient and macronutrient groups. Clusters 1, 2, and 3 accounted for 5, 53, and 13 food crops, respectively (for crop names of each cluster, please see Table S1). Crops that resembled cluster centers were: (1) spinach and coriander; (2) lettuce and pineapples; and (3) barley and peanut. The scales for each parameter were deduced after reviewing the literature, as the units were not specified in the source. Nevertheless, the three clusters can be easily differentiated among the parameters and nutrient groups. Starting with energy, it can be seen that cluster 3 achieved the highest value, i.e., 367 cal., followed by clusters 2 and 1, with 52 and 40 cal., respectively. Concerning macronutrients, cluster 3 also had higher values, with 56 g of carbohydrates, 14 g of proteins, 11 g of fat, and 5 g of fiber. For clusters 1 and 2, these values were lower, except carbohydrates were higher for cluster 2 (12 g). In more detail, the types of fats describe saturated and unsaturated fats (mono and poly). These stood out in cluster 2 as they were the highest of all the types, reaching a maximum of 0.21 g for monounsaturated fats and values above 0.05 g for the others. In the following, the micronutrients are described, and differentiated results were observed among the clusters. For example, in minerals, cluster 3 had the highest phosphorous (334 mg), while cluster 1 had the highest calcium (155 mg). Iron achieved the lowest value but was slightly higher for cluster 3 (5 mg). For vitamins, cluster 3 led for pyridoxine (0.17 mg); cluster 1 for vitamin B2 (0.21 mg); and cluster 3 for

thiamine (0.21 mg) and tocopherol (0.007 mg). Other relevant vitamin parameters were vitamin A, where cluster 1 peaked at 501 µg, along with ascorbic acid at 93 µg.

4.5. Food Crops and Plant Traits According to Clusters

Following the cluster descriptions, plant traits define the environments and uses of food crops. As the number of categorical parameters was greater than the numerical ones, Figure 5 shows, in the first two rows, a selection of different parameters in the form of bar graphs. The Y-axis shows, as percentages, the different categories of plant types, their uses, and their edaphoclimatic requirements. Therefore, it can be observed that cluster 1 is characterized mainly by herbs (100%), while cluster 2 is more heterogeneous and includes mostly herbs (45%), trees (27%), and shrubs (11%). Cluster 3 mainly shows a predominance of herbs (42%) and grasses (28%). With respect to life span, annual crops characterize cluster 3 (76%), and perennial crops, cluster 2 (65%). Cluster 1 is more heterogeneous, indicating a combination of food crops with annual (40%), biannual (40%), and perennial (20%) life spans. The following parameters show the different uses of food crops. In this sense, the three most important uses in clusters 1 and 2 indicate those related to food and beverages, medicinal, and material sectors (ranging from 13 to 29% of all food crops). These uses are also important in cluster 3, except for medicinal, which is replaced with animal food (17%). Other values that stand out are: cluster 1, food additive (21%); cluster 2, animal food (11%); and cluster 3, medicinal (15%). The next parameters are those related to pedology.

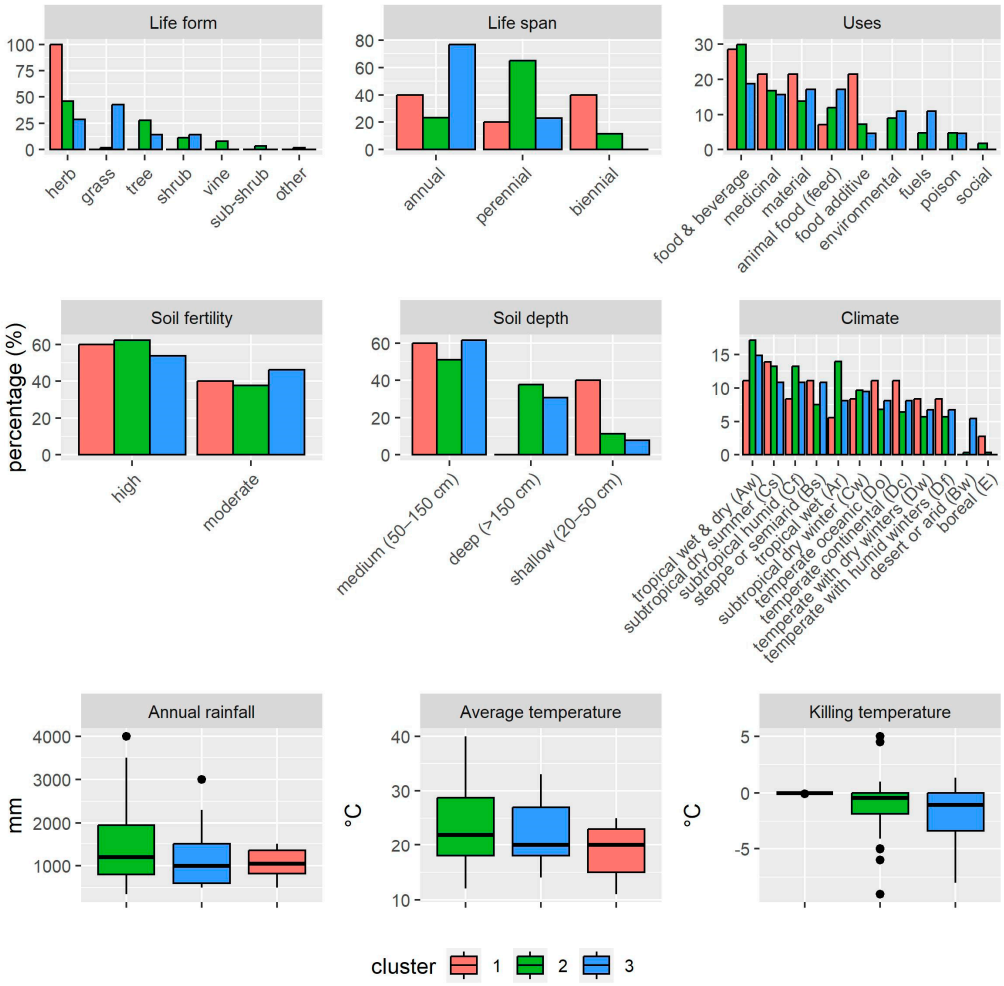

Figure 5. Plant traits observed in the clusters as percentages in the categorical parameters and boxplots in the numerical ones.

The first one is soil fertility, which characterizes all clusters with high soil fertility requirements at 53–62%. With moderate soil fertility requirements, all clusters indicate percentages between 37 and 46%. A related parameter is the soil depth, which indicates medium (50–150 cm) and deep (>150 cm) soil requirements for clusters 2 and 3, with percentages ranging from 30 to 61%. In contrast, cluster 1 shows only medium and shallow (20–50 cm) ranging percentages of 40–60%. The next parameter is climate, which is based on the Koppen–Geiger climate classification. This system establishes a hierarchical classification of climates based on a three-letter code, where the first character corresponds to one of the five general climate types, i.e., A: tropical, B: dry, C: temperate, D: continental and E: polar [46]. In this sense, cluster 1 represents more D types (38%), followed by C types (30%). In the case of cluster 2, types C (36%) and A (31%) are more relevant. Finally, in cluster 3, types C (31%) and D (29%) are the most representative. For the E types, a low proportion is observed, but cluster 1 has the lowest proportion of these (2%). The last row of Figure 5 shows the numerical variables as boxplots. On the left side, the annual precipitation rates are shown, where cluster 2 exhibits a more variable interquartile range (800–1950 mm), followed by cluster 3 (600–1500 mm) and cluster 1 (825–1350 mm). The average temperatures are similar, as the interquartile range of cluster 2 (18–28.7 °C) is more variable than that of cluster 3 (18–27 °C) and cluster 1 (15–23 °C). The last parameter, i.e., killing temperature, describes the temperature at which plants can be severely damaged or killed. Thus, it can be observed that cluster 1 shows no tolerance to temperatures below 0°, but cluster 2 shows more tolerance with an interquartile range of −1.8–0 °C, which is less variable than that observed in cluster 3, but the latter has a greater range of tolerance (−3.3–0 °C).

4.6. Planted Areas and Production Estimates of Food Crops in Clusters

The last result relates to the planted areas, which were obtained from the INEC tables. According to its source, these data were derived as estimates from the ESPAC complex survey for 15 permanent and 17 non-permanent crops whose processing is similar to that applied to the land use database (see Section 3.1). Since 29 food crops matched the clustered ones, they were summed according to each cluster. However, data could only be obtained for clusters 2 and 3, as none of the food crops in cluster 1 appeared in the INEC tables. In addition to the planted areas, these tables also included information regarding harvest, production, and sales data. All these estimates are shown in Table 1, where both clusters 2 and 3 have higher planted areas than harvested areas. In this regard, cluster 3 shows a slightly larger difference between planted and harvested areas, as it represents 7.1% of the total planted areas compared to 6.9% observed in cluster 2. In contrast, a larger difference is observed between clusters 2 and 3 when production and sales differences are compared. For example, in cluster 2, this difference represents 36% of the total production tons, while cluster 3 only represents 4%.

Table 1. Cluster surface areas and production estimates.

Cluster [1]	Food Crops (Count)	Planted (ha)	Harvested (ha)	Production (Tons)	Sales (Tons)
2	17	563,387	524,416	20,004,473	12,662,308
3	12	1,503,735	1,396,318	3,714,659	3,563,620

[1] ESPAC data were not available for Cluster 1.

5. Discussion

5.1. Possibilities and Challenges of Web Scraping, Data Imputation, and Database Integration

Web scraping allowed for the collection of 26 numerical and 17 categorical parameters from ECOCROP and FUNIBER databases in addition to those observed in the original ESPAC database. It took less than 15 min to search, retrieve, and organize the collected data when the request was directed at specific food crops. In this sense, the web-scraping technologies applied, i.e., Selenium and XML parsing, were favorable for our data-extraction

objectives and are recommended in the literature [56–58], but other open-source alternatives, such as BeautifulSoup, Scrapy, or Puppeteer for the Python [59] and Node.js [60] languages, are not less important. In addition, there are also commercial software alternatives which simplify web scraping when users are not familiar with programming such as Octoparse, ParseHub, or WebHarvy [61]. However, some particularities should be mentioned. In the FUNIBER database, for example, this procedure was not possible as multiple crop species concurred when using the search criteria using the crop name. It was necessary to download the entire base to later link the crop common names that were used in the database to specific species. This task required more time than expected (15–30 min.), in addition to the time spent on the web-scraping supervision, as the routine occasionally stopped due to the late response of the server. Most of the time required to implement the methods of the proposed approach was spent on data cleaning (Table 2), without taking into account the time spent on debugging the web-scraping routines. This is because the need to review the procedure resulted in a decrease in the number of food crop species as more databases were collated, but an increase in the number of associated parameters. These missing data were fixed with MIDAS, but its application required calibration until a network architecture and node number resulted in the best data imputation result.

Table 2. Food crops, data parameters, and task times.

Food Crops	Data Parameters (Count)		Missing Cases (%)	Operation	Task Time (Approx.) [1]
	Numeric	Categorical			
107	1	5	0	ESPAC raw data collection	short
89				Data cleaning, scientific names assignation	large
72	9	17	4.9	ECOCROP web scraping	short
64	17	0	11	FUNIBER web scraping	medium
72	27	22	0	MIDAS data imputation	medium

[1] Depends of the number of items. Short: <30 min; medium: 30–60 min; large: >60 min.

Although this can be tedious, the results of the MIDAS algorithm were certainly more reliable since the accuracy was beyond our expectations. Other imputation tools based on linear models, such as AMELIA [62], imputeTS [63], or Multivariate Imputation via Chained Equations [64] may not fit our analysis well, as our data were noisy and nonlinear relationships between variables were expected. However, MIDAS is not the only alternative and users of this approach are encouraged to try other methods based on deep learning, e.g., Generative Adversarial Imputation Networks [65], Fancyimpute [66], or Deep-Fill [67]. When performing experiments with these tools, it is recommended to pay attention to model tuning as, in our case, it was difficult to decide which categorical or numerical variables should be included in the imputation model to increase performance.

On the other hand, it was observed that data gaps were mainly due to divergences between datasets, e.g., in the case of species names due to the absence of scientific names in the databases. This limitation is not intrinsic to the web-scraping and data-imputation procedures, but a consequence of their assumption regarding the quality and reliability of the datasets. Therefore, it is strongly recommended that potential users of this approach evaluate critical aspects of the target database before collecting their data. Among these we can recommend:

- Metadata and methodological documents should be available and well-documented. In our case, the documentation of the FUNIBER database was not available, thus limiting our understanding of the reliability of this dataset, as well as that of the units of measurement of the nutrients analyzed.
- The existence of at least one binding parameter (e.g., scientific names, unique identifiers, date/time). This is required to be as unambiguous as possible to perform database integration. When a binding parameter is absent, it is recommended to

attribute this information by geocoding, which could also benefit from other spatial datasets (e.g., satellite imagery and GIS data).

- Ensure that data are open to the public and that no copyright is infringed when collecting data through web scraping. This may seem obvious, but it is common that there are no statements about the use of the data, so users should observe the presence of explicit mechanisms to stop web scraping such as captchas, robots.txt, IP blocking, and speed limiting.

5.2. Insights into the Nutritional Value of Widespread Planted Crops in Ecuador and Their Vulnerability to Climate Change

Classifying food crops according to their nutritional importance is relevant when talking about food security and the ability of individuals to use food to meet their nutritional needs [68]. This study gives unprecedented insights into the nutritional value of crops that are cultivated in Ecuador. The resulting clusters show that most food crops that are grown in the country are those qualified as "cash crops", i.e., those destined to market food production derivates (e.g., cacao, rice, sugar cane, corn, bananas). Even though our results show that many food crops that are rich in energy and macro- and micronutrients (except for vitamins A and C) are cultivated in the country, their productivity is difficult to estimate since plant areas are not available for some species. For example, this is true for cluster 1 and partly for the food crops of cluster 2. These nutrient-rich food crops are mainly cereals (e.g., barley, rice, rye, wheat, corn, oatmeal) and other grains (e.g., soy, peanut, quinoa, Andean lupin). Only a minority (7%) of the food crop species are high in vitamin A and C contents. This last group is composed of vegetables (e.g., spinach, turnip, carrots) and herbs (e.g., parsley and coriander), which were observed only in cluster 1. Of the 7 crops that occupy large farmland areas (more than a total of 40 ha at the national level according to both the INEC reports and the database estimates), 4 (57%) are part of the nutrient-rich cluster. Most of these crops are food staples (including two corn species and rice), while the remaining one is cocoa, which is currently an export product, but hardly eaten regularly to supply the nutritional needs of the local population. Other crops that currently occupy large farmland areas are not particularly nutritious but are important at a macroeconomic level as they are export products. These include two banana species and sugarcane. Published studies on food security in Ecuador coincide with the results obtained in this study by emphasizing the importance of cereals and grains [69]. Other authors, however, gave an even higher priority to less nutritious food staples, such as potatoes [70]. Food crops from cluster 1, rich in vitamins A and C, have not been given as much attention in the literature, regardless of their importance for nutrition, even when food staples are usually poor in micronutrient contents. Even though our data suggest that some highly nutritious crops are currently cultivated at both large and small scales in Ecuador, it is not clear which are the actual areas dedicated to their cultivation, and thus it is hard to estimate whether yields are enough to cover the country's needs. The discrepancies between the land use areas per food crop reported in the INEC tables and the numbers that were derived from the land use database raise questions about the validity of the data. On the other hand, the results on nutritious contents and clusters obtained in this study are based on the FUNIBER database, which has poor documentation about the methodologies that were used to obtain the data. Thus, the results on the nutritional content of the crops presented in this article should be further studied.

In a different way, the ECOCROP provided information on the environmental requirements of crops, which were well-documented, giving us information on the crops' tolerance to different environmental factors such as temperature, rainfall, soil type, and altitude. This dataset was useful for understanding that ESPAC crops were more related to annual herbaceous life forms with requirements for fertile soils with medium depths (50–150 cm). Moreover, climate requirements indicated the predominance of tropical (e.g., Aw, Ar) and subtropical (e.g., Cs, Cf) types according to the Koppen–Geiger classification. These climates types are located in areas close to the equator but also in coastal areas, where

during the wet season there is significant rainfall that favors the growth of vegetation and agricultural crops. However, during the dry season, in the subtropical type, there is little or no rainfall and the temperature is usually high. These climates are influenced by several factors, such as proximity to the equator, the movement of the Intertropical Convergence Zone (ITCZ), and the topography of the region [71]. As a result, the crops associated with these climates are also vulnerable to the effects of climate change, such as changes in rainfall patterns, more frequent droughts and floods, and temperature variations [72]. Therefore, Ecuador's listed food crops, which depend on these climate types, can be expected to be affected by a reduction in agricultural productivity due to higher temperatures [73], changing rainfall patterns, and more frequent extreme weather events such as droughts and floods [74,75]. In addition, changes in weather patterns can make it difficult for farmers to plan planting and harvesting schedules, resulting in lower crop yields and reduced incomes [76]. The distribution and abundance of pests and diseases, which can affect crops and livestock, can also be a problem, while water availability can impact agricultural production. If food security in Ecuador and similar countries is highly dependent on these climate types, then it is important to adopt strategies that could include developing drought-tolerant crop varieties [77], improving water management practices [78], promoting climate-smart agriculture [79], and implementing early warning systems for extreme weather events [80]. Finally, it is no less important to continue efforts to mitigate climate change by reducing greenhouse gas emissions to limit the magnitude of future impacts on food security in these regions.

5.3. Open Data and Requirements for Food Security

There is a growing tendency to share and publish datasets relevant to food security worldwide, as it is recognized as an important strategy to support decision making [81]. Nonetheless, the results of the agricultural censuses in Ecuador continue to be partly confidential even for research purposes. The United Nations member states have adopted the Sustainable Development Goals (SDGs) as a commitment and goal before 2030. A high priority was given to SDG2 "End hunger, achieve food security and improved nutrition and promote sustainable agriculture". In 2015, the Ecuadorian government adhered to the SDG agenda [82], which calls for a coordinated effort of the public, civil, and private sectors. As part of the efforts to work towards the goals and due to intrinsic interest to have reliable and timely information about the agricultural sector (which contributes to a relevant percentage of the gross domestic product), the ESPAC survey has been carried out yearly since 2002 (and since 2014 under the FAO framework). This census adheres to the national statistics law, which establishes that only numerical summaries, global concentrations, totalizations, and, in general, impersonal data, will be published [83]. Hence, all impersonal data should be of open access, not only to be congruent with the law but also to foster the collaboration among sectors that is enacted in the SDG agenda [84]. However, in practice, the ESPAC data content could only be partly obtained for this study. Important limitations to the use of the data and access to specific technical details were experienced in this work, even after contacting the developers of the datasets. Although the effort of implementing the ESPAC survey and the publication of part of the data must be recognized, we suggest that future data collections in the field of agriculture in Ecuador should consider additional criteria to design the surveys (e.g., scientific names of food crops), along with open discussions with academia to incorporate a holistic method to describe the food system, in addition to economic and statistical approaches.

Databases from non-governmental organizations used in this work enriched those obtained from public institutions. This was possible because of their open data policies that allow for innovation with novel approaches, such as the one proposed in this research. However, as mentioned before, such databases need to be thoroughly evaluated for their use in scientific projects. When methodological and metadata files are absent, a validation process is required. Despite these limitations, there are vast amounts of data to be collected and used, with proper documentation. In this regard, we point out that the ECOCROP

database was well-documented and organized at the time of our study, allowing for its use for research purposes.

6. Conclusions

In this paper we present an innovative approach to building food safety datasets. By combining web scraping to collect data and interpolating the gaps with the MIDAS algorithm, we were able to produce an enriched version of a public database with null information in terms of environmental and nutritional information. This approach can be applied to several areas of research, being relevant in its contribution to the automation of a process that, if performed manually, involves more work time and a higher probability of making mistakes. However, these advantages can only be fully exploited when high-quality databases linked to the Sustainable Development Goals are open to the public, especially in areas such as food security. Access to such data can assist in prioritizing goals, improving action plans, and measuring their outcomes. Open data on food security can also aid farmers to make informed decisions, facilitate the transition to organic production, manage food prices, adapt to climate change, promote healthy diets, identify food shortages, and comprehend farm–ecosystem interactions.

In our case study in Ecuador, achieving the curation of the food dataset was challenging as the quality and accessibility of the ESPAC data were limited and did not allow us to fully characterize the production areas, including the nutritional values for the crops obtained from FUNIBER. However, with the plant traits the results were better, since the ECOCROP database allowed us to achieve a better idea of the environmental constraints, which can be useful for designing strategies aimed at achieving food security in the country. Therefore, we suggest that future research on this topic explores additional indicators, such as the quantities of the nutrient-rich crops not identified in this study, along with the economic and physical accessibility of these products and the sustainability of their production, to better understand Ecuador's food security prospects.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agriculture13051015/s1. Table S1. Names of food crops analyzed in this research, together with their cluster and codes used to join ESPAC and FAO databases.

Author Contributions: Conceptualization, F.S. and N.A.; methodology, F.S.; software, F.S.; validation, F.S. and N.A.; formal analysis, F.S. and N.A; investigation, F.S. and N.A; resources, F.S.; data curation, F.S. and N.A; writing—original draft preparation, F.S.; writing—review and editing, N.A.; visualization, F.S.; supervision, F.S. and N.A; project administration, F.S.; funding acquisition, F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a research grant (INV-0014-016) from the Indoamerica University: https://uti.edu.ec/~utiweb/ (accessed on 1 May 2023).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here:

- ESPAC: https://www.ecuadorencifras.gob.ec/estadisticas-agropecuarias-2/ (accessed on 1 May 2023).
- ECOCROP: https://gaez.fao.org/pages/ecocrop (accessed on 1 May 2023).
- FUNIBER: https://www.composicionnutricional.com/composicion-nutricional (accessed on 1 May 2023).
- Processed data and scripts developed in this research can be found at: https://github.com/FSantosCodes/foodSecurityDataCuration (accessed on 1 May 2023).

Acknowledgments: The author thanks INEC, FUNIBER, and FAO for making data available. Special thanks to the R community and developers for making available all the software used in this work. Thanks also to five anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Prosekov, A.Y.; Ivanova, S.A. Food Security: The Challenge of the Present. *Geoforum* **2018**, *91*, 73–77. [CrossRef]
2. Azman Halimi, R.; Barkla, B.J.; Andrés-Hernandéz, L.; Mayes, S.; King, G.J. Bridging the Food Security Gap: An Information-Led Approach to Connect Dietary Nutrition, Food Composition and Crop Production. *J. Sci. Food Agric.* **2020**, *100*, 1495–1504. [CrossRef] [PubMed]
3. Ziv, G.; Baran, E.; Nam, S.; Rodríguez-Iturbe, I.; Levin, S.A. Trading-off Fish Biodiversity, Food Security, and Hydropower in the Mekong River Basin. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5609–5614. [CrossRef] [PubMed]
4. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793, ISBN 0-470-52679-3.
5. Salmoral, G.; Khatun, K.; Llive, F.; Lopez, C.M. Agricultural Development in Ecuador: A Compromise between Water and Food Security? *J. Clean. Prod.* **2018**, *202*, 779–791. [CrossRef]
6. Misra, N.N.; Dixit, Y.; Al-Mallahi, A.; Bhullar, M.S.; Upadhyay, R.; Martynenko, A. IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. *IEEE Internet Things J.* **2022**, *9*, 6305–6324. [CrossRef]
7. Muzenda, T.; Dambisya, P.M.; Kamkuemah, M.; Gausi, B.; Battersby, J.; Oni, T. Mapping Food and Physical Activity Environments in Low- and Middle-Income Countries: A Systematised Review. *Health Place* **2022**, *75*, 102809. [CrossRef] [PubMed]
8. Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N.C.; Ping, P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes* **2019**, *10*, 87. [CrossRef] [PubMed]
9. Cravero, A.; Pardo, S.; Galeas, P.; López Fenner, J.; Caniupán, M. Data Type and Data Sources for Agricultural Big Data and Machine Learning. *Sustainability* **2022**, *14*, 16131. [CrossRef]
10. Kumar, G.; Basri, S.; Imam, A.A.; Khowaja, S.A.; Capretz, L.F.; Balogun, A.O. Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. *Appl. Sci.* **2021**, *11*, 8275. [CrossRef]
11. Hariri, R.H.; Fredericks, E.M.; Bowers, K.M. Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges. *J. Big Data* **2019**, *6*, 44. [CrossRef]
12. Singrodia, V.; Mitra, A.; Paul, S. A Review on Web Scrapping and Its Applications. In Proceedings of the 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2019; pp. 1–6.
13. Wertheim-Heck, S.; van Bossum, J.; Levelt, M. Meeting the Growing Appetite of Cities–Delivering an Evidence Base for Urban Food Policy. In Proceedings of the IFoU 2018 Reframing Urban Resilience Implementation: Aligning Sustainability and Resilience, Barcelona, Spain, 10–12 December 2018.
14. Hillen, J. Web Scraping for Food Price Research. *Br. Food J.* **2019**, *121*, 3350–3361. [CrossRef]
15. Herforth, A.; Venkat, A.; Bai, Y.; Costlow, L.; Holleman, C.; Masters, W.A. *Methods and Options to Monitor the Cost and Affordability of a Healthy Diet Globally Background Paper for The State of Food Security and Nutrition in the World 2022*; FAO Agricultural Development: Rome, Italy, 2022.
16. Diouf, R.; Sarr, E.N.; Sall, O.; Birregah, B.; Bousso, M.; Mbaye, S.N. Web Scraping: State-of-the-Art and Areas of Application. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 6040–6042.
17. Lall, R.; Robinson, T. The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Anal.* **2022**, *30*, 179–196. [CrossRef]
18. Lin, W.-C.; Tsai, C.-F.; Zhong, J.R. Deep Learning for Missing Value Imputation of Continuous Data and the Effect of Data Discretization. *Knowl. -Based Syst.* **2022**, *239*, 108079. [CrossRef]
19. Zhao, B. Web Scraping. In *Encycl. Big Data*; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–3.
20. Liu, M.; Li, S.; Yuan, H.; Ong, M.E.H.; Ning, Y.; Xie, F.; Saffari, S.E.; Volovici, V.; Chakraborty, B.; Liu, N. Handling Missing Values in Healthcare Data: A Systematic Review of Deep Learning-Based Imputation Techniques. *arXiv* **2022**, arXiv:2210.08258.
21. INEC Encuesta de Superficie y Producción Agropecuaria Contínua—ESPAC. Available online: https://www.ecuadorencifras.gob.ec/estadisticas-agropecuarias-2/ (accessed on 31 October 2022).
22. FAO ECOCROP. Available online: https://gaez.fao.org/pages/ecocrop (accessed on 28 October 2022).
23. FUNIBER Base de Datos Internacional de Composición de Alimentos. Available online: https://www.composicionnutricional.com/composicion-nutricional (accessed on 28 October 2022).
24. Dogucu, M.; Çetinkaya-Rundel, M. Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *J. Stat. Data Sci. Educ.* **2021**, *29*, S112–S122. [CrossRef]
25. Munzert, S.; Rubba, C.; Meißner, P.; Nyhuis, D. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2014; ISBN 978-1-118-83481-7.
26. Wu, D.; Chau, K.T.; Wang, J.; Pan, C. A Comparative Study on Performance of XML Parser APIs (DOM and SAX) in Parsing Efficiency. In Proceedings of the 3rd International Conference on Cryptography, Security and Privacy; Association for Computing Machinery, New York, NY, USA, 19–21 January 2019; pp. 88–92.

27. Lamothe, M.; Guéhéneuc, Y.-G.; Shang, W. A Systematic Review of API Evolution Literature. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]
28. Postman What Is Postman? *Postman API Platform.* Available online: https://www.postman.com/product/what-is-postman/ (accessed on 21 April 2023).
29. Shete, D.; Bojewar, S.; Sanghvi, A. Survey Paper on Web Content Extraction & Classification. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2–4 April 2021; pp. 1–6.
30. Brewer, R.; Westlake, B.; Hart, T.; Arauza, O. The Ethics of Web Crawling and Web Scraping in Cybercrime Research: Navigating Issues of Consent, Privacy, and Other Potential Harms Associated with Automated Data Collection. In *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*; Lavorgna, A., Holt, T.J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 435–456, ISBN 978-3-030-74837-1.
31. Lin, W.-C.; Tsai, C.-F. Missing Value Imputation: A Review and Analysis of the Literature (2006–2017). *Artif. Intell. Rev.* **2020**, *53*, 1487–1509. [CrossRef]
32. Rodrigues, L.S.; Vespa, T.G.; Eleutério, I.A.R.; Oliveira, W.D.; Traina, A.J.M.; Traina, C. MiDaS: Extract Golden Results from Knowledge Discovery Even over Incomplete Databases. In *Proceedings of the Computational Science—ICCS 2022*; Groen, D., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 653–667.
33. Boursalie, O.; Samavi, R.; Doyle, T.E. Evaluation Methodology for Deep Learning Imputation Models. *Exp. Biol. Med.* **2022**, *247*, 1972–1987. [CrossRef]
34. Abiri, N.; Linse, B.; Edén, P.; Ohlsson, M. Establishing Strong Imputation Performance of a Denoising Autoencoder in a Wide Range of Missing Data Problems. *Neurocomputing* **2019**, *365*, 137–146. [CrossRef]
35. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Association for Computing Machinery. New York, NY, USA, 5–9 July 2008; pp. 1096–1103.
36. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. *Learning Internal Representations by Error Propagation*; California Univ San Diego La Jolla Inst for Cognitive Science: La Jolla, CA, USA, 1985.
37. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *Neural Netw.* **2012**, *2*, 1–18.
38. Bennett, D.A. How Can I Deal with Missing Data in My Study? *Aust. N. Z. J. Public Health* **2001**, *25*, 464–469. [CrossRef] [PubMed]
39. Gjorshoska, I.; Eftimov, T.; Trajanov, D. Missing Value Imputation in Food Composition Data with Denoising Autoencoders. *J. Food Compos. Anal.* **2022**, *112*, 104638. [CrossRef]
40. Costa, A.F.; Santos, M.S.; Soares, J.P.; Abreu, P.H. Missing Data Imputation via Denoising Autoencoders: The Untold Story. In *Advances in Intelligent Data Analysis XVII*; Duivesteijn, W., Siebes, A., Ukkonen, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 87–98.
41. Getz, K.; Hubbard, R.A.; Linn, K.A. Performance of Multiple Imputation Using Modern Machine Learning Methods in Electronic Health Records Data. *Epidemiology* **2023**, *34*, 206–215. [CrossRef] [PubMed]
42. R Development Core Team. *The R Project for Statistical Computing, Version 3.4.3*; R Development Core Team: Vienna, Austria, 2017; Available online: https://www.r-project.org/ (accessed on 1 May 2023).
43. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4.
44. Wickham, H. Reshaping Data with the Reshape Package. *J. Stat. Softw.* **2007**, *21*, 1–20. [CrossRef]
45. Lumley, T. Analysis of Complex Survey Samples. *J. Stat. Softw.* **2004**, *9*, 1–19. [CrossRef]
46. Software Freedom Conservancy Selenium. Available online: https://www.selenium.dev/ (accessed on 31 October 2022).
47. Harrison, J. RSelenium: R Bindings for "Selenium WebDriver"; 2022. Available online: https://cran.r-project.org/web/packages/RSelenium/index.html (accessed on 1 May 2023).
48. Lang, D.T. XML: Tools for Parsing and Generating XML Within R and S-Plus; 2022. Available online: https://cran.r-project.org/web/packages/XML/index.html (accessed on 1 May 2023).
49. Doerr, B.; Sutton, A.M. When Resampling to Cope with Noise, Use Median, Not Mean. In Proceedings of the Genetic and Evolutionary Computation Conference, Association for Computing Machinery. New York, NY, USA, 13–17 July 2019; pp. 242–248.
50. Robinson, T.; Lall, R.; Stenlake, A. RMIDAS: Multiple Imputation Using Denoising Autoencoders; 2022. Available online: https://cran.r-project.org/web/packages/rMIDAS/index.html (accessed on 1 May 2023).
51. Mac, H.; Truong, D.; Nguyen, L.; Nguyen, H.; Tran, H.A.; Tran, D. Detecting Attacks on Web Applications Using Autoencoder. In Proceedings of the Ninth International Symposium on Information and Communication Technology; Association for Com-puting Machinery, New York, NY, USA, 6–7 December 2018; pp. 416–421.
52. Bubeck, S.; Sellke, M. A Universal Law of Robustness via Isoperimetry. *J. ACM* **2021**, *70*, 1–18. [CrossRef]
53. MacQueen, J. Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Auckland, CA, USA, 1967; pp. 281–297.
54. Thorndike, R.L. Who Belongs in the Family. *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
55. Kassambara, A.; Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. 2020. Available online: https://cran.r-project.org/web/packages/factoextra/index.html (accessed on 1 May 2023).

56. Myllymaki, J. Effective Web Data Extraction with Standard XML Technologies. In Proceedings of the 10th International Conference on World Wide Web; Association for Computing Machinery, New York, NY, USA, 1–5 May 2001; pp. 689–696.
57. Manjari, K.U.; Rousha, S.; Sumanth, D.; Sirisha Devi, J. Extractive Text Summarization from Web Pages Using Selenium and TF-IDF Algorithm. In Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 15–17 June 2020; pp. 648–652.
58. Han, S.; Anderson, C.K. Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. *Cornell Hosp. Q.* **2021**, *62*, 89–104. [CrossRef]
59. Khder, M.A. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *Int. J. Adv. Soft Comput. Its Appl.* **2021**, *13*, 145–168. [CrossRef]
60. Chang, Z. A Survey of Modern Crawler Methods. In Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence, Virtual Event Japan, 11–13 March 2022; pp. 21–28.
61. Matta, P.; Sharma, S.; Uniyal, N. Comparative Study Of Various Scraping Tools: Pros And Cons. In Proceedings of the 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 11–13 February 2022; pp. 1–5.
62. Honaker, J.; Joseph, A.; King, G.; Scheve, K.; Singh, N. *Amelia: A Program for Missing Data*; Department of Government Harvard University: Cambridge, MA, USA, 1999.
63. Moritz, S.; Bartz-Beielstein, T. ImputeTS: Time Series Missing Value Imputation in R. *R J.* **2017**, *9*, 207. [CrossRef]
64. Hallam, A.; Mukherjee, D.; Chassagne, R. Multivariate Imputation via Chained Equations for Elastic Well Log Imputation and Prediction. *Appl. Comput. Geosci.* **2022**, *14*, 100083. [CrossRef]
65. Kim, J.; Tae, D.; Seok, J. A Survey of Missing Data Imputation Using Generative Adversarial Networks. In Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 19–21 February 2020; pp. 454–456.
66. Rubinsteyn, A.; Feldman, S. Fancyimpute: An Imputation Library for Python. 2016. Available online: https://github.com/Iskandr/Fancyimpute (accessed on 1 May 2023).
67. Shiri, I.; Sheikhzadeh, P.; Ay, M.R. Deep-Fill: Deep Learning Based Sinogram Domain Gap Filling in Positron Emission Tomography. *arXiv* **2019**, arXiv:1906.07168.
68. Roy, R.N.; Finck, A.; Blair, G.; Tandon, H. *Plant Nutrition for Food Security: A Guide for Integrated Nutrient Management*; FAO Fertilizer and Plant Nutrition Bulletin: Rome, Italy, 2006; Volume 16, p. 368.
69. Ochoa, F.B. ¿Hacia Un Modelo Agroalimentario Único? Diversidad e Identidades Espaciales En El Consumo de Alimentos En Ecuador. *Tsafiqui Rev. Científica En Cienc. Soc.* **2019**, *10*, 68–83. [CrossRef]
70. de los Santos Villalobos, S. *Inducción de Mutaciones: Estado Del Conocimiento En El Mejoramiento de Plantas En América Latina y El Caribe*; Editorial Fontamara: Coyoacán, Mexico, 2021; ISBN 978-607-736-684-3.
71. Garreaud, R.D.; Vuille, M.; Compagnucci, R.; Marengo, J. Present-Day South American Climate. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **2009**, *281*, 180–195. [CrossRef]
72. Reyer, C.P.; Adams, S.; Albrecht, T.; Baarsch, F.; Boit, A.; Canales Trujillo, N.; Cartsburg, M.; Coumou, D.; Eden, A.; Fernandes, E. Climate Change Impacts in Latin America and the Caribbean and Their Implications for Development. *Reg. Environ. Chang.* **2017**, *17*, 1601–1621. [CrossRef]
73. Blackmore, I.; Rivera, C.; Waters, W.F.; Iannotti, L.; Lesorogol, C. The Impact of Seasonality and Climate Variability on Livelihood Security in the Ecuadorian Andes. *Clim. Risk Manag.* **2021**, *32*, 100279. [CrossRef]
74. Sanabria, J.; Carrillo, C.M.; Labat, D. Unprecedented Rainfall and Moisture Patterns during El Niño 2016 in the Eastern Pacific and Tropical Andes: Northern Perú and Ecuador. *Atmosphere* **2019**, *10*, 768. [CrossRef]
75. Sáenz, C.; Litago, J.; Wiese, K.; Recuero, L.; Cicuéndez, V.; Palacios-Orueta, A. Drought Periods Identification in Ecuador between 2001 and 2018 Using SPEI and MODIS Data. *Eng. Proc.* **2021**, *9*, 9024. [CrossRef]
76. Mendelsohn, R. The Impact of Climate Change on Agriculture in Developing Countries. *J. Nat. Resour. Policy Res.* **2009**, *1*, 5–19. [CrossRef]
77. Nuccio, M.L.; Paul, M.; Bate, N.J.; Cohn, J.; Cutler, S.R. Where Are the Drought Tolerant Crops? An Assessment of More than Two Decades of Plant Biotechnology Effort in Crop Improvement. *Plant Sci.* **2018**, *273*, 110–119. [CrossRef] [PubMed]
78. Winterbottom, R.; Reij, C.; Garrity, D.; Glover, J.; Hellums, D.; McGahuey, M.; Scherr, S. *Improving Land and Water Management*; World Resources Institute: Washington, DC, USA, 2013.
79. Lipper, L.; Thornton, P.; Campbell, B.M.; Baedeker, T.; Braimoh, A.; Bwalya, M.; Caron, P.; Cattaneo, A.; Garrity, D.; Henry, K. Climate-Smart Agriculture for Food Security. *Nat. Clim. Chang.* **2014**, *4*, 1068–1072. [CrossRef]
80. Ebi, K.L.; Schmier, J.K. A Stitch in Time: Improving Public Health Early Warning Systems for Extreme Weather Events. *Epidemiol. Rev.* **2005**, *27*, 115–121. [CrossRef] [PubMed]
81. Restrepo, D.S.; Pérez, L.E.; López, D.M.; Vargas-Cañas, R.; Osorio-Valencia, J.S. Multi-Dimensional Dataset of Open Data and Satellite Images for Characterization of Food Security and Nutrition. *Front. Nutr.* **2022**, *8*, 796082. [CrossRef] [PubMed]
82. UN PROCESO DE NEGOCIACION INTERGUBERNAMENTAL HACIA LA AGENDA DE DESARROLLO POST. 2015. Available online: https://sdgs.un.org/statements/ecuador-13900 (accessed on 20 November 2022).

83. Portal Único de Trámites Ciudadanos Ley de Estadística | Ecuador—Guía Oficial de Trámites y Servicios. Available online: https://www.gob.ec/regulaciones/ley-estadistica (accessed on 20 November 2022).
84. OECD. Proceedings of the Getting Governments Organised to Deliver on the Sustainable Development Goals, New York, NY, USA, 18 July 2017. Available online: https://www.oecd.org/gov/SDGs-Summary-Report-WEB.pdf (accessed on 20 November 2022).