Linking Heterogeneous Data for Food Security Prediction

Hugo Deléglise[1,3,4(✉)], Agnès Bégué[1,3,4], Roberto Interdonato[1,3,4], Elodie Maître d'Hôtel[2,3,4], Mathieu Roche[1,3,4], and Maguelonne Teisseire[1,4,5]

[1] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
hugo.deleglise@cirad.fr
[2] MOISA, Univ Montpellier, CIHEAM-IAMM, CIRAD, INRAE, Institut Agro, Montpellier, France
[3] CIRAD, UMR TETIS, 34398 Montpellier, France
[4] CIRAD, UMR MOISA, 34398 Montpellier, France
[5] INRAE, Montpellier, France

Abstract. Identifying food insecurity situations timely and accurately is a complex challenge. To prevent food crisis and design appropriate interventions, several food security warning and monitoring systems are very active in food-insecure countries. However, the limited types of data selected and the limitations of data processing methods used make it difficult to apprehend food security in all its complexity.

In this work, we propose models that aim to predict two key indicators of food security: the food consumption score and the household dietary diversity score. These indicators are time consuming and costly to obtain. We propose using heterogeneous data as explanatory variables that are more convenient to collect. These indicators are calculated using data from the permanent agricultural survey conducted by the Burkinabe government and available since 2009. The proposed models use deep and machine learning methods to obtain an approximation of food security indicators from heterogeneous explanatory data. The explanatory data are rasters (population densities, rainfall estimates, land use, etc.), GPS points (of hospitals, schools, violent events), quantitative economic variables (maize prices, World Bank variables), meteorological and demographic variables. A basic research issue is to perform pre-processing adapted to each type of data and then to find the right methods and spatio-temporal scale to combine them. This work may also be useful in an operational approach, as the methods discussed could be used by food security warning and monitoring systems to complement their methods to obtain estimates of key indicators a few weeks in advance and to react more quickly in case of famine.

Keywords: Machine learning · Neural network · Heterogeneous data · Food security

I. Koprinska et al. (Eds.): ECML PKDD 2020 Workshops, CCIS 1323, pp. 335–344, 2020.
https://doi.org/10.1007/978-3-030-65965-3_22

1 Introduction

Hunger in Africa is growing again after several years of decline. In West Africa, progress in the fight against hunger was stable between 2000 and 2014. During this period, the prevalence of undernutrition gradually decreased from 12.3% to 10.7% before rising again to nearly 15% in 2017 [4]. The prevalence of severe food insecurity characterized by feeling hungry but not eating increased from 20.7% in 2014 to 29.5% in 2017 [3]. Burkina Faso is one of the most food-insecure countries in West Africa, with a prevalence of undernutrition of 21.3% between 2015 and 2017 [3]. The country is heavily affected by the "triple burden of malnutrition", a concept that highlights the complexity of malnutrition and its different expressions: undernutrition, micronutrient deficiency and excess weight/obesity. In 2017, the prevalence of wasting in children, stunting in children and obesity in adults was 7.6%, 27.3% and 4.5% respectively, the first two being among the highest in West Africa [4]. The reasons for the deterioration of the food situation in Burkina Faso in recent years are multiple and interrelated. A first reason is that climate change has led to an increase of extreme weather events such as droughts and floods that affect food availability [17]. A second reason is that conflicts in the Sahel displace populations and cause the fall of food production and distribution channels [9]. Both phenomenons hinder an economic downturn aggravated by an already fragile global economic context.

To prevent food crisis and design appropriate interventions, several food security warning and monitoring systems like the GIEWS (Global Information and Early Warning System) system created by the Food & Agriculture Organisation (FAO), and the FEWSNET one (Famine Early Warning Systems Network) founded by the United States Agency for International Development were set up in the second half of the last century by NGOs and state organizations, and are now very active in food-insecure countries. They do publish regular bulletins on the food situation at regional and national scale. These systems aim at preventing and treating food crisis through the establishment of targeted and appropriate food aid programs.

Given the difficulty of predicting food crisis, some aspects of these systems hinder more accurate and early predictions. First of all, to classify the level of food security (FS) of a territory, the different sources of information are exclusively combined and synthesized manually according to pre-established rules. This exclusively human intervention is time-consuming and limits the complexity and the updating of the decision rules. Moreover, these systems integrate mainly meteorological and remote sensing data, the integration of data from other fields related to FS (commodity prices, violent events, etc.) and other types (time series, high-resolution images) should make it possible to describe it more completely.

The objective of this paper is twofold: (i) To define original and efficient machine learning techniques for the processing of heterogeneous data in the context of FS. (ii) To enrich remote sensing data by linking them to data from different domains in order to make them more suitable for the analysis of complex FS phenomena.

Machine Learning methods are increasingly used to extract relevant information from complex and heterogeneous FS-related data, and several studies have attempted to detect food insecurity and crisis using machine learning techniques [1,10,14] with encouraging but improving results. A group of machine learning methods called deep learning is increasingly being used and is very effective in analysing complex and heterogeneous data [7]. Deep learning has been used with conclusive results for the analysis of FS-related topics such as poverty [15], drought [13] or market prices [12] but has not yet been used with convincing results in the field of FS.

In this work, we propose models that aim to predict two key indicators of FS: the food consumption score and the food diversity score. These indicators are calculated using data from the permanent agricultural survey conducted by the Burkinabe government and available since 2009. The proposed models use deep and machine learning methods to obtain an approximation of FS indicators from heterogeneous explanatory data. The explanatory data are rasters (Population densities, rainfall estimates, land use, etc.), GPS points (of hospitals, schools, violent events), quantitative economic variables (maize prices, World Bank variables), and also meteorological and demographic variables. An important issue was to perform pre-processing adapted to each type of data and then to find the right methods and the right spatio-temporal scale to combine them.

This work can be useful on 2 levels: on the one hand in a basic research approach. Indeed, the problem related to the combination of heterogeneous data is a current question of research, particularly in the field of FS; on the other hand, in an approach operational, the methods applied could be used by FS warning and monitoring systems in complement their methods to obtain estimates of key indicators a few weeks in advance and to be able to react more quickly in the event of famine.

2 Material and Method

2.1 Measuring Food Security

Food security is a complex and multifactorial concept, resulting from multiple and interrelated factors (e.g., climate, economy, wars). Food security holds "when all people have, at all times, physical and economic access to sufficient, safe and nutritious food" [16]. From this definition, four components emerge: (i) the availability in sufficient quantities of food of an appropriate nature and quality; (ii) the access of all persons to the resources necessary to acquire the food necessary for a nutritious diet; (iii) the stability of access to food over time despite natural or economic shocks; (iv) the appropriate use of food (storage, cooking, hygiene, etc.). These components can be appreciated at different levels, through data sources at the national, regional, household or individual level. There are a large number of food security indicators, and the use of several indicators is recommended because of the complexity of food security [2]. Hoddinott [5] estimated the number of food security indicators at about 450.

There are also a large number of proxies indicators related to one or more components of food security, such as vegetation indices, rainfall, food prices, local population densities, number of violent events, road conditions, number of schools and hospitals, etc.

The aim of this work is to propose machine learning models able to use FS proxies as input to predict FS indicators that are time-consuming and costly to obtain with classical methodologies, i.e., with data collected at the household level.

2.2 Study Data

Response Data. The response variables are derived from the Permanent Agricultural Survey, which has been conducted annually in routine by the Burkinabe Ministry of Agriculture since 1982 in Burkina Faso. For this study, we take into account the data that are available from 2009 to 2018 (personal communication, 2018). The resulting dataset contains information from 46400 farm households, i.e. an average of 4640 farm households per year distributed in 351 communes. A farm household is defined as a household practising one of the following activities: temporary crops (rainfed and off-season crops), fruit growing, animal husbandry. In this paper, we focus on two indicators based on answers to household surveys: the Food Consumption Score (FCS) and the Household Dietary Diversity Score ($HDDS$). They provide information on the frequency, quantity and quality of food, and are among the most popular indicators for researchers and organizations [6,11,18]. These indicators are averaged by commune and considered from 2009 to 2018, representing 3066 observations.

Food Consumption Score (FCS): This indicator is a proxy of the quantity of nutrient and energy intake. It is an estimate of the cumulative frequency of the different food groups consumed over 7 days within each household surveyed. The frequency of consumption of each food group is weighted by its nutritional value (Eq. 1; Table 1). Several thresholds to differentiate between households are commonly used. We choose thresholds set by the World Food Program (WFP): acceptable (>42), limit (28–42), and low (<28) [19]

$$FCS = \sum_{i=1}^{9} x_i.p_i \tag{1}$$

$x_i \in$ {Frequency of consumption for each food group i}, $p_i \in$ {Weighting of food groups}

Household Dietary Diversity Score ($HDDS$): It is an indicator of food consumption frequency and diversity more focused on the nutritional quality of the diet. It is an estimate of the number of food groups consumed in the last 24 h. There is no consensus on the choice of the number of groups to use and their boundaries. For example, WFP uses the same groups as the ones used for the FCS, while FAO uses a classification of 12 food groups [8]. The choice of food classification depends on the context (putting more emphasis on products rich in

Table 1. Food groups and their weights for the calculation of the Food Consumption Score (FCS). *Source:* [19]

Food group	Weighting
Cereals and tubers	2
Pulses	3
Vegetables and leaves	1
Fruits	1
Animal proteins	4
Dairy products	4
Sugars	0.5
Oils	0.5
Condiments	0

vitamins A, calories, etc.) and the available data. We use the FAO methodology to calculate the HDDS (Eq. 2; Table 2).

$$HDDS = \sum_{i=1}^{12} x_i \tag{2}$$

$x_i \in \{0\text{: food } i \text{ not consumed, } 1\text{: food } i \text{ consumed}\}$.

Table 2. Food groups for the calculation of the Household Dietary Diversity Score ($HDDS$). *Source:* [8]

Food group
Cereals
Roots and tubers
Vegetables
Fruits
Meat products
Eggs
Fish and seafood
Legumes, nuts and seeds
Milk and dairy products
Oils and fat
Sweets
Condiments, épices et boissons

Explanatory Data. First, FS proxies are pre-processed to extract relevant explanatory variables by commune, which is the smallest administrative boundary for which the response variables are spatialized. Some proxies have a finer

granularity and must be aggregated by commune; other proxies are available at a coarser granularity and must be interpolated on every commune. Then, for each commune and year, the explanatory variables obtained are selected by retaining only the explanatory variables significantly correlated with the response variable under consideration (p-value $<$ 0.05). The selected explanatory variables are classified into 4 groups according to their spatio-temporal granularity to be independently processed by a deep learning method for the prediction of the response variable:

Time Series [Multiple Values Per Year; One Value Per Commune]

- Smoothed Brightness Temperature (SMT); Source: National Oceanic and Atmospheric Administration (Noaa); Frequency: 7 days; Spatial resolution: 4km
- Rainfall estimate; Source: Tropical Rainfall Measuring Mission (Trmm); Frequency: 10 days; Spatial resolution: 6 km
- Maize price; Source: Société Nationale de Gestion du Stock de Sécurité alimentaire (SONAGESS); Frequency: 1 month; Spatial resolution: 64 markets. These data are interpolated on the centroids of each commune using the nearest k neighbour method.

=> These three time series are aggregated into monthly time series (May to November of the year in which the FS indicator is collected and of the previous year)

Conjunctural Data [One Value Per Year; One Value Per Commune]

- Meteorological data (daylight, temperature, humidity, evapotranspiration, wind); Source: Knoema platform; Frequency: 1 year; Spatial resolution: 10 stations. These data are interpolated on the centroids of each commune using the nearest k neighbour method.
- Population density raster; Source: Afripop, Frequency: 1 year; Spatial resolution: 100 m. Extraction by commune of quartiles, spatial autocorrelation, differential entropy and Gini coefficient.
- World Bank data (GDP growth, Consumer Price Index, military spending, etc.); Source: World Bank; Frequency: 1 year; Spatial resolution: Burkina Faso
- Mean SMT, rainfall estimates and maize price by commune

Structural Data [One Value Per Commune]

- Land cover map; Source: European Space Agency (ESA); 2016; Spatial resolution: 20 m. Calculation of the proportions of each type of soil per commune
- Hospitals, schools; Source: Open Street Map; 2018; Spatial resolution: Dot vectors. Calculation of numbers of hospitals and schools per 1000 habitants.
- Violent events; Source: Armed Conflict Location & Event Data Project (ACLED); 2018; Spatial resolution: Dot vectors. Calculation of the number of protests, riots, civil violence and total violence per 1000 habitants.

High Spatial Resolution Data [Multiple Values Per Commune]

– Population density raster; Source: Afripop, Frequency: 1 year; Spatial resolution: 100 m. The raster is split into 10×10 pixel patches.

Finally, each variable is centred reduced in relation to communes and years (consists of subtracting the mean and dividing it by the standard deviation).

2.3 Proposed Frameworks

We perform three types of regression analyses to predict FCS and $HDDS$. To assess performance, we randomly select 85% of the dataset for model learning and 15% for testing.

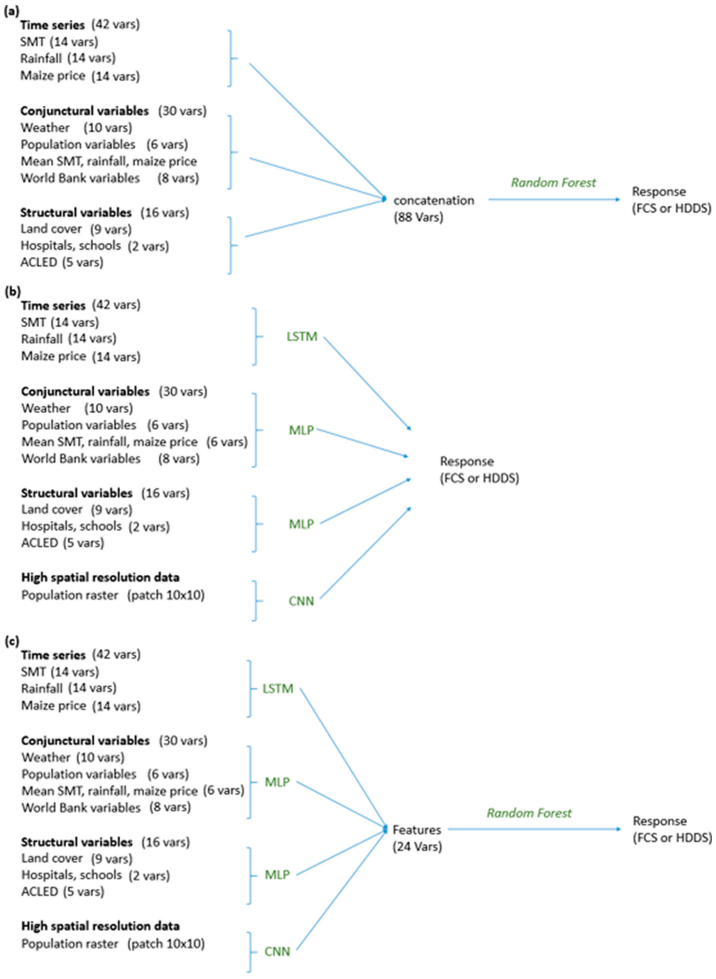

Fig. 1. Architecture of the three machine learning models (a), (b) and (c) used.

- (a) We apply a Random Forest directly on initial variables (for time series, conjunctural variables, structural variables and all variables).
- (b) We use four deep learning models separately on each group of variables. We apply a Long Short Term Memory (LSTM) on time series, a Multilayer Perceptron (MLP) on conjunctural and structural data and a Convolutional Neural Network (CNN) on high spatial resolution data.
- (c) We apply a random forest on features extracted by the deep learning models (for deep learning models associated with time series, conjunctural variables, structural variables, high spatial resolution variables, all variables) (Fig. 1).

3 Results

First, we find that the performance (R^2) is not high, not exceeding 0.38 for the $HDDS$ and 0.35 for the FCS (Table 3). These results show that the prediction of these FS indicators is a complex issue. At present, no published paper has attempted to predict these FS indicators, a WFP team worked on the prediction of the FCS in Burkina Faso for comparable results (R^=0.34)[1]. Second, we note that it is the structural variables (spatial distribution of schools, hospitals, violent events and land use) that seem to provide the most information: for the prediction of the FCS, the second best performance is obtained using only the structural variables; for the prediction of the $HDDS$, the addition of

Table 3. Performance (R^2) of the 3 types of models - (a): yellow; (b): green; (c): blue - for Food Consumption Score and Household Dietary Diversity Score prediction.

Model	FCS	HDDS
Random forest(Time series)	0.20	0.25
Random forest(Conjunctural variables)	0.27	0.32
Random forest(Structural variables)	0.30	0.38
Random forest(All variables)	0.35	0.37
LSTM(Time series)	0.20	0.21
MLP(Conjunctural variables)	0.10	0.11
MLP(Structural variables)	0.03	0.08
CNN(High spatial resolution data)	0.05	0.06
Random forest(Features(Time series))	0.19	0.20
Random forest(Features(Conjunctural variables)	0.08	0.13
Random forest(Features(Structural variables)	0.30	0.37
Random forest(Features(High spatial resolution data)	0.08	0.10
Random forest(All variables)	0.29	0.36

[1] https://wfp-vam.github.io/HRM/.

non-structural variables do not increase performance. Finally, the use of neural networks is complex for this type of multifactorial indicators and do not allow better performance than using classical machine learning methods for the current time.

4 Conclusion

This study proposes methods to obtain an approximation of FS indicators by integrating heterogeneous data. We faced two scientific obstacles: 1) the choice of input data and the preprocessing to be applied to them. In order to take into account all the facets of FS, we integrated several types of variables (vegetation index, meteorological, economic, demographic variables, etc.) with different spatio-temporal granularities, and we had to perform suitable treatments to extract relevant information; 2) The choice of methods to combine the different variables. We used machine learning and deep learning methods adapted to each group of data (LSTM adapted to time series and CNN adapted to images) and also used deep learning methods to extract features of the same dimension and therefore combinable. The best performances of this study outperform the only comparable (unpublished) work, this is mainly due to the diversity of the data collected and the pre-processing performed. The structural data (spatial distribution of schools, hospitals, violent events and land use) are the most informative in relation to FS. The future work will consist in improving the architecture and settings of the hyper-parameters of the deep learning models in order to take more into account the other types of data and better predict FS.

Acknowledgement. This study was conducted with the help of the Ministry of Agriculture, Water Resources, Sanitation and Food Security of Burkina Faso, the World Food Program (WFP) and the Société Nationale de Gestion du Stock de Sécurité Alimentaire (SONAGESS) which provided data. This work was supported by the French National Research Agency under the Investments for the Future Program #DigitAg, referred as ANR-16-CONV-0004.

References

1. Barbosa, R.M., Nelson, D.R.: The use of support vector machine to analyze food security in a region of Brazil. Appl. Artif. Intell. (2016). ISSN 10876545
2. Coates., J.: Build it back better: Deconstructing food security for improved measurement and action. Global Food Secur. **2**(3), 188-194 (2013). ISSN 22119124
3. FAO and ECA: Addressing the threat from climate variability and extremes for food security and nutrition. FAO (2018). ISBN 9789251311578
4. FAO, FIDA, OMS, WFP, and UNICEF. L'état de la sécurité alimentaire et de la nutrition dans le monde en 2018 : renforcer la Résilience face aux changements climatiques pour La sécurité alimentaire et la nutrition. FAO (2018). ISBN 978-92-5-130840-0
5. Hoddinott, J.: Choosing Outcome Indicators Of Household Food Security. International Food Policy Research Institute (1999)

6. Jones, A.D., Nguren, F.M., Pelto, G., Young, S.L.: What are we assessing when we measure food security? A compendium and review of current metrics. Adv. Nutr. (2013). ISSN 0022–3166
7. Julio, J.V.: Extreme learning machines with heterogeneous data types. Neurocomputing (2018). ISSN 18728286
8. Kennedy, G., Ballard, T., Dop, M.-C.: Guide pour mesurer la diversité alimentaire au niveau du ménage et de l'individu. FAO (2013)
9. Lacher, W.: Organized crime and conflict in the Sahel-Sahara region. Carnegie Endowment for International Peace (2012)
10. Lukyamuzi, A., Ngubiri, J., Okori, W.: Tracking food insecurity from tweets using data mining techniques. In: Proceedings of the 2018 International Conference on Software Engineering in Africa - SEiA 2018 (2018)
11. D. Maxwell, B. Vaitla, and J. Coates. How do indicators of household food insecurity measure up? An empirical comparison from Ethiopia. Food Policy **47**, 107-116 (2014). ISSN 03069192
12. Min, W., Ping, L., Lingfei, Z., Yan, C.: Stock market trend prediction using high-order information of time series. IEEE Access **7**, 28299-28308 (2019)
13. Mumtaz, A., Ravinesh, C.D., Nathan, J.D., Tek, M.: Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting. Comput. Electron. Agricult. **152**, 149–165 (2018)
14. W. Okori and J. Obua. Supervised Learning Algorithms For Famine Prediction. Appl. Artif. Intell. **25**(9), 822-835 (2011). ISSN 2078–0958
15. Shailesh, M.P., Tushar, A., Narayanan, C.K.: Multi-task deep learning for predicting poverty from satellite images. In: The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (2018)
16. Shaw, D.J.: World Food Security: A History Since: Palgrave MacMillan (2007). ISBN 10: 0230553559 (1945)
17. Tapsoba, A., Combes Motel, P., Combes, J.-L.: Remittances, food security and climate variability : the case of Burkina Faso. HAL (2019). ISSN 2114–7957
18. Vhurumuku, E.: Food security indicators - WFP. In: Integrating Nutrition and Food Security Programming for Emergency Response Workshop (2014)
19. Wiesmann, D., Bassett, L., Benson, T., Hoddinott., J.: Validation of the world food programme's food consumption score and alternative indicators of household food security. Int. Food Policy Res. Inst. (IFPRI) (2009)