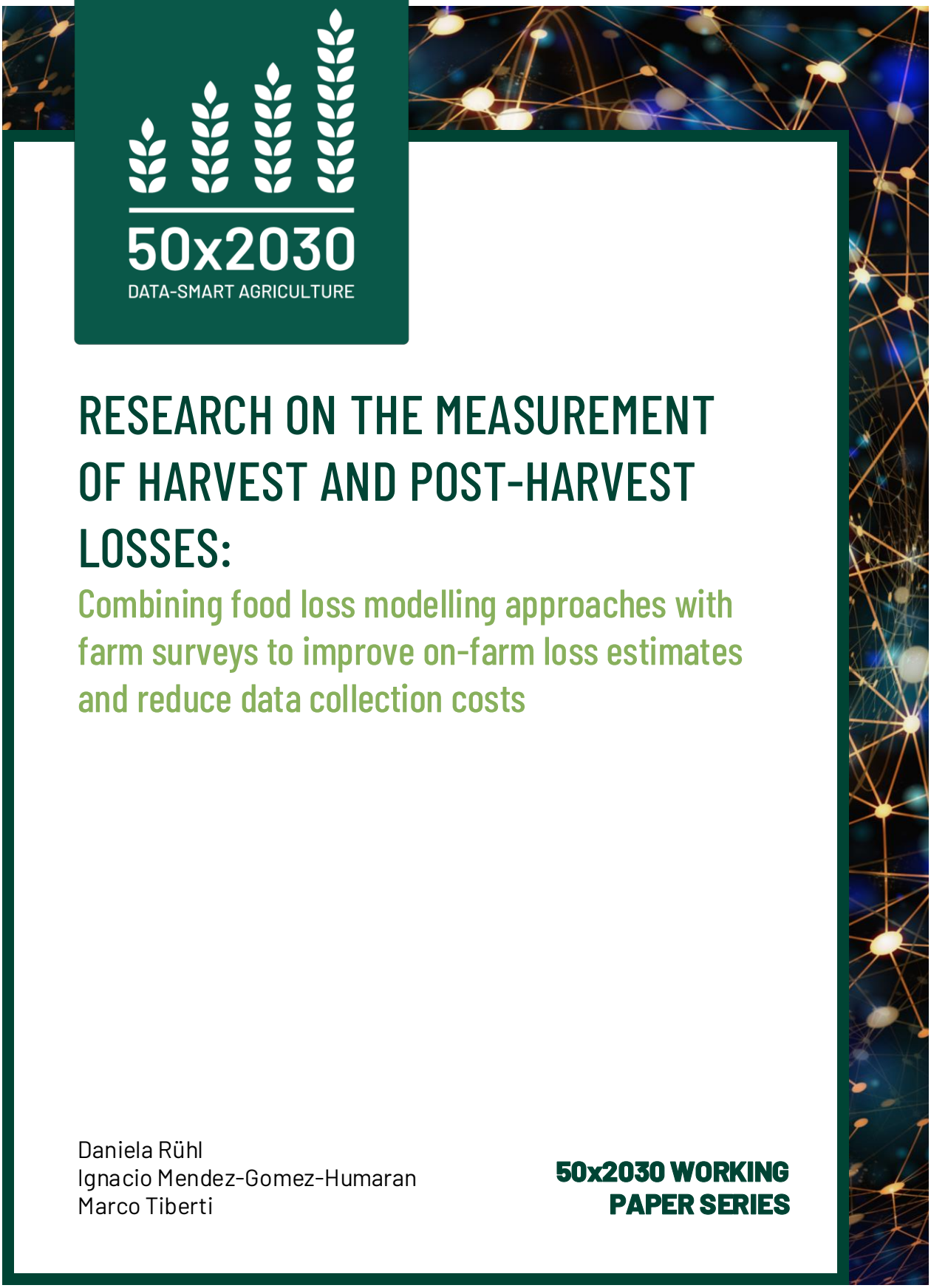


50x2030
DATA-SMART AGRICULTURE

RESEARCH ON THE MEASUREMENT OF HARVEST AND POST-HARVEST LOSSES:

Combining food loss modelling approaches with farm surveys to improve on-farm loss estimates and reduce data collection costs

Daniela Rühl
Ignacio Mendez-Gomez-Humaran
Marco Tiberti

50x2030 WORKING PAPER SERIES

ABSTRACT

Sustainable Development Goal 12 “responsible production and consumption’’ has a sub-indicator 12.3.1a Food Loss Index, which focuses on countries measuring and reporting on food losses from farm up to but excluding the retail level. While there is growing awareness of the issue of food losses at the political level, official post-harvest loss data for informing policymaking and reporting on the SDG Indicator is scarce. Notably, on-farm loss measurement is complex, and farm and household surveys often face several challenges in assessing and measuring these. As a result, food loss data collection can be burdensome for countries and their farm surveys

In this context, approaches that help to improve the farm loss estimates on the one hand, and reduce data collection costs on the other hand, can be relevant to countries. This research assesses the possibility of using modelling approaches in combination with farm surveys to improve the loss estimates, while reducing data collection costs associated with the farm survey. Two approaches to combining survey data with modelled estimates were tested for this research:

1. Identify and test a ***modelling approach to be used to support sub-sampling*** of the loss module in the farm survey, aiming to improve the loss estimates obtained from the sub-sample, and at the same time reducing data collection costs.

2. Identify and test a ***modelling approach for the estimation of food losses in consecutive year-to-year survey rounds,*** with the objective to identify the contribution of structural drivers as a potential tool for the indirect estimation of food losses.

The latter approach, combined with further research and development, would potentially make it possible to predict food losses based only on identified structural drivers collected regularly in the farm surveys.

For the first modelling approach to support sub-sampling, the proposed method is based on simpler post-stratification models using the Classification and Regression Tree (CART) method and the year-to-year estimation models are built using generalized structural equation models (GSEMs). The loss models were tested on four selected surveys, two food loss pilot surveys conducted in Malawi and Zimbabwe (Global Strategy to Improve Agricultural and Rural Statics, GSARS), and the Living Standard Measurement Study-Integrated Survey on Agriculture (LSMS-ISA) surveys conducted in Malawi and Nigeria.

For the first modelling approach, the performance of each model is assessed on the different sizes of sub-samples to improve the sample-based estimates, either by model-based estimates or by model-based data imputation. For the CART method, the research indicates that the model-based estimates improve the loss estimates of the sub-samples because of the post-stratification error reduction, thereby constituting a cost-effective complement to sub-sampling strategies. Model-based imputations should only be used on a reduced number of missing observations. The models perform best when the survey invests in obtaining more details on-farm loss data and considers some key variables identified as relevant for on-farm loss models. When using sub-sampling strategies, it is important to invest in more detailed questionnaires. Some considerations are derived from this research on how to design the questionnaires.

The use of GSEM is useful to generate food losses estimates that are modelled based on a set of determining factors available from the farm survey. This procedure was implemented in two rounds of the LSMS-ISA survey in Malawi and Nigeria and showed mixed results. Nevertheless, these models can be helpful to understand the contribution of several determinant and driving variables, covering socioeconomic farm characteristics, harvesting and post-harvest practices, and environmental factors, as well as their changes over time. The models were established using two survey rounds with a percentage of post-harvest losses. They are also useful to evaluate total and partial contributions of determinants on losses, such as conditional independent effects and possible non-independent effects based on the covariance between some of the determinants found. For the estimation of post-harvest losses, GSEM can be used as a procedure to estimate the percentage of post-harvest losses for survey rounds that collected information of food losses.

Table of Contents

ABSTRACT 2
Tables and figures 5
Acknowledgements. 7
Abbreviations 8
1.0 Introduction 9
1.1 Background and rational 9
1.2 Objectives and scope of the research 10
2. Literature review 12
2.1 Approach 12
2.2 Characterization of the literature sources 12
2.3 Identifying relevant modelling approaches 14
2.3.1 General findings 14
2.3.2 On-farm food loss models used to estimate food loss drivers 14
2.3.3 On-farm food loss models used to improve food loss survey data 17
2.3.4 Identifying explanatory factors of on-farm food losses 18
2.3.5. Conclusion of the literature review 24
3. Food loss modelling approach to be combined with farm surveys 24
3.1. Support sub-sampling of food loss modules in farm surveys with modelling approaches using post-stratification 25
3.1.1 The process of testing the modelling approaches 25
3.1.2 Evaluate the gains of the modelling approach for sample size reduction 26
3.1.3 Datasets available from farm surveys in Malawi and Zimbabwe 27
3.1.4. Results of applying the modelling approach: estimation of the food loss models for Malawi 32
3.1.5 Conclusions on the use of modelling approaches with sub-sampling food losses in farm surveys 38
3.2 Modelling approach for estimating farm losses with two independent survey rounds as alternative to reduce data collection costs 41
3.2.1 Overall modelling approach to be tested 41
3.2.2 Datasets used to test the modelling approach for Malawi 42
3.2.3. Results from applying the modelling approach to Malawi 43

3.2.4 Conclusions on modelling losses from independent survey rounds 57
4. Discussion and recommendations 58
References 60
Annex I: Overview of the papers screened for the literature review on determining factors 64
Annex II: List of indicators mentioned in the literature as determining factors of harvest and post-harvest losses 69
Annex III: Technical notes on statistical procedures 72
Brief description of exploratory factor procedures used 72
Annex IV: Additional country examples on the modelling approach tested for sub-sampling losses in the farm or household surveys 72
Models obtained GSARS Zimbabwe − maize: 72
Models obtained for the Nigeria LSMS-ISA surveys 74
Results obtained for the GSARS farm loss surveys in Zimbabwe 75
Results obtained for the Living Standard Measurement Studies in Nigeria 76
Annex V: General structural estimation model: estimation of the food loss model for Nigeria 77

Tables and figures

Figure 1: CART classification for Malawi GSARS Survey 33
Figure 2 CART classification for Malawi IHS4 Survey 35
Figure 3 Exploratory Factor Analysis Malawi LSMS-ISA 44
Figure 4 Exploratory Factor Analysis after Kaiser Criterium to select main factors 45
Figure 5 Graphical representation of the structural relationships between the variables 48
Figure 6 Final path diagramme for the reduced model 52
Figure 7 "LF2 precipitation" indirect effect on PHL through production 53
Figure 8 "LF3 Rrinfall ending year" indirect effect on PHL through production 53
Figure 9 "LF2 temperature" indirect effect on PHL through production 53
Figure 10 "Sample" indirect effect on PHL through production 54
Figure 11"Sample" indirect effect on PHL through LF3 rainfall ending year 54
Figure 12 "Sample" indirect effect on PHL through LF5 rainfall starting year 54
Figure 13 "Sample"indirect effect on PHL through LF2 precipitation 55
Figure 14 Total effect of survey sample round on post-harvest losses 55
Figure 15 Post-harvest loss model-based estimate for the whole period of the two included survey rounds 56
Figure 16 Model-based estimates of post-harvest loss for each survey round 56
Figure 17 Result of the linktest for the identified GSEM model (test model specifications) 56
Figure 18 CART classification Zimbabwe GSARS 73
Figure 19 CART classification Nigeria GHS 2015/16 74
Figure 20 Exploratory Factor Analysis, environmental factors, Nigeria 77

Figure 21 Selection of main factors as by Kaiser criterion 78
Figure 22 The second EFA obtained with reduced number of environmental variables 78
Figure 23 Varimax rotated eigenvalues, GSEM Nigeria 79
Figure 24 Path diagramme for linear relationships, GSME Nigeria 80
Figure 25 Path diagramme for the reduced model, GSME Nigeria 83
Figure 26 Mean estimate of post-harvest loss for the overall period, Nigeria 83
Figure 27 The post-harvest loss estimates for each survey round, Nigeria 84
Figure 28 Linktest for the final GSEM model, Nigeria 84

Acknowledgements

This paper was produced with financial support from the 50x2030 Initiative to Close the Agricultural Data Gap, a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030.

With the technical support from UN Food and Agriculture Organization and the World Bank. It was drafted by Daniela Rühl (FAO), Ignacio Mendez-Gomez-Humaran (Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Mexico), Marco Tiberti (World Bank) and edited by Sharon Mayienga (FAO) with guidance from Carola Fabi (FAO).

The authors want to thank Franck Cachia (FAO) for formulating the research component and the research idea that led into this research, and Marco Costantini (World Bank) for the support in the preparation of the analysis data. They also acknowledge the support of FAO and the World Bank in providing resources that enabled the process to be a success.

Abbreviations

50x2030	50x2030 Initiative for Data smart Agriculture
AGRIS	International System for Agricultural Science and Technology
AGRISurvey	Agricultural Integrated Survey Programme
CART	Classification and Regression Tree
CI	confidence interval
EFA	Exploratory Factor Analysis
FAO	Food and Agriculture Organization of the United Nations
FL	food losses
GSEM	generalized structural equation models
GHS	General Household Survey
GSARS	Global Strategy to Improve Agricultural and Rural Statistics
HIS	Integrated Household Survey
HS	household survey
LF	latent factors
LSMS	Living Standards Measurement Study
LSMS – ISA	Living Standards Measurement Study – Integrated Survey on Agriculture
PFM	Principal Factor Method
PHL	post-harvest loss
RCRE	Research Centre for Rural Economy of the Ministry of Agriculture and Rural Affairs of China
SEM	standard error of the mean

1.0 Introduction

1.1 Background and rational

As part of Sustainable Development Goal Indicator 12.3.1.a and the corresponding Food Loss Index, a major discussion has emerged on how to measure and monitor food losses at the country-level, covering the food supply chain from production up to, but not including, retail. The latest modelled food loss estimates published on the United Nations Statistics Division website[1] indicate losses at 13.2 percent, for 2021. Although losses differ considerably among commodities and countries, farms are one of the most critical loss points, as they have direct impacts on farmers' incomes, food security and natural resources (FAO, 2019).

Generating survey data at the farm level is one way to produce reliable estimates of harvest and post-harvest losses (PHL), orient decision-making and monitor progress towards reducing food losses. Although costly, surveys are necessary to obtain information on food losses that reflect actual reality, as opposed to a normative or theoretical measure or a metric that reflects recommended or best practices. On-farm loss measurement is complex, and farm and household surveys can encounter several challenges in assessing and estimating it, as outlined by Kitinoja *et al.* (2018), Xue *et al.* (2017), Delgado, Schuster and Torero (2017), Delgado, Schuster and Torero (2020) and Johnson *et al.* (2018). The multiple factors causing food losses, the different timings, operations and activities at which losses may occur, the considerable differences in the scale and cause of losses among commodities, typologies of actors, agro-ecological factors and management practices make measuring farm losses extremely burdensome. Collecting information on losses in farm and household surveys often requires breaking down the farm operations and asking the producer to quantify the losses for each operation (GSARS, 2018a). This can be time-consuming, especially given that these questions need to be asked for each of the farm's activities and crops and, in certain cases, for each plot (for harvest losses, for example). It also adds to the respondents' burden if a relatively long loss module is integrated into a broader farm or household survey, which could, therefore, undermine data quality. If losses were to be assessed through physical measurements or other methods to get more reliable data (Delgado, Schuster and Torero, 2017), in complement to or instead of farmer declarations, the interviewers' burden would be even more cumbersome, as these operations require more time and highly skilled enumerators. Because of these challenges, properly assessing on-farm losses can result in a relatively high burden on the farm or household surveys and the overall data collection effort.

On the other hand, several initiatives build models to estimate food losses. These models generally constitute a cost-effective strategy, and are reproducible and transparent (if variables, model structure and input data are described). They make it possible to link practices and production conditions and loss percentages, and provide a consistent explanatory framework for losses, which facilitates interpretation. Models, however, are often based on a partial representation of reality, for example, by incorporating only agronomic and production variables without considering the economic environment. They are often based on very partial data, drawn from field experiments, or limited datasets that might not cover all main food loss drivers. The consequence of this is that models alone are seldom able to provide representative information on losses.

[1] See https://unstats.un.org/sdgs/dataportal/database.

As part of the activities included in the 50x2030 Initiative to close the agricultural data gap (hereafter, 50x2030 Initiative) an optional questionnaire module for collecting data on harvest and post-harvest losses on the farm was designed (50x2030 Initiative, 2021). This module, which largely builds on the experience learned in the framework of the Global Strategy to Improve Agricultural and Rural Statistics (GSARS, 2018a), combines declarative and physical measurements and can be added to or integrated into to the other 50x2030 survey instruments, depending on the country's needs and demand.

To integrate the optional module on PHL into the modular 50x2030 Initiative survey system and, in parallel, optimize fieldwork implementation and data collection costs, sub-sampling of certain variables and modules is recommended in the 50x2030 Initiative sampling guidelines (50x2030 Initiative and FAO, 2021). The assessment of losses can thereby concentrate on a relatively small sub-sample of farmers and the freed resources invested into a more precise assessment of losses, either by detailing declarations or by using other methods to improve the estimates, such as physical measurements or visual scales.

On the other hand, the rotation of the different modules, including the food loss module, in the integrated farm or household survey is another strategy proposed in the 50x2030 Initiative sampling guidelines (50x2030 Initiative and FAO, 2021). Under this approach, countries can optimize data collection costs by implementing certain modules only in every second or third survey round.

In this context, farm loss modelling can constitute an additional instrument to support the sub-sampling strategy and the rotation of the loss module.

1.2 Objectives and scope of the research

The objective of this research is to come up with relevant modelling approaches to estimate or improve the estimation of agricultural losses at the country-level using mostly farm-level information obtained from large-scale farm or household surveys. These modelling approaches, used in combination and in complement with survey data, may also help to improve the overall cost-efficiency of statistics on farm losses by reducing data collection requirements. Apart from the objective to reduce data collection costs, the models also help to better identify the causal factors of losses (e.g., addressing labour shortages or promoting certain types of storage facility). The two main areas of application in the focus of this research are further described, as follows:

1. Identify and test a ***modelling approach to be used in combination with sub-sampling*** the loss module in the farm survey, aiming to improve the loss estimates obtained from the sub-sample, which, in turn, can reduce data collection costs,

   Specifically, the models are assessed to determine if they can improve the loss estimates obtained from smaller samples, either by using model-based estimates based on post-stratification or model-based estimates with imputation. To do this, the loss model is built on a sub-sample that uses a set of explanatory variables collected in the survey. The potential determinants are standard indicators that characterize the farm and its production system and may include, among other inputs, socioeconomic characteristics, such as age and level of education, harvesting methods and number of harvesting days, post-harvest technology used, information on the type of storage facility used, storage duration, use of pest control products during storage, and information on weather and production conditions. This modelling exercise entails examining

whether farm-level post-harvest loss models built on the available set of variables are sufficiently reliable to be used for prediction purposes. Afterwards, it is assessed whether modelling approaches can add to the results based on sub-sampling, for instance by improving the estimates and/or allowing a further reduction of the sub-sample (Affognon *et al*., 2016).

2. Identify and test a ***modelling approach for the estimation of food losses in consecutive year-to year survey rounds to identify structural drivers of losses, a potential tool for indirect estimation of food loss***, to be used to consider predictions based only in identified structural driver contributions farm surveys. This too would reduce the costs and frequency of post-harvest specific surveys.

   For this purpose, the models are assessed as to whether they can produce sufficiently reliable food loss estimates, while relying only on a set of explanatory variables collected in the survey to estimate food loss indirectly. By doing this, the survey data of two consecutive survey rounds of the Living Standard Measurement Study-Integrated Survey on Agriculture (LSMS-ISA) surveys are used for Malawi and Niger. A GSEM loss model is estimated and calibrated using a year-to-year approach. This model can be used to estimate losses indirectly for a survey round using only the explanatory variables considered. The input data are provided by the variables identified as being relevant in the LSMS-ISA surveys.

Apart from the requirement of providing options to reduce data collection costs and improve food loss estimates from farm or household surveys, the modelling approaches are intended to be easy to apply, making it more feasible for them to be used by national statistics officers after they receive capacity building on statistical modelling

In the rest of the paper, the most relevant results and conclusions are presented. First, a literature review was conducted to screen the most relevant determining factors for on-farm losses identified to date and the results are presented in chapter 2. In Chapter 3, the final modelling approaches that provide the best results are presented on both areas of application, describing the survey data used, the model specifications and the results. Only one country example is presented; the other country examples are added in the annex. The document closes with recommendations and conclusions.

2. Literature review

2.1 Approach

As a first step, a literature review was conducted to identify the main determining factors of on-farm losses identified. Of special use were journal publications of studies that used regression models on sub-national food loss surveys to identify the factors that cause losses.

The main sources used for the literature review were openly available research publications identified in Google Scholar, the AGRIS-library[2] and the FAO library[3]. Agricultural crops, especially grains, pulses, roots, tubers, fruits and vegetables, were the main commodity groups focused on for the literature review Animal products, such as meat, milk and eggs, as well as fish and fish products, were not screened specifically, and the first exploratory search did not generate any results on these commodity groups.

The review focused on on-farm food loss surveys and model approaches used to estimate post-harvest on-farm losses. Additionally, studies treating only on-farm storage losses were included in the literature review. The research was conducted between November 2020 and January 2021 on articles that were published from 1990 onwards.

Search and selection of the relevant literature

For the literature review, 126 articles were selected and screened, out of which 62 of them were evaluated as relevant publications that provided inputs about determining factors of on-farm losses and food loss models. Among the publications that were considered not relevant, several of them were related to food waste at the retail and household levels; some authors use the terms food loss and food waste interchangeably, and for this reason some articles related to food waste at the retail and consumer levels appeared in the search for food loss articles. Other articles not included in the review were on estimating food losses at the farm level and analysing causing factors without using a modelling approach. Another group of articles screened but discarded analysed different post-harvest technologies and their relevance for food losses, but without using food loss data or estimating the impact of these on food loss levels. A smaller group of papers was discarded because they focused on food losses at off-farm stages only, especially losses at markets places.

The key words used for the literature review are food losses, food loss determinants, food loss drivers or causing factors, food loss regression and food loss models. Articles that provide a food loss literature review are used as a first orientation; for instance, Affognon *et al*. (2016); Agarwal *et al.* (2021); Kader (2005); and Xue *et al*. (2017) review the food loss literature based on the methods and approaches used, modelling approaches being one of them.

2.2 Characterization of the literature sources

Commodity groups and commodities

The distribution by commodities shows that grains and pulses are the most represented commodities within the pool of articles found. Staple crops, such as rice, maize and wheat, are the most studied grains,

[2] See www.fao.org/agris/agris-and-network.
[3] See www.fao.org/library/resources/fao-library-discovery/en/.

while soya and lentils are only included in a few studies. Another group of commodities that are to some extent covered in the pool of articles are roots and tubers. Staple crops, such as potato, sweet potato, cassava and yams, are more frequently represented in the studies. No studies were found on animal products, fish and fish products, which may explain why there is a lack of sufficient food loss data of these two commodity groups in general.

Regions and countries

Considering the regional distribution, the pool of studies is highly concentrated in the African and Asian region, while with only few studies focus on Europe, North America, Oceania and Latin America. In the African region, a considerable number of papers on loss modelling approaches at the farm loss focus on Nigeria, Ethiopia, Kenya, Ghana and the United Republic Tanzania. In Asia, Bangladesh, India and Nepal are the countries most covered on this topic. This result is not surprising given the that the farm level represents a more relevant critical food loss point to most countries in the African and Asian region, while the research focus in most countries in Europe and North America tends to be at the retail and consumption level.

Table 1. Summary of the relevant papers identified during the literature review that use modelling approaches with post-harvest loss survey data

General characteristics of the relevant literature reviewed	Number of papers
By commodity groups and commodities	
Grains (rice, wheat, maize, soya, lentil, among others)	21
Roots and tubers (potato, sweet potato, yam, onion, carrots, among others)	12
Vegetables (tomato, cabbage, spinach, leeks, among others)	13
Fruits (banana, mango, among others)	6
Regions	
Asia (India, Bangladesh, China, Nepal, Sri Lanka, among others)	15
Africa (Kenya, Ghana, Uganda, United Republic of Tanzania, Ethiopia, Nigeria, Zimbabwe, Malawi, South Africa, among others)	>25
Latin America and the Caribbean (Brazil)	1
Europe and North America	1
Years	
2000– 2009	5
2000–2014	7
2015–2021	>25

Years of publication

The distribution of the articles by year of publication indicates interest in this area has picked up. As shown in table 1, most studies were published from 2015 onward and significantly less studies were published during the periods 2000−2009 and 2010−2014. This may indicate a recent rise in interest in this research topic and that further studies are likely.

2.3 Identifying relevant modelling approaches

2.3.1 General findings

The literature review shows that several types of food loss models and areas of applications were studied. The selection of most relevant approaches for the purpose of this research can be grouped as follows:

- **First group of articles:** characterized by models that are microlevel applications at the farm level, that make use of survey data to estimate the determinants and explanatory factors of on-farm food losses.
- **Second group of articles:** focuses on models and approaches that aim to improve on-farm food loss estimates from farm survey data.

Table 2. Summary of the different areas of applications of modelling approaches and food loss survey data identified in the literature review

On-farm food loss drivers and determinants	Number of papers
Papers that use regression models to identify food loss drivers within sub-national food loss survey	34
Papers that use regression models to identify food loss drivers/determinants within national surveys and complementary data sources	6
Models to improve on-farm food loss data from surveys	**Number of papers**
Papers that use models to apply different food loss measurement approaches (physical, by enquiry)	3
Papers that use model-based imputation for food loss data gaps in the collected surveys	1

The first two groups of models were considered the most relevant for the purpose of this research. Accordingly, for the literature review, these types of studies were sought. This may explain why this is the group with largest number of articles. The other group of food loss models were screened to identify interesting and complementary elements on the model specifications, although they were not analysed with further detail. The following contains a description on these groups of food loss models and the most relevant aspects are extracted to feed into the food loss models proposed in the context of this research.

2.3.2 On-farm food loss models used to estimate food loss drivers

As already highlighted, the first group of models is the most relevant for the objectives of this research and is the focus on identifying relevant studies for the literature review. The common characteristic of

this group of models is that they are built on farm loss surveys and apply food loss models for different purposes. Most papers conducted a regression analysis with the goal to identify the main explanatory factors of food losses. They were conducted in the context of a sub-national food loss survey on a specific crop and region. Only a few regression models were built on national survey data of food losses.

While the main objective of the research is to develop food loss modelling approaches based on national survey data, the lack of food loss indicators collected through nationally representative farm or household surveys is a limitation. Only three articles were identified, the studies were from the United Republic of Tanzania and China.

In the case of Unite Republic of Tanzania, two of the related published articles made use of the National Panel Survey on Households conducted by the National Bureau of Statistics in collaboration with the Ministry of Agriculture, Food Security and Cooperatives.[4] Both studies focused on the link between storage technology adaptation and food losses. Ndiritu and Ruhinduka (2019) examined the influencing factors on a binary technology adaptation choice function, food losses are treated only as a conceptual objective and the choice of technology will eventually have an impact without estimating a relation between the technology and food losses and Ngowi and Selejio (2019) established a post-harvest loss model, using a binary variable if the producer/household experienced food losses. The probabilistic choice model can be derived and estimated by the binary Logit model, which is used to determine the role of each variable in explaining the variation in the dependent variable.

In the case of China, Qu *et al.* (2020) makes use of a national survey that was conducted by China Agricultural University jointly with the Research Centre for Rural Economy of the Ministry of Agriculture and Rural Affairs of China in 2016 (China Agricultura University and Research Centre for Rural Economy, 2016).[5] The survey collected data on food losses from harvesting up to post-harvest operations of grains by declaration, as well as other relevant socioeconomic indicators. The model used to estimate influencing factors of food losses was able to differentiate traditional farmers from modern farmers, which is an additional element in national surveys, as they have a more comprehensive coverage, which includes all types of farmers that participate in national production.

The approaches differ from study-based food loss survey data with respect to the food loss model structure that in this case is limited by the information gathered within the national survey. The latter is probably less related to concrete food loss factors and more related to general household and production characteristics, but it benefits from a broader population and regional coverage.

Most of the study surveys collected cross-sectional food loss data of single commodities and sub-national regions, while on a few of them covered a larger number of crops. The literature review found approximately 40 articles. Several countries were represented in between two and six research articles. The greatest contribution to this research topic comes from Ethiopia and Nigeria, followed by Bangladesh, Kenya, India, Nepal, Uganda and Ghana. Several studies build on each other. For instance, Kumar *et al*. (2006) and Basavaraja, Mahajanashetti and Udagatti (2007) were cited by various articles and represent

[4] Access National Bureau of Statistics https://www.nbs.go.tz/index.php/en/census-surveys/poverty-indicators-statistics/national-panel-survey; Microdata: https://microdata.worldbank.org/index.php/catalog/2862.

[5] CAU and RCRE (2016). Joint Survey on Grain Harvest Loss of Farm Households by Grain Economy Research Group of China Agricultural University (CAU) and Rural Fixed Observatory Point Office of Research Center for Rural Economy (RCRE) of Ministry of Agriculture and Rural Affairs of China. Ministry of Agriculture and Rural Affairs of China. [Unpublished raw data].

one of the first applications of regression analysis to determine food loss drivers. Some of the mentioned studies used the same approach for off-farm stages, but they were not further examined for the purpose of this research.

Estimation strategy: The estimation strategies focus on microlevel approaches used to estimate the impact of different explanatory factors on post-harvest food losses at the farm level. The most used estimation approach is a multiple linear regression, using a set of explanatory variables to explain the dependant variable of food losses. This approach was used, for instance, by Kumar *et al.* (2006) and Basavaraja Mahajanashetti and Udagatti (2007) on grains, roots and tubers in India, Begum, Hossain and Papanagiotou (2012) and Khatun *et al*. (2014) on rice, wheat and tomato in Bangladesh, Arun and Ghimire (2019) and Paneru, Paudel and Thapa (2018) on a variety of commodities in Nepal, Adisa *et al.* (2015) on yam, Babalola *et al*. (2010) on tomato in Nigeria, Tadesse, Bakala and Mariam (2018) on potato in Ethiopia and Ambler, Brauw and Godlonton (2018) on cereals in Malawi.

Some authors opted to use a double and semi logarithmic multiple regression analysis, such as Folayan, Babalola and Ilesa (2013) on maize in Ethiopia, Aidoo *at al.* (2014) on tomato in Ghana, and Huang *et al.* (2017) on grains in China. Ansah and Tetteh (2016) used a fractional logistic regression model because of the proportional nature of the dependent variable, which, in this case, is the post-harvest management, used as a way of assessing the inverse of food losses. Hossain and Miah (2009) suggested a Cobb-Douglas production model to estimate the coefficients of the factors influencing potato storage losses in Bangladesh.

Shee *et al*. (2019) and Garikai (2014) used an ordered probit model, employing food loss categories that place loss percentages into four food loss categories. These categories were built from food loss percentage data collected in the respective study survey. Amentae *et al.* (2016) and Falola *et al*. (2017) used a tobit regression model under which only a binary category of food losses (low and high food losses, experience and do not experience losses) was applied. Kikulwe *et al.* (2018) made use of a tobit censored regression model to solve the limitation in the dataset of a significant number of producers who reported zero food losses.

Data: Most of the surveys collected food loss data by declaration together with other socioeconomic characteristics of the producer or household, agronomic indicators of the production or post-harvest management, and climate factors. Accordingly, the data for the regression models were obtained from the same survey with only very few studies that combined the survey data with climate data from other sources, such as, for example, Ambler, Brauw and Godlonton (2018). In general, the surveys conducted for the food loss studies had a sample size of approximately 100 to 300 households, which depended on the target population, regional coverage and available resources.

Dependent variable: Almost all surveys estimated total post-harvest losses aggregated for all post-harvest activities, while only some studies disaggregated losses by each operation or concentrated on one specific operation only. Storage losses were one of the operations of special interest, such as Kimenju and De Groote (2010) on maize, Falola *et al.* (2017) on yam and Hossain and Miah (2009) on potato storage. Ambler, Brauw and Godlonton (2018) conducted the regression model on aggregated post-harvest losses and by each post-harvest operation for maize, soya and groundnuts. The differences between them becomes clear on the choice of independent variables, which are more specific if the study and focus is

on a single post-harvest activity. Two studies, namely Kikulwe *et al*. (2018) and Qu *et al.* (2020), included harvest losses complementary to post-harvest losses, on banana in Uganda and on grains in China.

The most frequently used dependent variable is loss quantities in kilograms, as applied by Aidoo *et* al. (2014) and Alidu, Ali and Aminu (2018) for tomato in Ghana, Tadesse, Bakala and Mariam (2018) for potato in Ethiopia and Folayan, Babalola and Ilesa (2013) for maize in Nigeria in Ghana. Four studies considered food loss in kilograms per hectare, which is a way to relate the loss quantity to the size of the farm; they were Kumar, Basavaraja and Mahajanshetti (2006) and Basavaraja, Mahajanashetti and Udagatti (2007) in India and Begum, Hossain and Papanagiotou (2012) and Khatun *et al.* (2014) in Bangladesh. On the other hand, approximately six studies used food loss percentages as the dependant variable, dividing food loss quantities by the total quantity produced, including Mebratie *et al.* (2015) and Amentae *et al*. (2016) in Ethiopia and Paneru, Paudel and Thapa (2018) and Arun and Ghimire (2019) in Nepal. Each of these three approaches can have different implications in terms of the relevance of food loss drivers. Food loss quantities are likely to be positively related to production volume and factors influenced by the size of the farm. Food loss percentages are more commonly used to identify structural food losses usually caused by the type of production system, climate and agronomic practices. Food loss percentages tend to diverge between the latter to a greater degree than food loss quantities. On a per acre basis, food loss quantities could be interpreted similar to food loss percentages, although it does not take into account the differences in productivity that are counted in when using food loss percentages. Kikulwe *et al.* (2018) applied a regression analysis on the drivers of food loss quantities and food loss percentages, which provides the possibility to compare the changes in the significance of explanatory factors between both concepts.

On the other hand, three studies, Shee *et al.* (2019), Maziku (2019) and Garikai (2014), used food loss categories of minimum, low, medium and/or high food losses. These were either consulted directly as a categorial question to the producer or built on the food loss percentages the producer declared. Some studies, such as Morris, Kamarulzaman and Morris (2019) and Falola *et al.* (2017) on food losses in yam and plantain in Nigeria, need to be analysed separately as they build the regression analysis on the adaptation or use of technologies that are directly linked to higher or lower losses.

Other approaches: Apart from these estimation strategies, a few studies diverge by using other approaches to identify food loss determinants. Meena *et al.* (2009) used a Cronbach's alpha coefficient of reliability test to identify the factors responsible for post-harvest farm losses (Likert-type scale with a five-point Likert continuum). Kaminski and Christiansen (2014) applied the effects of the proximate determinants in a reduced form and causal analysis using (more) exogenous explanatory variables. Dharmathilake *et al.* (2019) conducted an acreage response analysis to identify relationships between the scale of farming operation and the post-harvest losses.

2.3.3 On-farm food loss models used to improve food loss survey data

Only a few studies targeted improving food loss estimates from survey data by using modelling approaches. A few studies combined different data collection methods to improve overall food loss estimates. This includes surveys that collected food loss data using different methods, such as declarations, physical measurement and visual scales, used models to identify biases or to obtain better food loss data through statistical pooling. Another application to improve survey data of food losses filled food loss data gaps or non-responses in the farm/household survey by using a food loss imputation model.

Model approaches to improve on-farm food loss data from surveys using different data collection methods:

As stated in the guidelines to measure harvest and post-harvest losses of cereals and pulses (GSARS, 2018b), food losses can be collected by farmers´ declarations or by actual/physical measurements. Delgado, Schuster and Torero (2017) and Delgado, Schuster and Torero (2020) analysed the differences and advantages and disadvantages among them and other approaches through a pairwise comparison of means. Jha *et al.* (2015) conducted national food loss surveys in India on several commodities by collecting both declared and physically measured food losses to improve overall food loss estimates by statistical pooling. No study was found to have applied a modelling approach to estimate the bias of one food loss data collection method against the other as suggested in chapter 1 as being one area to be covered by this research

Model approaches to improve on-farm food loss data from surveys by filling data gaps in surveys:

Apart from using combined methods of data collection to improve food loss estimates, data gaps and non-responses can be a problem in collecting food loss survey data for farm and household surveys. Hengsdijk and De Boer (2017) used a modelling approach to impute missing data of food losses within a national household survey. This represents one way to apply this research. The study used a random forest approach for the imputation exercise.

2.3.4 Identifying explanatory factors of on-farm food losses

One key result obtained from the literature review is a synthesis of the most relevant explanatory factors regarding on-farm losses. Based on the studies screened, an assessment matrix of the commonly used and relevant food loss drivers was generated. The results are presented in this section. Table 3 shows a general summary on the main thematic areas covered by most food loss models found in the literature review, which include socioeconomic characteristics of the household or producer, production characteristics, post-harvest activities, and weather and climate factors. Indicators on the post-harvest activities appear to be the most relevant group, while the factors on the production activity, weather and socioeconomic characteristics show mixed results.

Table 3. Groups of the main explanatory factors of food losses by the identified papers in the literature review

Categories of explanatory factors	Example indicators	≈Relevance
Socioeconomic characteristics (≈21 different indicators)	Age, education, gender	Less significant
	Income, credit access	Mixed results
	Communication, geographical area	Mixed results
	Family size	Significant
Characteristics of the agricultural activity (≈33 different indicators)	Size of the farm, production quantity	Significant with mixed results
	Farming experience, cooperation	Mixed results

	Input, labour	Less significant
	Timing of the harvest	Significant
Post-harvest activities (≈20 different indicators)	Market connection	Significant
	On-farm storage	Significant
	Packaging, transportation, threshing	Mixed results
Weather and climate (≈10 different indicators)	Weather conditions at harvesting	Significant
	Agro-ecological conditions/ zones	Mixed results
	HH/producer uses weather information	Less significant

2.3.4.1 Socioeconomic characteristics

Age and education: Age and education are the most used indicators to assess the relevance of the socioeconomic factors influencing food losses. To some extent, these indicators are used because they are mostly available in household and farm surveys. Nevertheless, the relevance of these indicators is limited and they are significant in most studies, independently of the region and commodity group. Arun and Ghimire (2019) in Nepal and Ansah and Tetteh (2016) on yam in Ghana, which found relevance for both indicators, are the only regressions that identified a relevance of age and educational level at the 5 and 10 percent significance level. The effect of education on losses was overly identified as being negative, which is in line with the hypothesis that more years of education may contribute to lower losses.

The age level of the producer, household head or respondent show mixed effects, some being positive and others being negative. To some extent, it is argued that older producers count with more years of experience in farming and post-harvest management. Because of that, food loss may be negatively correlated with age. On the other hand, younger producers may be adopting improved farming methods, which could be the reason why age is influencing food losses positively. The age level can also be related to the size of the farm and type of production system, as small scale subsistence farmers tend to be older than commercial farmers.

Experience in farming: A related indicator to the level of education is the years of experience in farming. The relevance of this indicator in the screened studies results is significant in all of them, which is a more stable correlation compared to the years of education. Although higher experience in farming is expected to have a negative impact on the level of food losses, with reduced losses when the respondent has more years of experience, the direction of the effect is mixed. The number of studies found to have either positive or negative impacts are equal but the negative correlation is generally confirmed at higher significance levels. The explication for the mixed results of the sign can be compared to those given for the age of the respondent, with older farmers more often being small scale subsistence farmers.

Gender: The gender of the producer, household head or respondent is another explanatory factor used in several regression analysis. In all of the studies, the gender on post-harvest levels was found to have a significant influence, varying at the 1 percent and 5 percent significance levels. The sign of the effect shows mixed results and the explanations of the relevance of the gender on the loss level are not straightforward. To some extent, gender can be related to general characteristics of the production

system, with women being excluded from the access to production inputs, technology and knowledge, which would assume that losses are positively correlated. On the other hand, it is argued, that women have more knowledge on post-harvest handling and are better able to conserve the food and reduce losses.

Size of the family: The most significant indicator among the socioeconomic characteristics is the size of the family based on number of family members. The reason for including this indicator is the hypothesis that larger households have more readily available and cheaper sources of labour to manage post-harvest losses effectively compared to smaller-sized households. Accordingly, for most studies, a negative correlation between the size of the family and food losses is expected. This hypothesis is confirmed by half of the studies, while the other half identified a positive correlation, which could be that the larger the household, the more diverse the responsibility for ensuring proper handling of crop production; alternatively, it could be that households have had to acquire more diverse plots to feed the household, and, therefore, there is more chance of loss. Two studies found no significance of the size of the family on post-harvest losses.

Income level and from farming: The studies screened show a significant impact of income and credit on the level of food losses of the studies screened. Compared to the other socioeconomic indicators, the type of indicator on income and credit varies, for instance, average annual income of the respondents, total household income, percentage of income coming from farming or agriculture as a primary occupation.

The income level is shown to be less significant to explain food loss levels. The literature suggests that wealthier households are better equipped to avoid crop loss, either because they have more resources or because they have access to better information. This is not confirmed by all studies, which either show no significance of positive correlation. No significance can be the result of a relatively homogenous sampling population, especially considering that most studies focus specifically on small-scale farmers and sub-national regions. A positive correlation can be argued in a similar way: The focus on small-scale farmers can imply that farmers, who produce more and generate higher incomes, need to handle larger quantities of crops in post-harvest operations, which can be the reason behind higher losses compared to subsistence farmers.

The percentage of income from farming appears to be significant and positively affects food losses, while the dummy variables on the households that have agriculture as the primary occupation is significant, but have mixed signs of effect. Agriculture as a primary occupation can indicate that households engage in more specialized farming, which can mean that more knowledge and production means are available.

Access to credit: Having access to credit is used by only two studies, and in both studies, there is a significant relevance. The sign of the effect is negative, which confirms the hypothesis that households that achieved access to credit are more formalized and organized and have access to production means and knowledge in harvesting and post-harvest management.

2.3.4.2 Characteristics of the agricultural activity

Size of the farm: Regarding the size of the farm, the most used indicators are the cropping area (ha), area of land cultivated (ha) or the area allocated for the crop being studied. On the one hand, it is argued that the greater the output a farmer results in greater probability of experiencing post-harvest losses. Then,

the sign of the coefficient is expected to be positive. On the other hand, the size of the farm can be positively related to the level of required investment, production and whether post-harvest management capacities are in place, which would imply a negative correlation between the size of the farm and food losses. With only a few exemptions, the studies reviewed included the farm size as explanatory variable, with mixed results. Several studies found no significance of the indicator for food losses, while others identified high significance at the 1 percent level. No pattern can be observed, independently of the country or region, the crop or commodity group, or the type of dependent variable (food loss quantities or food loss percentages) used in the regression analysis.

Quantity produced: The hypothesis on the effect of the produced quantity on the level of food losses is similar to the size of the farm. On the one hand, related to large scale production, farmers are expected to make greater investments and generate larger incomes and as a result have the capacity to adopt effective post-harvest management techniques. On the other hand, an increase in quantity harvested results in more complex post-harvest handling and an increase in losses when harvest and post-harvest management is not effective in place. The studies show that the impact of the quantity produced on food loss levels is highly significant, with most studies showing a significance at the 1 percent level. The sign of the effect is confirmed to be either negative or positive.

Special care is needed when interpreting the results of this explanatory factor, which differ considerably if the dependant variable is defined as food loss quantity compared to food loss percentage. Food loss quantities are likely to be positively correlated with the production level, while food loss percentages are observed to have a negative correlation with the size of the farm. Assuming food losses are structurally caused, food loss percentages are net production, with percentage losses observed to be higher for small scale farmers and lower for large scale or commercial farmers, while food loss quantities fluctuate with the production volume. It is, therefore, more likely that food loss quantities increase in line with increasing production output.

Input and technology: Within this group of indicators, different approaches can be taken to analyse the relevance of production technologies on the level of food losses. Some examples are the area under irrigation, indicators on input usage, including cost of improved seed, fertilizer, chemicals, labour, varieties used or having access to the extension service, number of extension visits per year or access to extension contact. The significance of these indicators is shown to be mixed, but they tend to be less relevant when explaining food loss levels. Homogeneity of the households in terms of input and technology can be one explanation for identifying an impact. Another reason may be that losses are more related to post-harvest operations and probably less to specific production characteristics. The mostly widely used indicator is the variety of seeds used, which shows significance in some of studies in which the indicator was employed.

Labour: Although less frequently used, some studies included indicators on the agricultural labour input. A major aspect considered is if the labour availability during harvesting was adequate, which is not significant in India, but is significant in Bangladesh with a negative effect. The type of labour used for harvesting, either family labour or hired labour, is another explanatory factor. In the labour-intensive harvesting and post-harvest operations of fruits and vegetables, additional indicators were included. For instance, the number of active labour force, pre harvest working days (man days) and harvest working days (man days). This indicator was found not to be statistically significant.

Cooperation and organization: A few studies included membership in producer cooperation and farmer organizations. The indicators vary from membership in producer organizations and farmer-based organizations and, membership in cooperative to purchase inputs jointly and marketing co-operatives for selling outputs This type of indicator is more commonly applied in models covering food losses in fruits and vegetable. The results on the significance of the membership on food losses are mixed and tend to be more relevant when the dependant variable used are food loss quantities instead of food loss percentages. The overall sign of the effect is negative, which is in line with the expected impact assuming that the membership in a producer organization contributes towards more efficient marketing of the fruits and vegetables and thus lower post-harvest losses.

Time of harvesting: Some studies used the time of harvesting as an explanatory factor for fruits and vegetable food losses. One of the indicators is the time of harvest after maturity (days) or the age of the fruit at harvest (months). Others used the decision to harvest if the fruit or vegetable harvested was mature or if the criteria to harvest is that the fruit is mature. Alternatively, a threshold was established to identify early harvest. The time of harvesting had a high significance for food losses in all studies in which the indicator was used. The sign of the effect is positive overall, with higher losses related to the produce that was harvested in a more mature state.

2.3.4.3 Post-harvest activities

The difficulties farmers and households encounter while engaging in post-harvest activities are key causes of food losses, making related explanatory factors are expected to be relevant. This differs from the previous factors on the socioeconomic characteristics and production activities, which are a proxy for the overall production structure and capacity for post-harvest management.

On-farm storage of grains, roots and tubers: Storage is a relevant post-harvest operation for on-farm food losses, especially for grains, roots and tubers. The indicators used as explanatory factors for storage are the existence of a storage facility, storage period in months or days of storage, the storage structure or type of storage or a storage variable if the storage facility was adequate. Similar to the literature that broadly confirms the importance of the storage facilities for post-harvest losses food losses, the type of storage shows to have a significant impact, but contrary to the expected sign of the impact, the studies indicate a positive effect of the access to storage facilities or adequate storage facilities on the level of food losses, although some studies confirm the negative relationship.

The period or length of storing the crop is expected to be relevant, but the results are mostly not significant. When significant, the sign of the effect is positive, with higher food losses related to a higher storage period. On the other hand, it is also argued that crops stored for a long time, better management techniques must be applied to prolong the shelf life of the produce.

On-farm packaging and storage of fruits and vegetables: Post-harvest on-farm packaging can be a relevant explanatory factor of fruits and vegetable food losses. The indicator used in the two studies are whether packaging is conducted or whether the packaging is suitable and worthwhile. When the packaging is suitable, the explanatory factor is significant with a negative effect on the level of food losses. The results associated with the indicator that assesses food losses shows that packaging is not significant.

On- and off-farm transportation: Transportation may be a relevant indicator, although it is more critical for food losses at off-farm stages instead of on-farm stages. The studies that assessed the impact the

adequacy of transportation facilities tended to find no significance on grains, roots and tubers. A different result may be suggested for fruits and vegetables, but the indicator was not included in the screened studies. The ownership of transport was another indicator used that resulted in not being significant. One study conducted in the United Republic of Tanzania used the number of livestock owned by the household as a means of transportation to move maize to markets, with livestock being linked to lower losses (Maziku, 2019). The studies that include off-farm stages could provide more insights into the effect of transportation on retail or wholesale losses.

Market connection and sales: Several studies included explanatory factors to examine the influence of market linkages on the level of post-harvest food losses. The most frequently used indicator is the distance to markets, which can be related to transportation losses or a lack of access to markets. Losses can then occur at the farm level if access to markets is hampered, or off-farm during transportation due to longer distances. The time to the nearest market has a similar interpretation. The overall results show that the indicator is significant for post-harvest food losses, including grains, roots, fruits and vegetables,, and are independent of the type of dependent variable used.

Other indicators, the area under commercial crops, the frequency of sales and the percentage of production used for own consumption, or the years of experience in markets are factors that can be used as proxies for the farmer´s degree of market integration. Only the area under commercial crops appeared to be significant for the quantity of food losses, showing a positive effect, which could be linked implicitly to the size of the farm and the quantity produced.

The sale price or current prices are explanatory factors that can indicate the type of market the product is sold, imply differences in the quality of the product or refer to different market situations (supply or demand excess). The two studies that used the indicator identified a significance at the 10 percent level with a negative sign. Food losses are, therefore, lower when sale prices are higher, assuming that farmers have a greater possibility to sell their production when prices are high (lower supply or higher demand).

2.3.4.4 Weather and climate

In terms of weather and climate conditions, general literature on food loss confirms the different implications they can have on food losses during harvesting and during post-harvest operations. The impacts on the quality of the produce, certain climate conditions during harvesting that can contribute to damages during harvesting or that affect the crop in post-harvest operations, for instance the moisture level of gains, and general regional agroecological conditions could affect food losses.

Weather conditions prior to and at harvesting: One group of explanatory factors aimed at identifying the weather at harvesting, with indicators requesting if the weather during harvesting was favourable or if rainfall/precipitation occurred during harvest and post-harvest. Alternatively, the month of harvest is used as an indicator to relate food losses to climate conditions. The results show that this indicator is overly significant and the consensus is on the positive sign of the effect, indicating that rain during harvesting contributes to higher food loss levels.

Another group of indicators reviews general weather conditions during the production period, using either a binary indicator on the weather condition (good or bad) or precipitation at preharvest, with rainfall prior to harvest indicative of overall production conditions and as a proxy for humidity patterns. Both indicators are overly significant. In one study, preharvest precipitation has a negative effect on

losses. More rain during the growing season should be indicative of a better-quality harvest that is less likely to be damaged or completely lost. The positive sign is confirmed by the second study, with good weather conditions contributing to lower losses.

Agro-ecological conditions: The overall agro-ecological conditions can also be proxied with agro-ecological zones that may have been established. In Nepal, altitude was used, while other studies chose a more general approach with a district dummy. This latter can include other regional effects apart from agro-ecological conditions, as there are socioeconomic differences among regions. Each of the explanatory factors was found to be significant. A positive effect between altitude and losses was observed, with higher altitude causing higher losses and belonging to a certain district that relates to more favourable agro-ecological conditions being associated with lower losses.

Use of weather information: The use of weather information or past weather experiences was found to be another explanatory factor. The result is overly significant, although mixed with commodities under which no significance was identified or a significance was only at the 10 percent level only. The use of weather information shows a negative effect, indicating that losses are lower if producers use climate data for decision-making. The use of past weather experience is positively related to food losses, which may indicate that these producers are more commonly affected by climate.

2.3.5. Conclusion of the literature review

In conclusion, the literature review cast out a considerable number of studies that examined the driving or determining factors of harvest and post-harvest based on farm or rural household surveys. Out of the 40 journal publications screened for this purpose, it was possible to single out different groups of indicators that seem to have a significant impact on losses and should, therefore, be considered when using modelling-approaches for on-farm losses. Although it is not always possible to standardize the combination of determining factors between different types of crops and countries, a common set of variables can be summarized that are usually available in farm or household surveys. Some key indicators repeatedly show significance, i.e., size of the farm/production quantity; timing of the harvest; on-farm storage; and weather conditions at harvesting. Additionally, some general elements for the model structure could be assessed, with a special emphasis on the type of loss indicator chosen for the independent variable. Although several studies used total loss quantities, or production quantities divided by total area harvested, the percentage loss seems to be a better fit to explain losses related to structural causes of losses. Some insights were also obtained on the method used for the regression models. While several studies had chosen multiple linear regression, some papers discussed the proper treatment of zeros, a relevant point considered in the following sections.

3. Food loss modelling approach to be combined with farm surveys

This chapter presents two modelling approaches that can be used to improve on-farm food loss estimates and reduce data collection costs.

The first modelling approaches are meant to be used through sub-sampling the food loss module in the farm survey. The main result indicated for this area of application is to use a post-stratification method derived from a classification and regression tree (CART). This reduces standard errors for food loss

estimates based on sub-samples, thereby improving the estimates to compensate for the reduced sample size.

The second modelling approach to be tested is to be used for consecutive survey rounds with food loss modules that cannot be included in all survey rounds. To produce loss estimates for in between missing years, a generalized structural equation model (GSEM) could be used to estimate losses, using only the information from the farm survey.

3.1. Support sub-sampling of food loss modules in farm surveys with modelling approaches using post-stratification

As mentioned earlier, the first area of application is to use modelling approaches to support the strategy of sub-sampling of food loss modules in farm surveys. To do this, four data sets from farm and household surveys were available, namely two pilot surveys, which were conducted on cereal crops in Malawi and Zimbabwe as part of the Global Strategy to Improve Agricultural and Rural Statistics (GSARS) in 2018, and were specific post-harvest loss surveys, the Fourth Malawi Integrated Household Survey 2016/17 (IHS4) and the Nigeria General Household Survey 2015/16 (GHS 2015/16), which was conducted by the World Bank.

3.1.1 The process of testing the modelling approaches

Dependent variable

As a first step and based on the insights obtained from the literature review, multiple linear regression models were tested by choosing a set of independent variables related to harvest and post-harvest known to be relevant for losses. On the other hand, different dependent variables proposed in the literature were thereby assessed to determine whether to use total losses or percentage losses, total post-harvest losses or losses disaggregated by operation (harvest, cleaning, drying, storage, etc.). The first conclusion, on-farm loss percentages seem to be a better suited dependent variable than total on-farm losses for the given survey datasets. Percentage losses appear to better indicate the structural problems causing losses and the efficiency of handling the grains, while the quantity of losses is to some extent driven by the production volume.

Model approach

As the recorded percentage losses show a positive skewed distribution, the use of the natural log transformation of the percentage losses is suggested. A linear regression could be used to generate a model to predict mean percentage losses, but, in that case, predictions from the model are in log scale and the reverse transformation results in a bias of the estimated mean losses. These could be corrected by including a function of the variance of the errors in the estimated mean losses. Nevertheless, a Poisson regression model could be a better alternative to log-linear regression because the link function of this model is the natural log of the response. Additionally, a Poisson regression handles outcomes that are true zeros, while a log regression does not consider zeros because of ln(0) is -∞. Poisson distribution assumes that the expected response equals the variance of the response, so the use of robust standard errors is useful to handle these assumptions.

Post-stratification

The second step entailed testing if post-stratification could be used with the main idea to improve the efficiency of the parameter estimates obtained from the sample survey and respective sub-samples. As stated by Smith (1991), it can be a useful method to reduce variance and correct for possible bias, and in this case, no external information was used. To do this, the proposal was to use CART to generate the post-stratification. The output of the CART is a decision tree in which each end node represents a stratum with a final prediction for the outcome variable, in this case on-farm losses. The algorithm selects the relevant independent variables and their respective cutting point, where the difference of the mean response for the resulting groups are maximized. Consequently, part of the variance in the sample survey is explained by the mean differences between the resulting groups. To make use of post-stratification in the modelling approach, the results of the classification and regression tree are used to set the estimation model, where the classification variable is used as the predictor and the mean prediction is used as the estimator of the mean on-farm loss. Then, the models are tested to determine whether they (i) are sufficiently well-specified to provide reliable estimates, and (ii) reduce the standard error as an assumed effect of the post-stratification procedure.

Model specification test

To test if the model is well suited, the model specification test (linktest) (Pregibon, 1980) is used to evaluate if the proposed models are correctly specified and determine if the post-stratification is a good food loss predictor. This test uses the linear predictor value $X\hat{b}$ and linear predictor value squared $(X\hat{b})^2$ as the predictors to rebuild the model (X represents the predictor variables and $\hat{b}$ the estimated model coefficients). The variable $(X\hat{b})^2$ should have no predictive power and the estimated parameter should be zero. On the contrary, if $(X\hat{b})^2$ is significant, the linktest is significant, meaning that relevant variables have been omitted, or that the link function is not correctly specified. In this case, it implies a model with lack of fit, which is of limited use for prediction purposes. On the other hand, $X\hat{b}$ should be close to 1, which is considered a good linear predictor.

It is expected that this modelling approach not only provides good and reliable on-farm loss estimates, moreover they simplify the modelling procedure and can help to improve the efficiency of the mean estimate by a reduction of its standard error. These gains, in turn, can support data collection on relatively small sub-samples. Data collection costs could be optimized without considerably compromising data quality, resulting in a complementing strategy to be used to design national data collection of losses within national farm and household surveys.

3.1.2 Evaluate the gains of the modelling approach for sample size reduction

The main idea of this research is to make use of the improvements in the mean estimate obtained from post-stratification to reduce on-farm loss data collection to a sub-sample. The models are assessed based on their capacity to produce estimates, which should not deviate considerably from the estimates obtained from the full sample. Additionally, a measure of efficiency is defined to evaluate whether the model-based estimates provide considerable gains that can be used for a reduction of the sub-sample of losses. Here, the ratio of the model-based variance to the actual full sample variance of an estimator of losses is built to express the efficiency gains in terms of a reduction in the variance. The relative efficiency of a model-based estimate compared to the full sample-based estimate is then:

$$RE = 1 - \frac{V(\widehat{L_m})}{V(\hat{\hat{L}})} \quad (1)$$

Where $V(\widehat{L_m})$ represents the variance for the loss estimate based on the specified model (implying a post-stratification effect), and $V(\hat{L})$ represents the variance for the survey-based loss estimate from the full original sample. RE is then the percentage of variance reduction.

To test the possibility for a reduction of the sub-sample, a simulation is run on the given the survey datasets. To create sub-samples, a progressive random elimination of 10 percent of the sample to a maximum of 50 percent reduction is conducted, generating five sub-samples. For the full sample and for each of the sub-samples, model-based mean loss estimates are obtained and compared to the full sample estimate sampling theory-based using a measure of relative efficiency (RE):

Sample-based loss estimates from the full sample $\hat{L}$ and its standard error $\sqrt{V(\hat{L})} = \hat{\sigma}$ and sample-based loss estimates from sub-samples $\widehat{L_s}$ and standard errors $\sqrt{V(\widehat{L_s})} = \widehat{\sigma_s}$, (subsequently called sample-based estimates)

Compared with:

Model-based loss estimates from the full sample and sub-samples, $\widehat{L_m}$ and its standard error $\sqrt{V(\widehat{L_m})} = \widehat{\sigma_m}$ (subsequently called post-stratification model-based estimates)

In addition to the model-based estimates, the specified loss models can also be used to impute missing values. To simulate the imputation of possible data gaps, the on-farm loss model is used here to impute losses in the sub-samples to create the full imputed sample. This exercise has some known limitations, especially when imputation techniques are applied to a larger proportion of the sample and lead to an artificial reduction of the standard error of the mean estimates. Accordingly, the exercise is presented in the results section, but highlighting the limiting interpretation of the standard error.

For the purposes of comparison, the following is defined:

Sample-based loss estimates from the full sample composed of a model-based imputed sub-sample $\widehat{L_\iota}$ and its standard error $\sqrt{V(\widehat{L_\iota})} = \widehat{\sigma_\iota}$.

3.1.3 Datasets available from farm surveys in Malawi and Zimbabwe

3.1.3.1 Global Strategy to Improve Agricultural and Rural Statistics farm loss surveys in Malawi and Zimbabwe

The first set of available farm survey loss data comes from the field tests conducted for the "Guidelines on the measurement of harvest and post-harvest losses" (GSARS, 2018a) in Malawi (FAO, 2020a) and Zimbabwe (FAO, 2020b).

These farm loss surveys were implemented in 2017 and 2018 at a sub-national level, covering the Salima and Lilongwe districts in Malawi for maize, rice and groundnuts, and the Makonde district in Zimbabwe for maize. The analysis focused on maize only, the most important staple food in the surveyed countries, and the crop for which the most data are available in the available surveys. The sample size in Malawi

achieved 447 observations for maize crops, in the case of Zimbabwe 307 observations were obtained for maize. In each of these regions, agricultural production is the main source of livelihoods. Average area harvested is 1.2–2.7 hectares per household in Zimbabwe and 0.4− 0.6 ha in Malawi, predominantly rain-fed and based on manual harvesting methods (see Table 4 for a summary of descriptive statistics of the main variables).

Although data were collected for local, hybrid and composite maize, the modelling approach was applied to the aggregation of all varieties of maize. This is because on the one hand, recent empirical evidence highlights problems of misclassification with respect to farmers declaring seed varieties (Woosen *et. Al.* 2019, Wineman *et.al.* 2020). On the other hand, the sample sizes of the GSARS surveys are relatively small for testing modelling approaches. The main variable of interest − the percentage of losses over production was calculated as the total quantity of maize losses (from harvest to storage) divided by the quantity of land cultivated with maize.

Table 4 Descriptive statistics of the variables relevant for the model (GSARS)

Country		Malawi GSARS			Zimbabwe GSARS*	
Variable	N	Mean	Std Dev	N	Mean	Std Dev
Loss percentage (harvest+post-harvest)	356	9.23	10.62	307	4.84	10.68
Crop production (Kg)	358	1105.30	1591.86	307	5905.64	9030.31
Age	357	45.43	14.70	307	50.15	15.92
Harvest length (days in average)	352	5.12	4.51			
Area planted (ha)				305	1.52	1.15
Percentage area harvested	358	0.51	0.45			
Variable	N	%		N	%	
Household head − Gender	357			307		
Female	84	23.53%		69	22.48%	
Male	273	76.47%		238	77.52%	
Household head − Education level	358			307		
No education	76	21.23%		42	13.68%	
Primary school	229	63.97%		106	34.53%	
Secondary school	53	14.80%		159	51.79%	
Household head - Literacy	358			307		
Yes	271	75.70%		264	85.99%	
No	87	24.30%		43	14.01%	
Thresh/shell the harvest	354			307		

	Yes	341	96.33%	293	95.44%
	No	13	3.67%	14	4.56%
Clean/winnow the harvest		344		307	
	Yes	210	61.05%	228	74.27%
	No	134	38.95%	79	25.73%
Harvest drying method		328		307	
	No dry	111	33.84%	4	1.30%
	Manual	165	50.30%	302	98.37%
	Mechanical	52	15.85%	1	0.33%
Use of high-tech storage		358		307	
	No storage	24	6.70%	4	1.30%
	No	286	79.89%	303	98.70%
	Yes	48	13.41%		
Use of pesticides during storage		318		305	
	Yes	149	46.86%	253	82.95%
	No	169	53.14%	52	17.05%
Assistance from government or NGOs		345		307	
	Yes	165	47.83%	221	71.99%
	No	180	52.17%	86	28.01%
**The results for Zimbabwe can be found in annex 4.*					

3.1.3.2 Living Standard Measurement Study – Integrated Survey on Agriculture (LSMS-ISA) in Malawi and Nigeria

This study uses two datasets from nationally representative household surveys in Malawi and Nigeria: the Fourth Malawi Integrated Household Survey 2016/17 (IHS4);[6] and the Nigeria General Household Survey 2015/16 (GHS 2015/16)[7]. The surveys are part of the Living Standards Measurement Study – Integrated Survey on Agriculture (LSMS-ISA). They contain an integrated household and agricultural component. The household survey component collects detailed socioeconomic information, including household-level data on consumption, income, assets and housing, and individual-level data on demographics, education and health. The agricultural component collects detailed information, among other items, on agricultural inputs used and outputs produced and output disposition, at the plot-level. Particularly important for the analysis is that information on inputs are collected at the plot level; information on agricultural output is

[6] The microdata, survey report and basic information document about the GHS 2015/16 implementation are available at https://microdata.worldbank.org/index.php/catalog/2936.

[7] The microdata, survey report and basic information document about the GHS 2015/16 implementation are available at https://microdata.worldbank.org/index.php/catalog/2734.

collected at the crop/plot level; and output disposition is collected at crop level. In addition, the IHS4 2016/17 and GHS 2015/16 datasets include several exogenous climatological and geospatial variables. These include measures of distance, climatology, soil and terrain, and other environmental factors. Time-series data on rainfall and vegetation have also been used to describe the survey agricultural season relative to normal conditions.

The IHS4 2016/17 is the fourth wave of the Integrated Household Survey and includes 12,480 households surveyed in 780 enumeration areas. Households were visited one time throughout the 12 months of fieldwork between April 2016 and April 2017. The GHS 2015/16 is the third wave of the Nigeria General Household Panel and includes 4,581 households surveyed in 500 enumeration areas that were visited two times, between September 2015 and April 2016, one time right after the end of the planting activities and one time right after the end of the harvest activities relative to the 2015/16 rainy season. Given this two-visit setup, the length of the recall period of the GHS 2015/16 is shorter than length of the recall period of the IHS4 2016/17. As with the GSARS surveys in Malawi and Zimbabwe, the samples for the IHS4 2016/17 and GHS 2015/16 were restricted to maize (all varieties of maize were aggregated).

The Malawi IHS4 and Nigeria GHS 2015/16 samples were restricted to observations reporting positive values for losses, and therefore excluded zero observations. The very high percentage of zero losses shed doubts on the accuracy of such zero reported losses, particularly given the very low percentage of zero losses in the GSARS surveys. The assumption here is that in LSMS-ISA surveys, farmers, in some cases interviewed several months after the harvest, might have reported only substantial losses and omitted marginal losses (see section 4.3 for an explanation of the differences in the scope and methodologies between GSARS and LSMS-ISA surveys). Including "false" zero losses could lead to downward-biased estimates. Although excluding zero losses could lead to biased (in the case that a certain percentage of reported zero losses are "true" zeros) or partial estimates (which, by definition, apply to the specific case of farmers reporting non-zero losses), after cross-checking loss variables with other strongly correlated variables, such as whether the crop was stored, the decision was made to restrict the sample to positive loss observations for the estimation of models and the simulation of sub-sampling scenarios, with the clear statement that the results from LSMS-ISA surveys in the study refer to farmers reporting non-zero losses. To assess the potential bias introduced by the sample restriction, the probabilities of non-zero and zero reporting were tested using propensity scores technique and the significant difference in characteristics between the those reporting zero losses and those reporting positive losses. Both tests show that the two groups are not systematically and significantly different (not shown here, available upon request) (see Table 5 for a summary of descriptive statistics of main variables).

Table 5: Descriptive statistics of the variables relevant for the model (LSMS-ISA)

Country	Malawi IHS 4			Nigeria GHS 15/16*		
Variable	N	Mean	Std Dev	N	Mean	Std Dev
Loss percentage (post-harvest)	1852	12.76	19.79	253	10.12	13.84
Crop production (Kg)	1852	438.05	387.21	253	1048.88	1455.94
Household head – Age	1850	44.72	16.31	253	52.91	13.38
Harvest length (days in average)	1852	17.70	12.72	248	38.13	33.92

Area planted (ha)		1852	0.29	0.25	253	0.75	1.02
Plot distance to household (Km)		1678	1.24	7.43	251	1.45	3.77
Plot slope		1675	4.96	5.35	251	2.93	2.16
Plot elevation		1675	901.57	306.37	251	294.74	239.94
HH distance to market (Km)		1852	24.21	14.32	253	75.30	35.72
Variable		N	%		N	%	
Household head – Gender		1852			253		
	Female	578	31.21%		47	18.58%	
	Male	1274	68.79%		206	81.42%	
Household head – Literacy		1852			253		
	Yes	192	10.37%		87	34.39%	
	No	1660	89.63%		166	65.61%	
Improved seed		1852			253		
	Yes	853	46.06%		31	12.25%	
	No	999	53.94%		222	87.75%	
Agro-ecological zones		1852			253		
	Tropic-warm/semiarid	959	51.78%		91	35.97%	
	Tropic-warm/subhumid	535	28.89%		147	58.10%	
	Tropic-cool/semiarid	196	10.58%		12	4.74%	
	Tropic-cool/subhumid	162	8.75%		3	1.19%	
Region Malawi		1850					
	North	318	17.19%				
	Central	714	38.59%				
	Southern	818	44.22%				
Region Nigeria					253		
	North Central				40	15.81%	
	North East				40	15.81%	
	North West				54	21.34%	
	South East				84	33.20%	
	South				19	7.51%	
	South West				16	6.32%	

**The results for Nigeria are shown in annex 4.*

3.1.3.3 Differences in the datasets in the survey design regarding farm losses

The on-farm loss data obtained from the GSARS harvest and post-harvest loss surveys and the LSMS-ISA surveys differ from each other in terms of their design and data collection method. First, the GSARS data stems from a survey specifically designed to measure on-farm losses. With all the focus placed on the loss indicators, the questionnaire includes loss questions disaggregated by on-farm activities (harvest, threshing, winnowing/cleaning, and storage) and complementary data on socioeconomic, production and post-harvest characteristics. Additionally, on-farm loss data were collected by farmers' declarations and by physical measurement. On the other hand, the scope of the survey is the local level, where it is probable that the population is less diverse than at the national level.

The LSMS-ISA surveys, on the contrary, are designed to estimate agricultural production and productivity of the rural households in which losses are covered as one complementary indicator among various in the crop disposition section. It is collected by one sub-question on the destination of the harvested production, where farmers declare total post-harvest losses among quantities self-consumed, sold, given away as a gift, and used as seeds and as animal feed, without detailing the post-harvest activities. Harvest losses are not considered. The set of variables collected in the survey are less tailored to on-farm losses, but they cover a broader range of socioeconomic, production and environmental characteristics. The survey is nationally representative, whereby it is assumed to cover a more heterogeneous population compared to the GSARS surveys. Both surveys allowed for reporting loss and production quantities in local non-standard units, converted into kilograms using correspondence tables specific to each survey.

For the 50x2030 Initiative, a similar questionnaire structure to the GSARS harvest and post-harvest loss survey was developed for the corresponding loss module recommending a more detailed assessment of harvest and post-harvest losses. On the other hand, the 50x2030 Initiative seeks nationally representative surveys, whereby the LSMS-ISA survey helps to better understand the implications of nationwide data in the modelling approach.

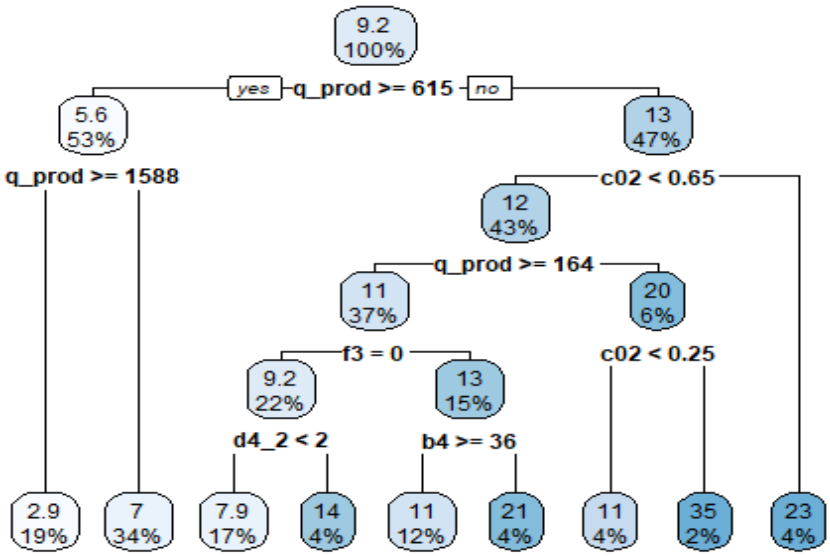
3.1.4. Results of applying the modelling approach: estimation of the food loss models for Malawi

3.1.4.1 Models obtained for the GSARS harvest and post-harvest loss surveys in Malawi

Based on sampling theory, the estimated percent loss of maize using data from the GSARS survey in Malawi is 8.66 percent harvest and post-harvest losses with a standard error of 0.514 percent (95 percent% CI: 7.66, 9.67). One of the best theory-based models for this sample, which includes selected variables and some interactions (not shown), predicted a mean estimate of 8.39 percent loss with a standard error of 0.502 (95 percent CI: 7.41, 9.38).

CART method: To improve the efficiency of the estimator, a regression tree was built to generate post-stratification criteria to specify the loss estimation model. The regression tree for the GSARS loss survey in Malawi is shown in Figure 1. It selected eight cutting points on five variables, namely the quantity of maize produced (q_prod), the age of the household head (b4), the percent of the area harvested (c02), whether the household received any assistance from the government (f3) and the drying method used (d4_2). This tree generated nine terminal nodes with different mean percentage losses. It can be observed that the population was divided in different classification groups, starting at the level of production (q_prod) either being above or below 615 kgs. The difference between loss levels can be seen, namely 13

Figure 1. CART classification for Malawi



percent for those with lower production, and 5 percent for those with larger production. The larger farmers were then further divided by those producing more than 1,588 kgs, showing a similar trend in terms of the loss percentages (2.9 percent for those who produce above that threshold, and 7 percent who produce less). For the producers with smaller production quantities, the classifications were built with the additional variables mentioned above. The highest loss percentages appeared to be among farmers who harvested less than 25 percent of an area and produced less than 164 kgs. Possibly, they faced damages that are also reflected in in losses at harvest and post-harvest.

Model specification: The classification established as post-stratification was used as a predictor variable in a Poisson model to generate a percent loss estimate. The output from this model is shown in table 6.

Table 5: Model with CART classification as regressor for the Malawi GSARS survey.

Poisson regression				Number of obs	=	356
				Wald chi2(8)	=	156.22
Log pseudolikelihood = -1511089.3				Prob > chi2	=	0.000
% HPHL	Coefficient	Robust Standard Error	Z	P	95% LL	95% UL

2	0.875	0.163	5.36	0.000	0.555	1.195
3	0.943	0.206	4.59	0.000	0.540	1.346
4	1.346	0.292	4.61	0.000	0.773	1.919
5	1.328	0.193	6.88	0.000	0.950	1.707
6	1.930	0.244	7.91	0.000	1.452	2.409
7	1.377	0.355	3.88	0.000	0.682	2.073
8	2.234	0.218	10.23	0.000	1.806	2.662
9	1.850	0.237	7.8	0.000	1.386	2.315
constant	1.097	0.147	7.44	0.000	0.808	1.386

This is a parsimonious model[8] that uses only one classification variable as the predictor, but it includes three independent variables in the classification criteria. To test the use of a Poisson model (natural log as link function) and the specification of the model with respect to the independent classification variable, the corresponding linktest is shown in Table 7. This test indicates that the specified on-farm loss model has very good functionality for estimating the mean percent losses, where the linear prediction $\hat{L}$ presented a significant coefficient equal to 1, meaning a perfect correspondence (1:1) to the observed percent losses (p = 0.034), and the square predicted $\hat{L}^2$ had no predictive power (p = 1), with an estimated coefficient equal to zero. This is the ideal situation for model-based predictions. The estimated percentage loss of maize using the Poisson model was also 8.66 percent, but it had a smaller standard error of 0.429 percent (95 percent CI: 7.8, 9.5). This improved variance can be attributed to the loss classifications identified in the post-stratification procedure of the CART method. It showed an efficiency increase of 30.3 percent from the sample-based standard error to the model-based standard error.[9]

Table 6. Model specification tests, results from the link tests Malawi GSARS

Malawi GSARS			
Predictor	Coefficient	Std. Err.	P
$\hat{L}$	1.000	0.471	0.034
$\hat{L}^2$	0.000	0.104	1.000

[8] Brief description: A parsimonious model is one that accomplishes the desired level of explanation or prediction with as few predictor variables as possible. For more details see www.statisticshowto.com/parsimonious-model/

[9] The efficiency increase or "model relative efficiency" (RE) is the percentual difference between the standard error obtained from the sample-based loss estimate on the full sample to the standard error obtained from the model-based loss estimates on the full sample or any size of the sub-samples.

3.1.4.2 Models obtained for the Malawi LSMS-ISA surveys

The sampling base estimate for the percent loss of maize in Malawi using data from the Living Standard Measurement Studies showed a point estimate of 14.5 percent with a standard error of 0.661 percent (95 percent CI: 13.2, 15.8). To improve the efficiency of the estimator for the post-harvest percentage losses of maize, the regression tree for the post-stratification criteria was derived, as shown in Figure 2.

The regression tree selected three cutting points on three variables: crop harvested production (imp_production); IHS4 2016 region location (region); and agro-ecological zone (hhgv_ssa_aez09). This tree arrived at four terminal nodes with different mean percentage losses. The Poisson model for this sample apparently fit properly; the linktest shows very good functionality in estimating the mean percentage of losses. Linear prediction $\hat{L}$ presents a correspondence of 1:1 to the observed food losses, but the coefficient appeared to be not significantly different from zero (p = 0.502). This was related to a bigger standard error of the estimated coefficient and probably implies weak predictive power. The square predicted $\hat{L}^2$ had no predictive power, and the coefficient was practically zero (p = 1), so the model passed the linktest, as shown in Table 8.

The estimated percentage loss of maize using this model was also 14.5 percent, but with a slightly smaller standard error of 0.621 percent (95 percent CI: 13.3, 15.7). This represented an efficiency increase of only 12 percent. The independent variables included in this model, which were not specific to harvesting procedures and referred to general aspects of production places, showed that recorded information is less related to food losses.

Figure 2: CART classification for Malawi IHS4 Survey

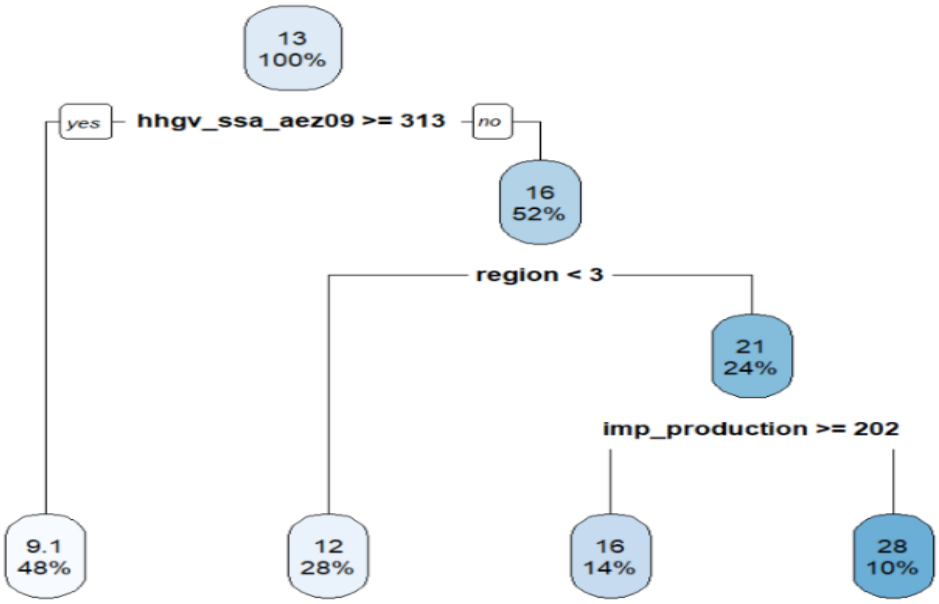


Table 7 Model specification test results (linktest), for the Malawi IHS4 sample.

Malawi IHS4			
Predictor	Coefficient	Std. Err.	P
$\hat{L}$	1.000	1.495	0.503
$\hat{L}^2$	0.000	0.264	1.000

3.1.4.3 Results for sub-sampling obtained for the GSARS farm loss surveys in Malawi

To evaluate changes in the relative efficiency of food loss estimates based on a progressive sample size reduction by simulating a reduction of survey data collection, models on sub-samples were estimated using the regression tree stratification obtained from the full sample. Table 9 shows the simulated sample reduction and the corresponding estimates for the percentage food loss for maize based on survey theory estimates and the model-based estimates.

Table 8 Estimates, standard errors and relative efficiencies for sub-samples – Malawi GSARS.

Sample reduction	Survey-based loss estimate		Post-stratification loss estimate (model-based)		Model relative efficiency (RE)
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_m$	$\hat{\sigma}_m$	
0%	8.66	0.514	8.66	0.429	30.3%
10%	8.58	0.543	8.58	0.441	26.4%
20%	8.68	0.584	8.68	0.474	15.1%
30%	8.42	0.592	8.42	0.481	12.6%
40%	8.51	0.671	8.51	0.556	-17.1%
50%	8.34	0.720	8.34	0.619	-44.7%

With a sample reduction of 30 percent, the model estimate still presented good efficiency of 12.6 percent compared to the survey estimate using the whole sample, meaning that using models can help to improve food loss estimates with an important reduction in survey data collection. An alternative to sample reduction is to impute missing values to enhance model estimates. In Table 10, survey estimates and model-based estimates using imputed missing values on the reduced part of the sample are presented.

In this case, an important concern arises because the use of imputed samples reduces the standard errors $\hat{\sigma}_i$. This can be the result of the use of imputed values estimated as the most reliable expected for each missing data point and can result in artificially smaller standard errors than expected, representing a hidden reduction on the coverage of probability intervals respect to the assumed a priori.

Table 9 Estimates, standard errors, and relative efficiencies for imputed sub-samples – Malawi GSARS.

Sample reduction	Survey-based loss estimate		Estimates with model-based imputation	
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_i$	$\hat{\sigma}_i$
0%	8.66	0.514	8.66	0.514
10%	8.58	0.543	8.93	0.506
20%	8.68	0.584	8.73	0.483
30%	8.42	0.592	8.79	0.476
40%	8.51	0.671	9.15	0.470
50%	8.34	0.720	9.05	0.438

3.1.4.4 Results for sub-sampling obtained for the Living Standard Measurement Studies in Malawi

To evaluate changes in the relative efficiency of food loss estimates based on a progressive sample size reduction, the LSMS-ISA from Malawi is reviewed. Table 11 shows the results of sample-based and model-based percent loss estimates for simulated sub-samples.

Table 10: Estimates, standard errors and relative efficiencies for sub-samples – Malawi IHS4

Sample reduction	Survey-based loss estimate		Post-stratification loss estimate (model-based)		Model relative efficiency
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_m$	$\hat{\sigma}_m$	
0%	14.52	0.661	14.52	0.621	12.0%
10%	14.58	0.700	14.58	0.656	1.5%
20%	14.32	0.741	14.32	0.699	-11.6%
30%	14.25	0.798	14.25	0.756	-30.7%
40%	13.79	0.837	13.79	0.777	-38.1%
50%	13.65	0.892	13.65	0.835	-59.4%

In this case, the gains obtained from model-based estimates are limited compared to a sample-based full sample estimate in the standard error. This can be due to the low quality of information related to on-

farm loss determinants obtained in the LSMS-ISA survey compared to GSARS, which considers specific post-harvest loss data collection.

In Table 12, survey and model estimates using imputed missing values on the reduced part of the sample are shown. As mentioned above, imputation can lead to artificially smaller standard errors.

Table 11. Estimates, standard errors, and relative efficiencies for imputed sub-samples – Malawi IHS4

Sample reduction	Survey-based loss estimate		Estimates with model-based imputation	
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_i$	$\hat{\sigma}_i$
0%	14.52	0.661	14.52	0.661
10%	14.58	0.700	14.37	0.626
20%	14.32	0.741	14.26	0.586
30%	14.25	0.798	13.98	0.538
40%	13.79	0.837	14.42	0.521
50%	13.65	0.892	13.86	0.470

3.1.5 Conclusions on the use of modelling approaches with sub-sampling food losses in farm surveys

The results from using modelling approaches to support sub-sampling strategies for on-farm loss data collection were generally positive in cases in which loss models built from a post-stratification procedure showed sufficiently good prediction performance and efficiency gains on mean estimates. The data-driven post-stratification models improved model-based estimates obtained from full sampling and sub-sampling compared to the survey-based estimates. Sub-sampling is thereby possible, without causing a considerable loss in the quality of the estimates. The models provided some scope for further reductions of the sub-sample.

These results were stronger for the GSARS farm loss survey datasets compared to the LSMS-ISA survey data, which is to some extent related to the specific survey design used in the GSARS farm loss survey that builds on a detailed assessment of food losses. On the other hand, the GSARS farm loss survey was conducted at a sub-national level, covering a less heterogeneous population compared to the nationwide conducted LSMS-ISA surveys. The set of variables chosen to specify the classification groups follow the general knowledge of on-farm loss causes in which small-scale farmers tended to have higher loss percentages compared to large-scale producers. Production levels were, therefore, a structural variable that may signal efficiency of harvest and post-harvest procedures. Socioeconomic variables, such as the age of the producer, provide additional criteria for explaining on-farm loss differences, as older farmers tended to face higher losses. Access to technical assistance and the drying methods used were other relevant variables found to be specific to the model. Apart from those chosen by the CART method, other variables from the GSARS farm loss survey appeared be relevant to explain farm losses, such as storage

technology, use of pesticides during storage, days of harvesting and harvest methods. These are aligned with the literature review and were identified when testing the theory-based farm loss models, but are less relevant for prediction purposes.

For the LSMS-ISA surveys, the model application in the case of Malawi resulted in little improvements obtained from modelling on the loss estimate, while the model-based estimations for Nigeria showed considerable improvements. It is, therefore, important to highlight the need for good quality data to obtain the expected gains from the use of the modelling approach. In the case of Malawi, the sample was spread over a twelve-month period and the recall period varied substantially within the sample if the fieldwork was completed after several months after the harvest period, while in Nigeria, two visits were undertaken close to the planting and harvest periods. The shorter periods between the survey and the harvesting period in the Nigeria survey likely helped make the responses more accurate (Smith, 1991; Pregibon, 1980; FAO, 2020a). The results obtained for the LSMS-ISA survey in Nigeria showed that the modelling approach worked on national surveys, although these might not focus on food loss measurement as the main survey objective and were, therefore, composed of a wider but less loss-specific set of variables. It is interesting to observe, that, apart from production quantities and area planted, the region and agro-ecological zone and the plot characteristics accounted for most of the effects identified in the CART Method for prediction purposes. These are valuable insights for the set of relevant variables to be included in farm surveys, a conclusion supported by the literature review, although a different approach was used by GSARS surveys where these factors have not been collected or included in the available datasets.

The applications with respect to model-based imputation procedures to extrapolate the estimates from the sub-sample to the full sample show that using them is limited for improvements of on-farm loss estimates. Model-based imputations above 10 percent of the sample-size reduction started to show an artificial reduction of the standard error of the full sample. Accordingly, the model-based imputation should not be applied to extrapolate larger datasets. The artificial reduction of standard errors can result in potential errors in the use of confidence intervals, representing a risk for decision making. Overall, model-based imputations showed better results than median-based imputations and can, therefore, be used as a complement to fill possible non-responses.

Some relevant conclusions can be derived from the information that can support farm loss modelling on sub-samples and are, therefore, recommended to be considered in survey design and implementation. As suggested in the 50x2030 Initiative sampling guidelines (50x2030 Initiative, 2021b), sub-sampling for the on-farm loss module in household and farm surveys is recommended for optimizing data collection costs. The saved resources could then be invested in a more precise assessment of losses, either by detailing declarations or by using other objective methods to improve the estimates. Indeed, for the given surveys, a more detailed loss module seems to provide better loss estimates and avoid unreliable zero-response rates, an area where further research is needed. Investments in better quality data with a reduced sample size also pay off in stronger loss models, which, in turn, helps to sustain the sub-sampling strategy. Finally, geo-references from LSMS-ISA are shown to be very useful for adding climate and plot characteristics to the survey data which, in turn, can be relevant for building farm loss models. Accordingly, given the minimal implementation burden for the enumerators, capturing Global Positioning System (GPS) coordinates of the surveyed household/dwelling and plots is highly recommended.

Recognizing, on the one hand, the importance of collecting data on losses to inform policymaking, and, on the other hand, the complexity of collecting such data, the results highlight possible gains from the integration of survey data with modelling to improve the quality of loss estimates and support sub-sampling strategies. The combination of sub-sampling a detailed module on post-harvest losses with a modelling approach can be a useful recommendation for large-scale household and farm surveys, and through this approach, the 50x2030 Initiative could present the ideal opportunity for scaling-up the collection of data on post-harvest losses.

3.2 Modelling approach for estimating farm losses with two independent survey rounds as alternative to reduce data collection costs

A second objective of using food loss modelling approaches with national farm surveys is the possibility to rotate the food loss module in the farm survey, a strategy used to reduce the lengths of the farm surveys and optimize survey costs. Some modules rotate from one survey round to the other, whereby food losses are collected only every second or third round of the farm survey. This can be relevant when farm surveys cover diverse thematic areas and each one has a specific set of indicators. To avoid jeopardizing the quality and completeness of the survey results due to fatigue of the respondents, a rotational modular survey system can be chosen. While a core module with only a few critical questions is kept for collecting the key variables, the rest of the modules can be used in alternate survey rounds. This strategy is suggested, for instance, in AGRI-Survey (GSARS/FAO, 2018) and the 50x2030 initiative´s guidelines (50x2030 Initiative, 2021b).

Within this strategy, a modelling approach could complement and foster this strategy. Loss modules can be applied in to two or more survey rounds with the application. On these survey-based variables, a modelling approach can be tested for producing model-based loss estimates for the survey rounds that do not apply the food loss module. To do this, two distant survey rounds can be used to construct a GSEM to generate predictions of the mean responses of food losses, using a wide variety of independent variables as determinants, which are included in the core module of the farm survey. These could include socioeconomic characteristics, such as age, gender, level of education, harvesting methods and number of harvesting days, the use of improved seed, post-harvest technology used, information on the type of storage facility used, storage duration, use of pest control products during storage and information on region, weather and soil production conditions.

3.2.1 Overall modelling approach to be tested

In terms of the overall approach, GSEM was used for this area of application. This modelling approach has the advantage of being able to predict the dependent variable (here food losses) by identifying the interrelated effects (or paths) of the independent variables that can be used to produce model-based indirect estimates of food losses for surveys that collect only information on the predictors considered in a GSEM model.

In this context and given the set of indicators from the farm and household survey, as a first step a GSEM model was built to create a system of relationships between predictor variables and the observed food loss using two distant year survey rounds that collects information of useful predictor variables and the food loss module. GSEM was then used to identify structural driver contributions to food loss. If the variables in the farm or household surveys are sufficient to specify a GSEM with reasonable prediction power, it can be a p tool for indirect estimation of food loss.

In 1934, Sewall Wright introduced a model based on diagrammatic causal trajectories between variables called Path Analysis. Later, Jöreskog and Sörbom (1982), Keesling (1973), and Loh (2011) developed a general model of structural equations, known as the LISREL model (Linear Structural Relations). Jöreskog extended Exploratory Factor Analysis (EFA) to Confirmatory Factor Analysis (CFA), generating the concept of Structural Equation Model (SEM).

The Structural Equation Model is a form of graphical modelling in which a system of relationships between variables or multivariate latent factors is established (structural component). These relationships are translated into a system of statistical equations that fit simultaneously. Latent factors represent abstract concepts that can be observed indirectly through their effects on observed indicators or variables (CFA component).

Furthermore, the combination of SEM modelling capabilities with the broader Generalized Linear Models (GLM) estimation framework, has been put together in a framework called generalized structural equation models (GSEM), which allow models to be built that include latent variables and a wide variety of response variables that are continuous, binary, ordinal or multinomial. Models are, for example, linear regression, gamma regression, logit, probit, ordinal logit, ordinal probit, Poisson, negative binomial and multinomial logit.

Once the model is calibrated, it can be used to predict a food loss estimate for the subsequent farm surveys without a food loss module. It might be necessary to update the recalibration of the model after the food loss data are collected in a subsequent farm survey.

3.2.2 Datasets used to test the modelling approach for Malawi

For this approach, the Living Standard Measurement Study – Integrated Survey on Agriculture (LSMS-ISA) from Malawi and Nigeria were used because the two consecutive survey rounds with food loss data for these two surveys were available. The LSMS-ISA datasets used for Malawi were the Integrated Household Survey 2016/17 (IHS4), and 2018/19 (IHS5). For Nigeria the General Household Survey 2015/16 (GHS 2015/16) and 2018/19 (GHS 2018/19) were used. Their variables and general descriptive statistics are presented in table 13.

Table 12: Descriptive statistics of the variables relevant for the model (LSMS-ISA) Malawi Integrated Household Survey 2016/17 (IHS4), and 2018/19 (IHS5)

Country	Malawi IHS 4 2016/17			Malawi IHS 5 2018/19		
Variable	N	Mean	Std Dev	N	Mean	Std Dev
Loss percentage (post-harvest)	7143	3.84	13.48	7465	3.09	10.50
Crop production (Kg)	8009	375.73	371.19	7926	372.42	377.67
Household head – Age	8006	44.81	16.39	7926	45.33	16.35
Harvest length (days in average)	7999	17.95	15.50	7908	21.95	19.79
Area planted (ha)	8009	37.19	24.87	7926	36.27	25.13
Plot distance to household (Km)	7372	1.14	5.38	7446	1.23	4.65
Plot slope	7355	5.28	5.53	7439	5.23	5.36
Plot elevation	7355	864.54	324.00	7439	844.17	314.93
HH distance to market (Km)	8008	23.33	14.55	7918	23.18	13.92
Variable	N	%		N	%	
Household head – Gender	8006			7926		

	Female	2455	30.66%	2547	32.13%
	Male	5551	69.34%	5379	67.87%
Household head – Literacy		8009		7962	
	Yes	720	8.99%	1995	25.06%
	No	7289	91.01%	5967	74.94%
Improved seed		8009		7962	
	Yes	3786	47.27%	3151	39.58%
	No	4223	52.73%	4881	60.42%
Agro-ecological zones		8008		7954	
	Tropic-warm/semiarid	3764	47.00%	3572	44.91%
	Tropic-warm/subhumid	2828	35.31%	2813	35.37%
	Tropic-cool/semiarid	792	9.89%	884	11.11%
	Tropic-cool/subhumid	624	7.79%	685	8.61%
Region Malawi		8006		7926	
	North	1360	16.99%	1241	15.66%
	Central	2686	33.55%	2767	34.91%
	Southern	3960	49.46%	3918	49.43%

3.2.3. Results from applying the modelling approach to Malawi

In this section, the results obtained for the Generalized Structural Equation Models and its model-based estimates in Malawi are presented. The results for Nigeria are available in the in the annex.

To build a generalized structural equation model, the first step entails establishing the theoretical relationship between the different sets of variables and post-harvest losses. Some of the factors influencing the level of food losses might be considered structural, as the size of the farm, the type of technology and practices used. For these, it is expected that they tend to produce similar levels of loss percentages in the given environment. Nevertheless, they are influenced by other factors that can affect the level of losses in the short-term. Here, the most relevant are weather conditions, the level of pest incidence, or market related fluctuations (less relevant if most of the crop production is for own consumption). Between these variables, and the variable of total production and post-harvest losses, multiple effects can exist. For the specification of GSEM, these variables are assessed and estimated.

As a first step, the **Exploratory Factor Analysis (EFA)** was used to identify relevant environmental factors among all observed environmental variables in the LSMS-ISA survey data. The overarching goal of EFA is to identify the underlying relationships between measured variables and identify a set of "unobserved" variables called factors. Usually, it is used on a large number of measured variables, which are assumed to be related to a smaller number of factors. In the context of this application of environmental variables,

EFA can explain environmental changes using a certain number of factors in the set of independent variables, which are possible to observe from one survey round to another. For instance, differences in rainfall or temperatures from one survey to another, when it is assumed that they affect the level of food losses. This technique uses the structure of the correlation between all observed variables to extract the maximum of the common variance. The variables that achieve the maximum common variance are grouped together and these groups are referred to as factors. They represent latent variables, which are not measured directly, but are unobserved underlying drivers. These factors are estimated and then used as the explanatory environmental variables in the structural path diagramme of the proposed structural equation model (see annex 3 for details of the EFA procedures used).

In these routines, EFA was applied to identify environmental factors that seemed to be relevant for the level of post-harvest losses. To do this, the Principal Factor Method (PFM) was used to analyse the correlation matrix. PFM seeks to find the fewest principal factors that can account for the common variance (correlation) of a set of variables. To select the number of factors that seem to be relevant, the so called “Kaiser criterion” is used, which selects the factors with an eigenvalue greater than one. Additionally, the cumulative proportion of the variance explained by the factors is used as complimentary criterion for selecting the factors, these are expressed as percentage, with the proportion showing the factor´s contribution to explaining the variance. To choose the factors to be considered, the criterion establishes a cutoff point at more than 80 percent of the cumulative proportion of the variance.

For example, for environmental variables recorded for Malawi LSMS-ISA in the two Integrated Household Surveys, the Exploratory Factor Analysis produced the following first output, which is shown in Figure 3:

Figure 3: Exploratory Factor Analysis Malawi LSMS-ISA

Factor analysis/correlation — Number of obs = 15,910
Method: principal factors — Retained factors = 5
Rotation: (unrotated) — Number of params = 70

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	6.11888	3.27067	0.4038	0.4038
Factor2	2.84821	0.34598	0.1880	0.5918
Factor3	2.50224	0.65122	0.1651	0.7569
Factor4	1.85101	0.84050	0.1222	0.8791
Factor5	1.01052	0.21875	0.0667	0.9458
Factor6	0.79177	0.67673	0.0523	0.9980
Factor7	0.11504	0.04377	0.0076	1.0056
Factor8	0.07127	0.06653	0.0047	1.0103
Factor9	0.00474	0.00917	0.0003	1.0106
Factor10	-0.00443	0.01011	-0.0003	1.0103
Factor11	-0.01455	0.00224	-0.0010	1.0094
Factor12	-0.01678	0.00633	-0.0011	1.0083
Factor13	-0.02311	0.00622	-0.0015	1.0067
Factor14	-0.02934	0.00510	-0.0019	1.0048
Factor15	-0.03444	0.00395	-0.0023	1.0025
Factor16	-0.03838	.	-0.0025	1.0000

For this output, the Kaiser criterion indicates many factors from which five are selected. As shown in Table 14, factors 1 to 5 have an eigenvalue that exceeds one. These five factors show a cumulative percentage of 94.58 percent of the total variance of the 16 variables included from the LSMS-ISA survey.

To make the results obtained from EFA more reliable, and to better understand this output, the Varimax (Kaiser, 1958) rotation method was used. The rotation method rotates the variables identified in the five factors between the factors, resulting in a percentage of variance explained by each factor instead of the eigenvalue, while the cumulative proportion of the explained variance remains the same. The rotated eigenvalues, representing the variance for each factor, are presented in figure 4:

Figure 4 Exploratory Factor Analysis after Kaiser Criterium to select main factors

```
Factor analysis/correlation                    Number of obs    =     15,910
    Method: principal factors                  Retained factors =          5
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         70
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	5.68709	2.87520	0.3753	0.3753
Factor2	2.81189	0.76855	0.1856	0.5609
Factor3	2.04335	0.06276	0.1349	0.6957
Factor4	1.98058	0.17263	0.1307	0.8265
Factor5	1.80795	.	0.1193	0.9458

There is a clear, though small change in the proportion of the variance for each of the five factors when applying the Varimax rotation method. For example, the proportion for factor 1 in the initial output was 0.4038, and in the rotated output result, it is 0.3753; for factor 5, the initial output showed a proportion of variance of 0.0667, in the rotated output result is 0.1193. The cumulative variance for all five selected factor remains at 94.58 percent%.

For the interpretation of common factors, it is relevant to outline the variables behind the five factors, which are singled out in EFA. To do this, the rotated factor loadings are analysed and represent measures of the association between the factor and observed variables. Rotated factor loadings for this example are shown in Table 14.

Table 13: Rotated factor loadings

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
hhgv_af_bio_1	0.1581	0.068	-0.0422	0.9716	0.0212	0.0242
hhgv_af_bio_8	0.1044	0.0368	-0.1227	0.9744	-0.0439	0.0213
hhgv_af_bio_12	0.0104	0.9053	0.0939	0.033	0.1772	0.1389
hhgv_af_bio_13	0.1628	0.9181	0.07	0.0598	0.0835	0.1151
hhgv_af_bio_16	0.1047	0.9598	0.0294	0.0436	0.0799	0.0586
hhgv_sq1	0.8473	-0.0991	0.192	0.0125	0.1331	0.2175
hhgv_sq2	0.9258	-0.0295	0.1371	0.0705	0.0828	0.1113
hhgv_sq3	0.8178	0.1666	-0.0611	-0.0316	0.004	0.2987
hhgv_sq4	0.9386	0.0646	0.0375	0.1491	-0.0012	0.0913
hhgv_sq5	0.964	0.1107	0.0392	0.1416	0.0172	0.0365
hhgv_sq6	0.9727	0.1109	0.0357	0.1164	0.0132	0.0266

hhgv_sq7	0.7683	0.3039	-0.1075	0.0018	-0.0181	0.3055
hhgv_h_in_tot	0.0824	0.1852	-0.0304	-0.0355	0.9459	0.0619
hhgv_h_in_wetQ	0.0317	0.1108	-0.4013	0.0113	0.8863	0.0399
hhgv_h_end_tot	0.08	0.1111	0.947	-0.0758	-0.1171	0.0651
hhgv_end_wetQ	0.094	0.06	0.937	-0.1069	-0.2032	0.0568

Rotated loadings that are as close to 1 or −1 as possible represent a high correlation between the observed variables and the latent factor. Low loadings are those that are as close to 0 as possible. The variable called "uniqueness" in Table 14, represents the proportion of variance of each observed variable not represented in the selected common factors. This proportion should be small, ideally under 0.4 or 0.5. If not, it indicates that the variable does not contribute to the explanation of the common factors and could be excluded from the analysis.

For the interpretation of the latent factors, at loadings over 0.7 or below -0.7 are shaded blue in Table 14. Factor 1 includes seven observed variables:

- hhgv_sq1 = Nutrient availability
- hhgv_sq2 = Nutrient retention capacity
- hhgv_sq3 = Rooting conditions
- hhgv_sq4 = Oxygen availability to roots
- hhgv_sq5 = Excess salts
- hhgv_sq6 = Toxicity
- hhgv_sq7 = Workability (constraining field management)

These variables are descriptors of the soil´s physical and chemical characteristics, so factor 1 is referred to as "soil patterns". Factor 2 includes three observed variables:

- hhgv_af_bio_12 = annual precipitation (mm)
- hhgv_af_bio_13 = precipitation of wettest month (mm)
- hhgv_af_bio_16 = precipitation of wettest quarter

These variables are precipitation indicators, so factor 2 is called "precipitation". Factor 3 includes two variables:

- hhgv_end_wetQ = total rainfall in wettest quarter (mm) within 12-month periods starting July 2015 (ihs4) and 2018 (ihs5).
- hhgv_h_end_wetqstart = start of wettest quarter in dekads 1-36, where the first week of July 2015 (ihs4) and 2018 (ihs5)=1.

These variables represent the start of the wettest quarter and the total rainfall in the second year of each survey time span, so factor 3 is called "Rainfall ending year". Factor 4 considered two variables:

- hhgv_af_bio_1 = Annual Mean Temperature (degC * 10)
- hhgv_af_bio_8 = Mean Temperature of Wettest Quarter (degC * 10)

These are annual temperature variables, so factor 4 is called "temperature". Finally, factor 5 includes the last two variables:

- hhgv_h_in_wetQ = total rainfall in wettest quarter (mm) within 12-month period starting July 2014 (IHS4) and 2017 (IHS5).
- hhgv_h_in_wetQstart = start of wettest quarter in dekads 1-36, where first week of July 2014 (IHS4) and 2017 (IHS5) =1.

These variables represent the start of the wettest quarter and the total rainfall in the first year of each survey time span, so factor 5 is called “rainfall starting year”.

Estimation of GSEM: For the next step, the construction of a path diagramme is presented. A graphical representation of the structural relationships between the variables based on a theoretical framework is presented in Figure 5. The path diagramme is developed to represent the linear relationships between the determinants of crop production volumes and food loss percentages, using graphical devices.[10] For this example, a graphical representation of the proposed path diagramme for the relationships between determinants of crop production and food losses are explained below.

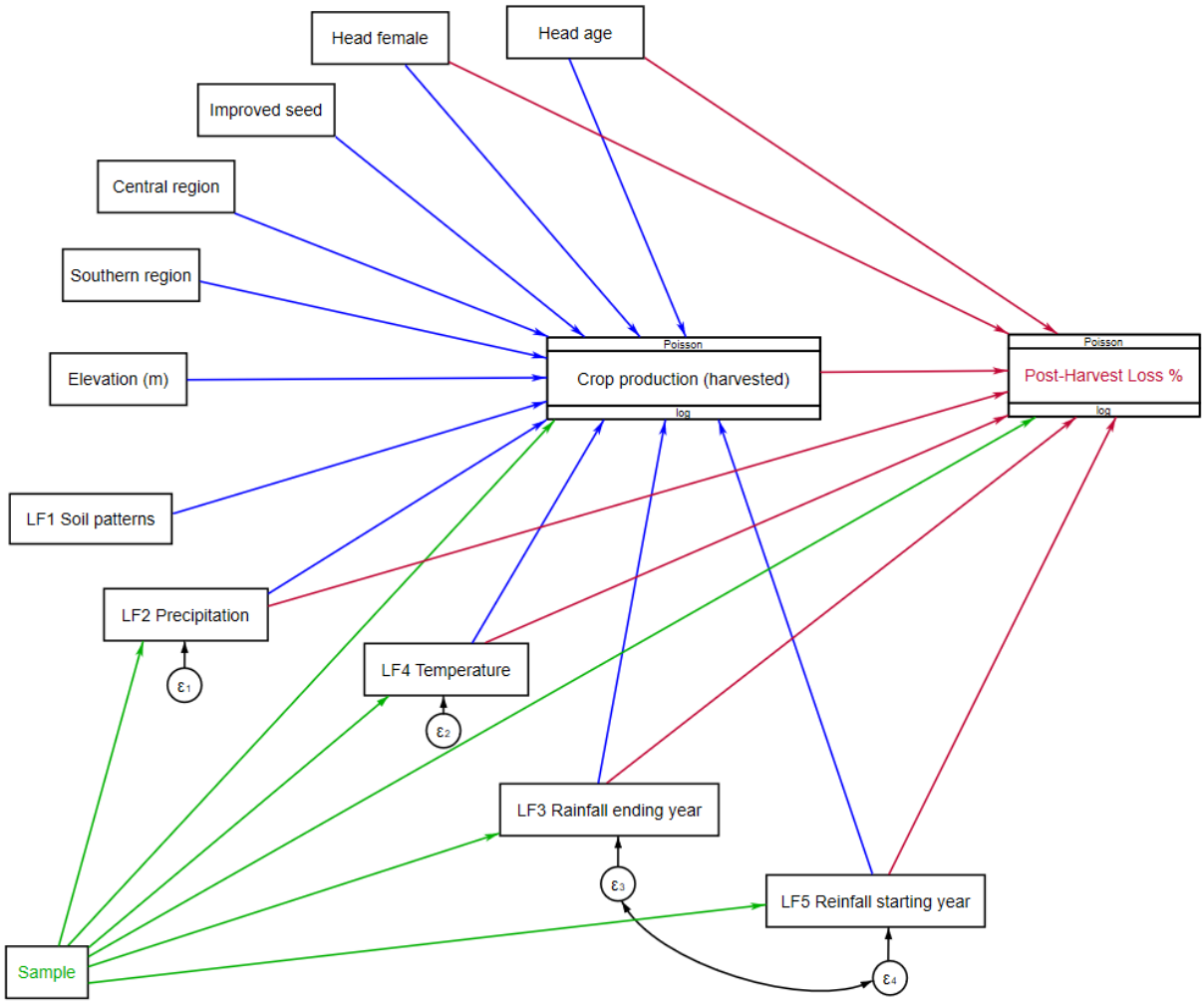
In this representation, the response variables are crop production volumes and post-harvest loss percentages. Notice that both variables are modelled using a Poisson regression (log link function and the Poisson distributional family).

In the bottom left of the diagramme, there is a box named “sample”, which is a dummy variable and represents changes between survey rounds (year-to-year). It has linear effects on four of the five environmental latent factors (LF2 to LF5) and a direct effect on the log post-harvest losses in percent (green arrows). Environmental latent factors receive an arrow representing the regression response, each with their specific error term ε_i. At the top of the diagramme, there are two social variables, age of the head of the household and a dummy variable indicating if their gender is female.

With blue arrows, there is a set of additional determinants in the medium upper left of Figure 5. These include the use of improved seed, the region of the country, the field elevation and the latent variable related to soil conditions. The blue arrows represent the linear contribution of the determinants on the log crop production, and the red arrows represent linear contributions on the log post-harvest losses in percent.

[10] The statistical programme STATA has a built-in device called “SEM Builder” under which path diagrammes for GSEMs can be created and analysed.

Figure 5: Graphical representation of the structural relationships between the variables



Fitting process: The second step is the fitting process using sampling weights. This can be used to correct bias and adjust standard errors. The coefficient estimates for the full model, are shown in Table 5.

Table 14: Coefficient estimates for the latent factors in the GSEM

Equation	Coef.	Robust Std. Err.	z	P	LB 95%	UB 95%
Crop production						
LF1 Soil patterns	-0.010	0.011	-0.90	0.369	-0.032	0.012
LF2 Precipitation	0.051	0.012	4.26	0.000	0.028	0.075
LF3 Rainfall ending year	-0.029	0.024	-1.22	0.221	-0.076	0.018
LF4 Temperature	-0.265	0.038	-6.95	0.000	-0.340	-0.190
LF5 Rainfall starting year	0.017	0.013	1.38	0.167	-0.007	0.042
age of household	-0.002	0.001	-4.29	0.000	-0.004	-0.001
Household female (yes)	-0.089	0.021	-4.31	0.000	-0.129	-0.048
improved seed (yes)	0.149	0.019	7.85	0.000	0.112	0.187
Elevation (m)	-0.0003	0.0001	-2.57	0.010	-0.001	0.000
Region central	0.032	0.040	0.80	0.425	-0.047	0.112
Region south	-0.288	0.047	-6.12	0.000	-0.380	-0.196
sample (year)	0.077	0.039	1.97	0.049	0.000	0.153
_cons	2.699	0.125	21.55	0.000	2.453	2.944
ln(area planted)	1	(exposure)				
Post-harvest loss in percent						
Crop production	-0.001	0.000	-8.17	0.000	-0.001	-0.001
LF2 Precipitation	-0.046	0.037	-1.25	0.211	-0.118	0.026
LF3 Rainfall ending year	-0.116	0.051	-2.25	0.024	-0.217	-0.015
LF4 Temperature	0.276	0.039	7.14	0.000	0.201	0.352
LF5 Rainfall starting year	-0.142	0.047	-3.00	0.003	-0.235	-0.049
age of household	-0.001	0.002	-0.32	0.749	-0.005	0.004
Household female (yes)	-0.056	0.080	-0.70	0.482	-0.213	0.100
sample (year)	-0.423	0.121	-3.50	0.000	-0.660	-0.186
_cons	1.799	0.118	15.31	0.000	1.569	2.030
LF2 Precipitation						
sample (year)	0.087	0.017	5.25	0.000	0.055	0.120
_cons	-0.108	0.011	-9.44	0.000	-0.130	-0.085
LF3 Rainfall ending year						
sample (year)	1.486	0.009	163.54	0.000	1.468	1.504
_cons	-0.860	0.007	-124.62	0.000	-0.874	-0.847
LF4 Temperature						
sample (year)	-0.032	0.019	-1.67	0.095	-0.069	0.006
_cons	-0.065	0.014	-4.51	0.000	-0.093	-0.037
LF5 Rainfall starting year						
Sample	-0.724	0.015	-48.68	0.000	-0.753	-0.695
_cons	0.215	0.009	23.19	0.000	0.197	0.234
var(e.f2)	0.868	0.013			0.844	0.893
var(e.f3)	0.308	0.005			0.299	0.317
var(e.f4)	0.963	0.011			0.941	0.985
var(e.f5)	0.658	0.011			0.636	0.680
cov(e.f3,e.f5)	0.237	0.005	45.82	0.000	0.227	0.247

Each equation in this GSEM can be identified and described, and their coefficients represent the direct effect between the independent variables and the response for each equation. In Table 15, the first

equation is a Poisson regression to model crop production, where the size of the planted area is used as an offset variable. Estimated coefficients show that precipitation had a positive effect and temperature had a negative effect on crop production, while female and older household heads had negative effects on crop production. The use of improved seeds, lower terrain elevation, and the southern region are related to higher crop production. A significant increase in crop production is observed for the second survey sample round (sample (year)).

The second equation is a Poisson regression to model post-harvest loss in percent in which a higher crop production is related to a lower percentage of post-harvest losses. Lower mean temperatures, shorter rainfalls for starting and ending years are related to a lower percentage of post-harvest losses. In the second survey (sampling round) there is a significant reduction in the percentage of post-harvest losses.

The third to sixth equation compare the weather conditions between both years, given that these are assumed to have an impact on the level of production and losses. The third equation shows an increase in precipitation between both years, the fourth shows a shorter rainfall period in the ending year for the second year, and the sixth equation a longer rainfall period in the starting year for the second year. The fifth equation does not show significant changes for temperature between both survey rounds. These last equations are linear regressions for the effect of survey rounds on environmental latent factors.

At the bottom of the coefficient estimation output are the estimates of residual variance for linear regression equations (the regression error terms) and the covariance between latent factors 3 and 5 (rainfalls at starting and ending years). These are represented as a double-headed arrow between error terms for both environmental latent factors.

Model reduction procedure: The third step is a model reduction procedure during which a step-by-step manual process is used to delete useless terms in the model without the loss of predictability efficiency for estimation purposes. In each step, the least significant term (higher p values) is carefully eliminated, checking to avoid changes in the remaining variable relationships. In each step, the path arrow with the least significant coefficient is eliminated in the graphical diagramme.

The next step is the analysis of the fitted coefficient for the step-in course. If no significant changes of the coefficients and their level of significance are evidenced, the process continues to eliminate the least significant arrow. When a variable in the model has no arrows representing relationships with other variables, it is directly deleted from the model. The deletion criteria used is based on significant values being over 0.1.

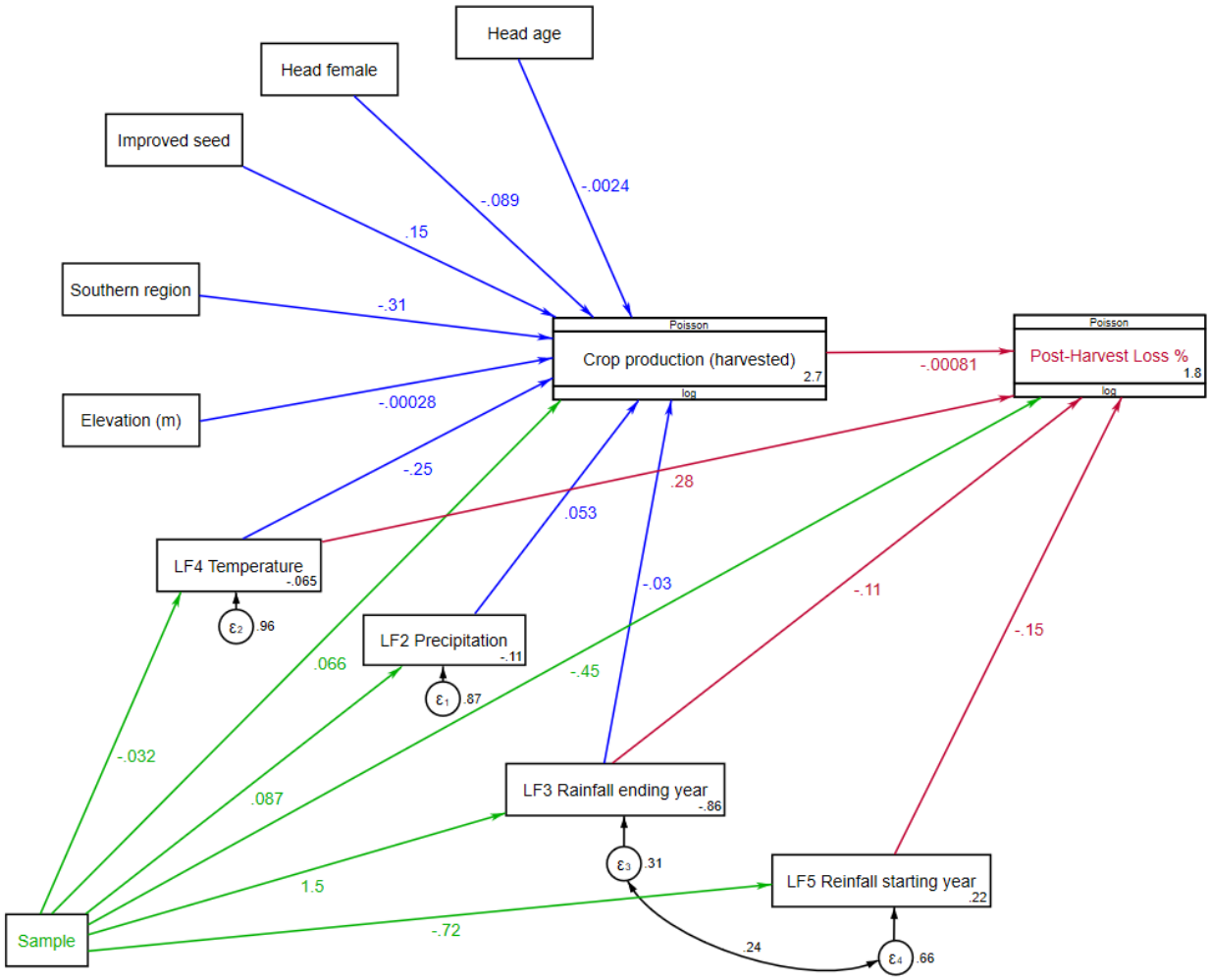
After this procedure, the first model assigned 12 paths to the crop production equation, while the final model contains only nine paths, i.e. process LF1 “soil patterns” was eliminated. The percentage of post-harvest losses had eight paths in the first model and ended with five paths after the reduction process. The rest of the model components remained unchanged. The coefficient estimates for the final reduced model are presented in Table 16.

Table 15: The coefficient estimates for the final reduced GSEM model

Equation	Coef.	Robust Std. Err.	z	P	LB 95%	UB 95%
Crop production						
LF2 Precipitation	0.053	0.012	4.43	0.000	0.030	0.076
LF3 Rainfall ending year	-0.030	0.018	-1.68	0.094	-0.064	0.005
LF4 Temperature	-0.252	0.035	-7.11	0.000	-0.321	-0.182
Age of household	-0.002	0.001	-4.30	0.000	-0.004	-0.001
Household female (yes)	-0.089	0.021	-4.32	0.000	-0.130	-0.049
improved seed (yes)	0.150	0.019	7.88	0.000	0.112	0.187
Elevation (m)	-0.0003	0.0001	-2.45	0.014	-0.0005	-0.0001
Region south	-0.314	0.026	-12.16	0.000	-0.364	-0.263
Sample (year)	0.066	0.029	2.26	0.024	0.009	0.124
_cons	2.689	0.111	24.16	0.000	2.471	2.907
ln(area planted)	1.000	(exposure)				
Post-harvest loss in percent						
Crop production	-0.001	0.000	-8.39	0.000	-0.001	-0.001
LF3 Rainfall ending year	-0.106	0.050	-2.12	0.034	-0.204	-0.008
LF4 Temperature	0.282	0.039	7.23	0.000	0.205	0.358
LF5 Rainfall starting year	-0.151	0.047	-3.23	0.001	-0.243	-0.059
sample (year)	-0.452	0.122	-3.70	0.000	-0.691	-0.212
_cons	1.753	0.081	21.59	0.000	1.594	1.912
LF2 Precipitation						
sample (year)	0.087	0.017	5.25	0.000	0.055	0.120
_cons	-0.108	0.011	-9.44	0.000	-0.130	-0.085
LF3 Rainfall ending year						
sample (year)	1.486	0.009	163.54	0.000	1.468	1.504
_cons	-0.860	0.007	-124.62	0.000	-0.874	-0.847
LF4 Temperature						
sample (year)	-0.032	0.019	-1.67	0.095	-0.069	0.006
_cons	-0.065	0.014	-4.51	0.000	-0.093	-0.037
LF5 Rainfall starting year						
sample (year)	-0.724	0.015	-48.68	0.000	-0.753	-0.695
_cons	0.215	0.009	23.19	0.000	0.197	0.234
var(e.f2)	0.868	0.013			0.844	0.893
var(e.f4)	0.308	0.005			0.299	0.317
var(e.f3)	0.963	0.011			0.941	0.985
var(e.f5)	0.658	0.011			0.636	0.680
cov(e.f3,e.f5)	0.237	0.005	45.82	0.000	0.227	0.247

Final path diagramme for the reduced model: This reduced model has the same interpretation of estimated direct effects on response variables obtained from the full model. The estimated coefficients for each direct path appear explicitly in Figure 6 and correspond to those obtained in the model output in Table 16. The fitted path diagramme includes coefficients for the direct effects.

Figure 6: Final path diagramme for the reduced model



Mediator variable: Furthermore, it is possible to evaluate indirect effects of some independent (x) variables on the dependent variable (y) via a mediator variable (m). In this example, the significant positive change of the crop production has a direct effect of the difference between the first and the second survey round ("sample" in Figure 6). In the path, the crop production is conditionally independent from the environmental latent factors, but there are three other paths through which the sample contributes to changes in crop production through environmental factors. For example, an indirect path from "sample" to the level of crop production is represented in the path diagramme by the connection between "sample" to "LF2 Precipitation" (green arrow), and then from "LF2 Precipitation to "crop production" (blue arrow), where "LF2 Precipitation" acts as a mediator variable. The indirect effect is the multiplication of the coefficients of the arrows in the path, with the green coefficient showing a value of 0.087 and blue coefficient of 0.053, so the indirect effect is $0.087 * 0.053 = 0.0046$. This is summarized in Figure 7.

Figure 7: "LF2 Precipitation" indirect effect on post-harvest loss through production

```
. nlcom _b[imp_production:f2]*_b[f2:sample]

       _nl_1:  _b[imp_production:f2]*_b[f2:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.0046268	.0013709	3.38	0.001	.00194	.0073137

It shows that there is a significant positive indirect effect of the survey round “sample” on “crop production”, mediated by LF2 Precipitation. Another estimated indirect effect of “sample” on “crop production” is mediated by “LF3 Rainfall ending year”, as shown in Figure 8.

Figure 8: "LF3 Rainfall ending year" indirect effect on post-harvest loss through production

```
. nlcom _b[imp_production:f3]*_b[f3:sample]

       _nl_1:  _b[imp_production:f3]*_b[f3:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.0440934	.0263038	-1.68	0.094	-.0956479	.0074611

In this case the indirect effect is negative but not significant. The estimated indirect effect of the “sample” on “crop production” mediated by LF4 Temperature is shown in Figure 9.

Figure 9 "LF2 Temperature" Indirect effect on PHL loss through production

```
. nlcom _b[imp_production:f4]*_b[f4:sample]

       _nl_1:  _b[imp_production:f4]*_b[f4:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.0079308	.0048889	1.62	0.105	-.0016512	.0175128

This indirect effect is positive, but also is not significant. To estimate the total effect of survey the round “sample” on "crop production”, the estimated overall effect is the sum of direct and indirect effects. This is estimated in Figure 10.

Figure 10: "Sample" indirect effect on post-harvest loss through production

```
. nlcom _b[imp_production:sample]+_b[imp_production:f2]*_b[f2:sample]+ _b[imp_
> production:f3]*_b[f3:sample]+ _b[imp_production:f4]*_b[f4:sample]

       _nl_1:  _b[imp_production:sample]+_b[imp_production:f2]*_b[f2:sample]+
> _b[imp_production:f3]*_b[f3:sample]+ _b[imp_production:f4]*_b[f4:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.0346812	.0198043	1.75	0.080	-.0041345	.0734969

This means that the total effect of the survey round "sample" on "crop production" tends to be positive but is not significant.

The significant indirect effects of survey round "sample" (green arrows) on post-harvest losses (red arrows), mediated by environmental factors are shown as followed. A negative significant effect mediated by "LF3 Rainfall ending year":

Figure 11:"Sample" indirect effect on post-harvest loss through LF3 Rainfall ending year

```
. nlcom _b[imp_per_phl:f3]*_b[f3:sample]

       _nl_1:  _b[imp_per_phl:f3]*_b[f3:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.1577304	.0742067	-2.13	0.034	-.3031728	-.012288

A positive significant effect mediated by "LF5 Rainfall starting year":

Figure 12 "Sample" indirect effect on post-harvest loss through LF5 Rainfall starting year

```
. nlcom _b[imp_per_phl:f5]*_b[f5:sample]

       _nl_1:  _b[imp_per_phl:f5]*_b[f5:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.1094681	.0340261	3.22	0.001	.0427781	.1761581

In addition, it is possible to estimate the indirect effects including two mediators (three connecting arrows), for example the effect of the survey round "sample" via environmental factors (green arrows) and crop production (blue arrows), and by connecting on post-harvest losses (red arrow). In the example, an estimated negative effect of "sample" on "post-harvest losses" via "LF2 Precipitation", via "crop production" is shown in Figure 13.

Figure 13: "Sample" indirect effect on PHL through LF2 Precipitation

```
. nlcom _b[imp_per_phl:imp_production]*_b[imp_production:f2]*_b[f2:sample]

       _nl_1:  _b[imp_per_phl:imp_production]*_b[imp_production:f2]*_b[f2:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-3.73e-06	1.20e-06	-3.12	0.002	-6.08e-06	-1.38e-06

This estimated coefficient is very close to zero and is, therefore, possibly not important for prediction even though it is shown to be significant.

The total effect of survey sample round on post-harvest losses summarized in Figure 14.

Figure 14: Total effect of survey sample round on post-harvest losses

```
. nlcom _b[imp_per_phl:sample]+_b[imp_per_phl:f3]*_b[f3:sample]+ ///
>  _b[imp_per_phl:f4]*_b[f4:sample]+ ///
>  _b[imp_per_phl:f5]*_b[f5:sample]+ ///
>  _b[imp_per_phl:imp_production]*_b[imp_production:f2]*_b[f2:sample]+ ///
>  _b[imp_per_phl:imp_production]*_b[imp_production:f3]*_b[f3:sample]+ ///
>  _b[imp_per_phl:imp_production]*_b[imp_production:f4]*_b[f4:sample]

       _nl_1:  _b[imp_per_phl:sample]+_b[imp_per_phl:f3]*_b[f3:sample]+ _b[imp_
> per_phl:f4]*_b[f4:sample]+ _b[imp_per_phl:f5]*_b[f5:sample]+ _b[imp_per_phl:i
> mp_production]*_b[imp_production:f2]*_b[f2:sample]+ _b[imp_per_phl:imp_produc
> tion]*_b[imp_production:f3]*_b[f3:sample]+ _b[imp_per_phl:imp_production]*_b[
> imp_production:f4]*_b[f4:sample]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.5086462	.0693084	-7.34	0.000	-.6444882	-.3728042

The total effect of time elapsed between the two survey rounds is a significant reduction in post-harvest losses, mainly through the changes in the weather conditions between both production cycles, and to a minor degree, through the direct impact of the levels of production.

Evaluation of the model obtained for Malawi: Following the establishment of the factor and mediator variables, GSEM is used to estimate post-harvest loss percentages. As a result, model-based estimates can be generated for following survey rounds for which food losses are not being collected. The model was, therefore, built on two possibly distant survey rounds.

The post-harvest loss model-based estimate for the whole period of the two included survey rounds obtained is that 3.85 percent of maize crop production being lost in the post-harvest period.

Figure 15: Post-harvest loss model-based estimate for the whole period of the two included survey rounds

```
Predictive margins                                Number of obs     =     15,720
Model VCE    : Robust

Expression   : Predicted mean (Post-Harvest Loss (over production) %
               (imputed)), predict(outcome(imp_per_phl))
```

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
_cons	3.846266	.1331041	3.585386	4.107145

The estimates of the post-harvest loss for each survey round are 4.86 percent in the first survey round, and 3 percent in the second survey round (both are model-based estimates). Their characteristics are summarized in Figure 16.

Figure 16: Model-based estimates of post-harvest loss for each survey round

```
Predictive margins                                Number of obs     =     15,720
Model VCE    : Robust

Expression   : Predicted mean (Post-Harvest Loss (over production) %
               (imputed)), predict(outcome(imp_per_phl))
over         : sample
```

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
sample				
sample 1	4.863093	.2478233	4.377369	5.348818
sample 2	3.008268	.1297711	2.753921	3.262615

Model specification tests: Finally, the linktest is used to validate the use of the GSEM model for estimation propose and its predictive power. Figure 17 present a summary of the output for the test.

Figure 17 Result of the linktest for the identified GSEM model (test model specifications)

```
Poisson regression                                Number of obs     =     14,589
                                                  Wald chi2(2)      =     242.28
Log pseudolikelihood =  -40316630                 Prob > chi2       =     0.0000
```

imp_per_phl	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	.2285528	.0414349	5.52	0.000	.1473419	.3097637
_hatsq	.0002866	.0035605	0.08	0.936	-.0066919	.0072651
_cons	.3550432	.110291	3.22	0.001	.1388768	.5712096

The model correctly passes this test; the squared prediction _hatsq has a non-significant coefficient and is almost zero, while the linearity in the prediction _hat is significantly positive. This indicates that the

model's specifications are sufficiently good and the model can be used to produce model-based estimations of post-harvest losses for the given LSMS-ISA survey in Malawi.

3.2.4 Conclusions on modelling losses from independent survey rounds

For this research GSEM models were used to produce post-harvest loss estimates at the farm level based on input variables from a national household or farm survey. The use of GSEM can be useful to understand the contribution of several determinant variables, including the environmental factors, and changes over time using two or more independent survey rounds on the percentage of post-harvest losses. In this regard, the model helps in evaluating total and partial contributions of determinants on losses, such as conditional independent effects and possible non-independent effects based on the covariance between some determinants.

Another very important use of GSEM, is that it gives an efficient and reliable procedure to estimate the percentage of post-harvest losses. In this regard, GSEM can distinguish between more structural factors influencing the overall level of losses and factors that can have a short-term immediate effect observed between survey rounds. The latter refers especially to weather factors that can have an immediate impact on the level of losses, with all other harvest and post-harvest practices being constant. This allows the model to factor in, first, a general level of losses given the structural characteristics of the farm, and second, the variables that drive the level of losses from one year to the other. Additionally, this model design makes it possible to distinguish between direct and indirect effects, whereby it can be considered that the level of losses is to some extent driven by the level of production as a mediator, and the indirect effect of variables that are key drivers of the level of production.

Based on the GSEM models tested in this research, the first relevant conclusion is that the variables collected in the given household surveys were sufficient to produce a model with predictive power and sufficiently good model specifications. One relevant input from the household survey came from the comprehensive set of weather variables provided in the data base. These were decisive for understanding the differences of the loss levels from one survey round to the next. On the other hand, the variables that are considered structural and thus influence the level of losses more in the medium term were covered by the following variables: age; gender; region; and type of seeds used. These variables are usually constant in the short term, even though they influence the overall level of losses; they only change in the medium to long term.

The variables that affect losses in the short term are mainly weather factors. Using weather factors, based on the set of available variables in the LSMS-ISA survey in Malawi, it was possible to aggregate a number of common factors that cumulatively represent 94.5 percent of the total variance of the weather variables and their differential influence in the food loss percentage for the two survey rounds. This supports the assumption that weather variables change the loss level from one year to the other. These variables can be obtained by capturing the GPS information of the farm and are therefore recommended to be included in farm and household surveys.

Although the models pass the specification tests and show an acceptable level of predictive power, they have not been tested for their ability to produce indirect food loss estimates for the survey rounds without the inclusion of the food loss module. To do this, the models require a longer series of survey rounds (preferably at least 2-3 survey rounds). In addition, further development is needed to generate a

prediction model that first needs to be calibrated, and then be connected to the survey variables that are to be used for doing the projection. These developments could be subject to further research.

The presented modelling approach can be of relevance to national statistics offices, especially when agricultural or household surveys are designed on rotating modules. Accordingly, the loss module might not be collected in each survey round. In this scenario, a calibrated prediction model could be a cost-effective option to generate loss estimates in-between the application of the loss module. The survey-based loss estimates that are collected, usually in a frequency of 3-4 years, show the medium-term structural trend of the level of losses (e.g., because of improvements of the production and post-harvest handling systems). The modelled-based estimates in-between the survey rounds most likely replicate the overall structural loss levels but they also reflect the loss variations that might be introduced due to weather conditions.

4. Discussion and recommendations

The research component to identify modelling approaches that can support the quality and cost-efficiency of harvest and post-harvest loss measurement in national farm and household surveys derived some insights and potential areas of use.

A first area of application is the use of modelling approaches to compensate for sub-sampling of food loss modules. Sub-sampling is one of the main recommendations of the 50x2030 agriculture survey design, as it enables countries to cover a diverse set of indicators without overloading the national farm and household surveys. The modelling approach used to improve the loss estimates of the survey with model-based estimates were overly positive. To a certain level, the tested farm and household surveys appear to cover sufficient variables to make it possible to explain parts of the variance and improve its quality in model-based estimates. The used modelling approach and its application requires less specialized capacities in the National Office of Statistics and can be adapted by a team of statisticians that are familiar with regression analysis. Sub-sampling, additionally, offers the possibility of collecting a more detailed loss questionnaire, which can improve the response rates and the accuracy of loss declarations. Though, some recommendations need to be considered for the questionnaire design regarding some of the variables that appear to support the model specification and estimation power.

On the other hand, the models to predict losses in between survey rounds that collect the loss module is still subject to further research. This exercise has made it possible to identify the first elements of the modelling approach, building on a structural equation model that can consider the structural drivers of losses and those with a short-term year-to-year impact. The changes in losses from one year to another only rely on weather variables, while other relevant short-term drivers of losses might be missing, such as pest infestation, diseases, overall supply and market situation. To incorporate them, corresponding variables are needed in the farm and household survey or from other external data sources. The prediction power of these models still needs to be assessed in more detail and built based on the variables that are collected in every survey round. The use of this modelling approach requires more specialized experience in statistical modelling, which might not be available in national statistics offices. Collaboration with more specialized research centres or University departments might, therefore, is needed,

complementary to broader capacity development of statistical modelling in the National Statistics Offices. Nevertheless, this modelling approach has the potential to reduce data collection costs and can be used to replace rotating food loss modules in integrated agricultural and household surveys.

References

50x2030 Initiative & FAO (2021a) A Guide to Sampling. Technical Note Series. 50x2030 Sampling Strategy_14 June COVER ACK_SG_tracked.pdf

50x2030 Initiative (2021b). The 50x2030 Initiative: Bringing together committed partners to fill the agricultural data gap. Rome: World Bank.

Adisa, R.S., Adefalu, L.L., Olatinwo, L.K., Balogun, K.S. & Ogunmadeko, O.O. (2015). Determinants of post-harvest losses of yam among yam farmers in Ekiti State, Nigeria. *Bulletin of the Institute of Tropical Agriculture, Kyushu University*. 38 (1):73-78.

Affognon, H., Mutungi, C., Sanginga P. & Borgemeister, C. (2016). Unpacking Postharvest Losses in Sub-Saharan Africa: a meta-analysis. *World Development,*66, 49−68.

Agarwal, M., Agwaral, S, Ahmad, S., Singh, R. & Jahayari, K.M. (2021). Food loss and waste in India: the knowns and the unknowns. *World Resources Institute India*. www.wri.org/publication/food-loss-and-waste-in-india

Aidoo R., Rita, A. & Mensah, O.J. (2014). Determinants of postharvest losses in tomato production in the Offinso North District of Ghana. *Journal of Development and Agricultural Economics*. 6(8):338−344.

Alidu, A-F, Ali, E.B. and Aminu, H. (2016). Determinants of post-harvest losses among tomato farmers in the Navrongo Municipality in the upper east region. *Journal of Biology, Agriculture and Healthcare* 6(12): 14−20.

Ambler, K., Brauw, A. & Godlonton, S. (2018). Measuring postharvest losses at the farm level in Malawi. *Australian Journal of Agricultural and Resource Economics*. 62 (1): 139−160.

Amentae, T.K., Tura, E.G., Gebresenbet G. & Ljungberg D. (2016). Exploring value chain and post-harvest losses of Teff in Bacho and Dawo districts of central Ethiopia. *Journal of Stored Products and Post-harvest Research*. 7(1),11−28.

Ansah, I.G.K. & Tetteh, B.K. (2016). Determinants of yam postharvest management in the Zabzugu District of Northern Ghana. *Advances in Agriculture*.

Arun, G.C. and Ghimire, K. (2019). Estimating post-harvest loss at the farm level to enhance food security: A case of Nepal. *International Journal of Agriculture Environment and Food Sciences*, 3(3):127−136.

Babalola, D.A., Makinde, Y.O., Omonona, B.T., & Oyekanmi, M.O. (2010). Determinants of postharvest losses in tomato production: a case study of Imeko – Afon local government area of Ogun state. *Journal of Life and Physical Science*. 3(2): 14–18.

Basavaraja, H., Mahajanashetti, S.B. & Udagatti, N.C. (2007) Economic analysis of post-harvest losses in food grains in India: a case study of Karnataka. *Agricultural Economics Research Association (India).* 20(1): 176−126.

Begum, E.A., Hossain, M. & Papanagiotou, E. (2012). Economic analysis of post-harvest losses in food-grains for strengthening food security in northern regions of Bangladesh. *International Journal of Applied Research in Business Administration and Economics.* 1(3): 56−65.

Dharmathilake, N.R.D.S; Rosairo, H.S.R; Ayoni, V.D.N & Herath, R.M. (2020). Implications of post-harvest losses and acreage response of selected upcountry vegetables from Nuwara-Eliya District in Sri Lanka on sustained food security. *Journal of Agriculture Sciences,* 15(1)

Delgado, L., Schuster, M. & Torero, M. (2017). The reality of food losses: a new measurement methodology. IFPRI Discussion Paper 1686. Washington, DC, International Food Policy Research Institute.

Delgado, L., Schuster, M. & Torero, M. (2020). Quantity and quality food losses across the value chain: a comparative analysis. *Food Policy*. 98: 101958.

Falola, A., Salami, M.F., Bello, A.A. & Olaoye, T.A. (2017). Effect of yam storage techniques usage on farm income in Kwara State, Nigeria. *Agrosearch*,. 17(1):54–65.

Folayan, J.A., Babalola, J.A.& Ilesa, A. (2013). Determinants of post-harvest losses of maize in Akure North Local Government Area of Ondo State, Nigeria. *Journal of Sustainable Society.* 2(1).

FAO (2019). *The State of Food and Agriculture. Moving Forward on Food Loss and Waste Reduction*. Rome. http://www.fao.org/3/ca6030en/ca6030en.pdf

FAO (2020a). Guidelines on the measurement of harvest and post-harvest losses. Estimation of crop harvest and post-harvest losses in Malawi. Maize, rice and groundnuts. Field Test Report. Rome. www.fao.org/3/cb1562en/cb1562en.pdf

FAO (2020b). Guidelines on the measurement of harvest and post-harvest losses. Estimation of crop harvest and post-harvest losses in Zimbabwe. Field Test Report. Rome. https://www.fao.org/documents/card/en/c/CB1554EN/

Garikai, M. (2014). Assessment of vegetable postharvest losses among smallholder farmers in Umbumbulu area of KwaZulu-Natal province. University of KwaZulu-Natal, Pietermaritzburg, South Africa Master's Thesis,. http://hdl.handle.net/10413/11918

Global Strategy to improve Agricultural and Rural Statistics (GSARS). (2018a). Handbook on the Agricultural Integrated Survey (AGRIS). AGRIS Handbook on the Agricultural Integrated Survey (fao.org)

Global Strategy to improve Agricultural and Rural Statistics (GSARS) (2018b). Guidelines on the measurement of harvest and post-harvest losses. Rome. https://www.fao.org/3/ca6396en/ca6396en.pdf

Hengsdijk, H., & De Boer ,W.J..(2017). Post-harvest management and post-harvest losses of cereals in Ethiopia*. Food Security.* 9, 945–958.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23 (3): 187−200.

Hossain, M.A., & Miah, M.A.M. (2009). Post-harvest losses and technical efficiency of potato storage systems in Bangladesh. National Food Policy Capacity Strengthening Programme. Final Report CF # 2/08.

Huang, T., Li, B., Shen, D., Cao, J. & Mao, B. (2017). Analysis of the grain loss in harvest based on logistic regression. *Procedia Computer Science*, 122, 698−705.

Jha, S.N., Vishwakarma, R.K., Ahmad, T., Rai, A. & Dixit, A. . (2015). Report on Assessment of Quantitative Harvest and Post-Harvest Losses of Major Crops/Commodities in India. ICAR-All India Coordinated Research Project on Post-Harvest Technology, ICAR-CIPHET 130..

Johnson, L. K., Dunning, R. D., Bloom, J. D., Gunter, C. C., Boyette, M. D., & Creamer, N. G. (2018). Estimating on-farm food loss at the field level: A methodology and applied case study on a North Carolina farm. *Resources, Conservation and Recycling*, 137, 243−250.

Jöreskog, K.G., & Sörbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research*. 19(4): 404–416.

Kader, A. A. (2005.) Increasing food availability by reducing postharvest losses of fresh produce. *Acta Horticulture*. 682, 2169−2176.

Kaminski, J., & Christiansen, L. (2014). Post-harvest loss in sub-Saharan Africa − What do farmers say? *Global Food Security*. 3 (3-4): 149−158.

Keesling, W. (1972). Maximum likelihood approaches to causal flow analysis. University of Chicago. PhD thesis.

Khatun, M., Karim, M. R., Khandoker, S., Hossain, T. M., & Hossain, S. (2014). (2014). Post-harvest loss assessment of tomato in some selected areas of Bangladesh. *International Journal of Business, Social and Scientific Research*. 1(3), 209−218.

Kikulwe, E. M., Okurut, S., Ajambo, S., Nowakunda, K., Stoian, D. & Naziri, D. (2018). Postharvest losses and their determinants: A challenge to creating a sustainable cooking banana value chain in Uganda. *Sustainability*, 10(7): 2381.

Kimenju, S, & De Groote, H. (2010). Economic analysis of alternative maize storage technologies in Kenya. Paper presented at the Third Conference of the African Association of Agricultural Economists (AAAE). September 19-23;Cape Town, South Africa. http://econpapers.repec.org/scripts/search/search.asp?ft=simon+kimenju

Kitinoja, L., Tokala, V. Y. & Brondy, A. (2018). Challenges and opportunities for improved postharvest loss measurements in plant-based food crops. *Journal of Postharvest Technology*, 6(4): 16-34.

Kumar, D.K., Basavaraja, H. & Mahajanshetti, S.B. (2006). An economic analysis of post-harvest losses in vegetables

Morris, K.J., Kamarulzaman, N.H. & Morris, K.I. (2019). Small-scale postharvest practices among plantain farmers and traders: A potential for reducing losses in rivers state, Nigeria. *Scientific African*, 4, e00086.

Loh, W.Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1): 14−23.

Maziku, P. (2019). Determinants for post-harvest losses in maize production for small holder farmers in Tanzania. *African journal of applied research*, 5(1): 1−11.

Mebratie, M.A., Haji, J., Woldetsadik, K., Ayalew, A, & Ambo, E. (2015). Determinants of postharvest banana loss in the marketing chain of central Ethiopia. *Food Science and Quality Management*, 37, 52−63.

Meena, M.S., Kumar, A., Singh, K.M., & Meena, H.R. (2016). Farmers' attitude towards post-harvest issues of horticultural crops. *Indian Research Journal of Extension Education*, 9(3): 15−19.

Ndiritu, S.W. & Ruhinduka, R.D. (2019). Climate variability and post-harvest food loss abatement technologies: evidence from rural Tanzania. *Studies in Agricultural Economics*, 121(1): 30−40.

Ngowi, E.R. & Selejio, O. (2019). Post-harvest loss and adoption of improved post-harvest storage technologies by smallholder maize farmers in Tanzania. *African Journal of Economic Review*, 7(1), 249−267.

Paneru, R.B., Paudel, G. & Thapa, R.B. (2018). Determinants of post-harvest maize losses by pests in mid hills of Nepal. *International Journal of Agriculture, Environment and Bioresearch*, 3(1): 110−118.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 29(1): 15−24.

Qu, X., Kojima, D., Nishihara, Y., Wu, L. & Ando, M. (2020). Impact of rice harvest loss by mechanization or outsourcing: Comparison of specialized and part-time farmers. *Agricultural Economics/Zemědělská Ekonomika*, 66(12): 542−549.

Shee, A., Mayanja, S., Simba, E., Stathers, T., Bechoff, A. & Bennett, B. (2019). Determinants of postharvest losses along smallholder producers maize and Sweet potato value chains: an ordered Probit analysis. *Food Security*, 11, 1101−1120.

Smith, T.M.F. (1991). Post-stratification. *The Statistician*, 40(3): 315−23.

Tadesse, B., Bakala, F. & Mariam, L.W. (2018). Assessment of postharvest loss along potato value chain: the case of Sheka Zone, southwest Ethiopia. *Agriculture & Food Security*, 7, 1-14.

Wineman, A., Njagi, T., Anderson, C.L., Reynolds, T.W., Alia, D.Y., Wainaina, P, Ayieko, M.W. *et al*. (2020). A case of mistaken identity? Measuring rates of improved seed adoption in Tanzania using DNA fingerprinting. *Journal of Agricultural Economics*, 71(3): 719-741.

Wossen, T., Abdoulaye, T., Alene, A., Nguimkeu, P., Feleke, S., Rabbi, I. Y., Manyong, V. *et a*l. (2019). Estimating the productivity impacts of technology adoption in the presence of misclassification. *American Journal of Agricultural Economics*, 101(1): 1-16.

Xue, L., Liu, G., Parfitt, J., Liu, X., Van Herpen, E., Stenmarck, Å., O'Connor, C. *et al*. (2017). Missing food, missing data? A critical review of global food losses and food waste data. *Environmental Science & Technology*, 51(12): 6618-6633.

Annex I: Overview of the papers screened for the literature review on determining factors

Table 16: Overview of the papers screened for the literature review on determining factors

Country	Food group	Crops	Stages of the supply chain	Operations covered in the stages	Type of model	Full Reference
Papers that estimate food loss drivers on sub-national food loss survey data						
Bangladesh	Vegetable	Tomato	Farm, trade	Post-harvest		Khatun, M., Karim, M. R., Khandoker, S., Hossain, T. M., & Hossain, S. (2014). Postharvest loss assessment of tomato in some selected areas of Bangladesh. *International Journal of Business, Social and Scientific Research*, 1(3): 209–218.
Bangladesh	Vegetable	Tomato	Farm, trade	Post-harvest		Khatun, M. & Rahman, M. (2020). Postharvest loss assessment of tomato in selected locations of Bangladesh. *Bangladesh Journal of Agricultural Research*, 45, 43-52.
Bangladesh	Grains	Rice and wheat	On-farm	Harvest, post-harvest, storage, transportation		Begum, E.A., Hossain, M. & Papanagiotou, E. (2012). Economic analysis of post-harvest losses in food-grains for strengthening food security in northern regions of Bangladesh. *International Journal of Applied Research in Business Administration and Economics,*. 1(3): 56–65.
India	Grains	Rice and wheat	On-farm	Post-harvest		Basavaraja, H., Mahajanashetti, S.B. & Udagatti, N.C. (2007) Economic analysis of post-harvest losses in food grains in India: a case study of Karnataka. *Agricultural Economics Research Association (India),* 20(1): 176–126.
India	Roots and tubers	Potatoes and onions	On-farm, middleman	Post-havest		
Nepal	Various	Rice, wheat, maize, potato, mustard, cabbage and lentil	On-farm	Post-harvest		Arun, G.C. and Ghimire, K. (2019). Estimating post-harvest loss at the farm level to enhance food security: A case of Nepal. *International Journal of Agriculture Environment and Food Sciences*, 3(3): 127–136.
Nepal	Grains	Maize	On-farm	Post-harvest, storage		Paneru, R.B., Paudel, G. & Thapa, R.B. (2018). Determinants of post-harvest maize losses by pests in mid hills of Nepal. *International Journal of Agriculture, Environment and Bioresearch*, 3(1): 110–118.

Kenya	Grains	Maize	On-farm	Storage		De Groote, H., Kimenju, S. C., Likhayo, P., Kanampiu, F., Tefera, T., & Hellin, J. (2013). Effectiveness of hermetic systems in controlling maize storage pests in Kenya. *Journal of Stored Products Research*, 53: 27-36.
Ghana	Vegetables	Tomato	On-farm	Post-harvest		Aidoo R., Rita, A. & Mensah, O.J. (2014). Determinants of postharvest losses in tomato production in the Offinso North District of Ghana. *Journal of Development and Agricultural Economics,* 6(8): 338−344.
Ghana	Vegetables	Tomato	On-farm	Post-harvest		Alidu, A-F, Ali, E.B. and Aminu H. (2016). Determinants of post harvest losses among tomato farmers in the Navrongo Municipality in the upper east region. *Journal of Biology, Agriculture and Healthcare*, 6(12): 14−20.
Ghana	Vegetables	Tomato	On-farm, retailers, wholesalers	Post-harvest		Amoako-Adusei, Ruth (2019). Tomato postharvest losses in Ghana: an economic analysis. This report is taken from an MSc thesis conducted at the Wageningen University and Research, published under The HortiFresh Program
Uganda	Grains, Roots and tubers	Maize, sweet potato	On-farm, off-farm	Post-harvest		Shee, A., Mayanja, S., Simba, E., Stathers, T., Bechoff, A. & Bennett, B. (2019). Determinants of postharvest losses along smallholder producers maize and Sweetpotato value chains: an ordered Probit analysis., *Food Security*, 11, 1101−1120.
Uganda	Fruits	Banana	On-farm, retail	Post-harvest		Kikulwe, E. M., Okurut, S., Ajambo, S., Nowakunda, K., Stoian, D. & Naziri, D. (2018). Postharvest losses and their determinants: A challenge to creating a sustainable cooking banana value chain in Uganda. *Sustainability*, 10(7): 2381.
Ethopia	Fruits	Banana	On-farm, wholesale, retail	Post-harvest		Mebratie, M.A., Haji, J., Woldetsadik, K., Ayalew, A, & Ambo, E. (2015). Determinants of postharvest banana loss in the marketing chain of central Ethiopia. *Food Science and Quality Management*, 37, 52−63.
Ethiopia	Cereals	Teff	On-farm	Post-harvest		Amentae, T.K., Tura, E.G., Gebresenbet G. & Ljungberg D. (2016). Exploring value chain and post-harvest losses of Teff in Bacho and Dawo districts of central Ethiopia. *Journal of Stored Products and Post-harvest Research,* 7(1): 11−28.
Ethopia	Roots and tubers	Potato	On-farm, local trader,	Post-harvest		Tadesse, B., Bakala, F. & Mariam, L.W. (2018). Assessment of postharvest loss along potato value chain: the case of Sheka Zone, southwest Ethiopia. *Agriculture & Food Security*, 7, 1-14.

			wholesaler, retailer			
Ethopia	Roots and tubers	Onion	On-farm, traders	Storage		Emana, B., Afari-Sefa, V., Kebede, D., Nenguwo, N., Ayana, A., & Mohammed, H. (2017). Assessment of postharvest losses and marketing of onion in Ethiopia. *International Journal of Postharvest Technology and Innovation*, 5(4): 300–319.
United Republic of Tanzania	Grains	Maize	On-farm	Post-harvest		Maziku, P. (2019). Determinants for post-harvest losses in maize production for small holder farmers in Tanzania. *African Journal of Applied Research*, 5(1): 1–11.
Nigeria	Grains	Maize	On-farm	Post-harvest		Folayan, J.A., Babalola, J.A.& Ilesa, A. (2013). Determinants of post-harvest losses of maize in Akure North Local Government Area of Ondo State, Nigeria. *Journal of Sustainable Society,* 2(1).
Nigeria	Fruits		Fruit marketers	Post-harvest		Busari, A.O., Idris-Adeniyi, K.M., & Lawal, A.O. (2015). Food security and post-harvest losses in fruit marketing in Lagos metropolis, Nigeria. *Discourse Journal of Agriculture and Food Sciences*, 3(3): 52-58.
Nigeria	Roots and tubers	Yam	On-farm	Post-harvest		Adisa, R.S., Adefalu, L.L., Olatinwo, L. ., Balogun, K.S., & Ogunmadeko, O. O. (2015). Determinants of post-harvest losses of yam among yam farmers in Ekiti State, Nigeria, *Bulletin of the Institute of Tropical Agriculture, Kyushu University*, 38(1): 073-078.
Nigeria	Roots and tubers	Yam	On-farm	Storage		Falola, A., Salami, M.F., Bello, A.A. & Olaoye, T.A. (2017). Effect of yam storage techniques usage on farm income in Kwara State, Nigeria. *Agrosearch*, 17(1): 54–65.
Nigeria	Fruits	Plantain	On-farm, traders	Post-harvest		Morris, K.J., Kamarulzaman, N.H. & Morris, K.I. (2019). Small-scale postharvest practices among plantain farmers and traders: A potential for reducing losses in rivers state, Nigeria. *Scientific African*, 4, e00086.
Nigeria	Vegetables	Tomato	On-farm	Post-harvest		Babalola, D.A., Makinde, Y.O., Omonona, B.T., & Oyekanmi, M.O. (2010). Determinants of postharvest losses in tomato production: a case study of Imeko – Afon local government area of Ogun state. *Journal of Life and Physical Science,* 3(2): 14–18.
Nigeria	Vegetables	Tomato	On-farm	Post-harvest		Kuranen-Joko, D.N., & Liambee, D. H. (2017). Determinants of Postharvest Losses among Tomato Farmers in Gboko Local Governments Area of Benue State. *CARD International Journal of Agricultural Research and Food Production (IJARFP)*, 2(4): 27-33.

Nigeria	Vegetables	Tomato	On-farm	Post-harvest		Ayandiji, A., Adeniyi, O. D., & Omidiji, D. (2011). Determinant post harvest losses among tomato farmers in Imeko-Afon local government area of Ogun State, Nigeria. *Global Journal of Science Frontier Research*, 11(5): 23-27.
Malawi	Cereals	Maize, soya, groundnut	On-farm	Aggregated post-harvest and by activity		Ambler, K., Brauw, A. & Godlonton, S. (2018). Measuring postharvest losses at the farm level in Malawi. *Australian Journal of Agricultural and Resource Economics,* 62 (1): 139–160.
South Africa	Vegetables	Cabbage and spinach, tomato	On-farm	Post-harvest		Garikai, M. (2014). Assessment of vegetable postharvest losses among smallholder farmers in Umbumbulu area of KwaZulu-Natal province. University of KwaZulu-Natal, Pietermaritzburg, South Africa Master's Thesis. http://hdl.handle.net/10413/11918
Bangladesh	Roots and Tubers	Potatoes	Farm, Storage, Trade	Storage		Hossain, M.A., & Miah, M.A.M. (2009). Post-harvest losses and technical efficiency of potato storage systems in Bangladesh. National Food Policy Capacity Strengthening Programme. Final Report CF # 2/08.
China	Grains	Various	On-farm	Harvest		Huang, T., Li, B., Shen, D., Cao, J. & Mao, B. (2017). Analysis of the grain loss in harvest based on logistic regression. *Procedia Computer Science*, 122, 698–705.
Fiji Islands	Vegetables	Tomato	On-farm	Post-harvest		Kumar, S., & Underhill, S. J. (2019). Smallholder farmer perceptions of postharvest loss and its determinants in Fijian tomato value chains. *Horticulturae*, 5(4): 74.
Ghana	Roots and Tubers	Yam	On-farm	Post-harvest		Ansah, I.G.K. & Tetteh, B.K. (2016). Determinants of yam postharvest management in the Zabzugu District of Northern Ghana. *Advances in Agriculture*.
Zimbabwe	Vegetables	Tomato	On-farm	Post-harvest		Macheka, L., Spelt, E.J., Bakker, E.J., van der Vorst, J.G. & Luning, P.A. (2018). Identification of determinants of postharvest losses in Zimbabwean tomato supply chains as basis for dedicated interventions. *Food Control*, 87, 135-144.
India	Vegetables	Various	On-farm	Post-harvest		Meena, M.S. & Singh, K. (2009). Farmer's attitude towards post-harvest aspects of horticultural crops. *Indian Research Journal of Extension Education*. 9, 15–19.
Sub-Saharan Africa	Grains	Maize	On-farm, markets	Storage		Kaminski, J., & Christiansen, L. (2014). Post-harvest loss in sub-Saharan Africa – What do farmers say? *Global Food Security,*. 3 (3-4): 149–158.

Sri Lanka	Vegetables	Carrots, cabbage and leeks	On-farm	Post-harvest		Dharmathilake, N.R.D.S; Rosairo, H.S.R; Ayoni, V.D.N & Herath, R.M. (2020). Implications of post-harvest losses and acreage response of selected upcountry vegetables from Nuwara-Eliya District in Sri Lanka on sustained food security. *Journal of Agriculture Sciences,* 15(1)
United Republic of Tanzania	Grains	Maize	On-farm	Storage		Ngowi, E.R. & Selejio, O. (2019). Post-harvest loss and adoption of improved post-harvest storage technologies by smallholder maize farmers in Tanzania. *African Journal of Economic Review*, 7(1): 249−267.
United Republic of Tanzania	Grains	Maize	On-farm	Storage		Ngowi, E.R. & Selejio, O. (2019). Post-harvest loss and adoption of improved post-harvest storage technologies by smallholder maize farmers in Tanzania. *African Journal of Economic Review*, 7(1), 249−267.
China	Grains	Rice	On-farm	Harvest and post-harvest		Qu, X., Kojima, D., Nishihara, Y., Wu, L. & Ando, M. (2020). Impact of rice harvest loss by mechanization or outsourcing: Comparison of specialized and part-time farmers. *Agricultural Economics/Zemědělská Ekonomika*, 66(12): 542−549.
Papers that fill food loss data gaps in surveys, apply modelling to reduce costs (smaller sample size)						
Ethopia	Cereals	Various	On-farm	Post-harvest		Hengsdijk, H. & De Boer ,W.J..(2017). Post-harvest management and post-harvest losses of cereals in Ethiopia. *Food Security,* 9, 945–958.

Annex II: List of indicators mentioned in the literature as determining factors of harvest and post-harvest losses

Table 17: List of indicators mentioned in the literature as determining factors of harvest and post-harvest losses

I) Household/ producer socioeconomic characteristics
Education
Education of producer/respondent/HH head (Number of years)
Secondary education (Dummy variable, 1= if got the secondary education, otherwise 0)
Age and gender
Age of HH head/ respondent (Years)
Sex of household head/gender household heat (1=male, 0= female)/Respondent (male or female)
Respondent is female (1=yes, 0=no)
Number of adult males working in the farm
Number of adult females working in the farm
Income and access to credit
Average annual income of the respondents/Total HH income
Inverse hyperbolic sine of total household asset value
% of income coming from farming
Occupation (Dummy variable, =1 if household head's primary occupation is farming, otherwise 0)/Agriculture as a primary occupation (Yes=1, No=0)
Agriculture as a secondary occupation (Yes=1, No=0)
Having credit access (Yes=1, No=0)
Communication
Ownership mobile (Y/N)
Having internet access (Yes=1, No=0)
Having electricity access (Yes=1, No=0)
Source of information
Family size and type of family
Family size/Household size (Number of person)
Type of family dummy [value '0' for joint family and '1' for nuclear family]
II) Agricultural Activity
Size of the farm and production
Farm size/Cropping area (ha)/Area of land cultivated (hectare)/area allocated for potato
Inverse hyperbolic sine of total land area owned (acres)
Farm number of plots
Quantity produced/ Total production/Quantity harvested
Inverse hyperbolic sine of production quantity (kg)
Inv hyp sine of gross value of agricultural output
Production per ha
Land under the crop (%)/ Area allocated for potato
Farming experience
Member in the farmer club (proxy for agricultural experience and access to agricultural advice service and other services)
Farming decision taken by the household (Yes=1, No=0)
Risk score (1-10) (whether respondents generally see themselves as a person who is fully prepared to take risks =10)
Input and technology

Having exclusively rainfed farming (Yes=1, No=0)
Area under irrigation
Input usage (cost of improved seed, fertilizer, chemicals, labor etc.)
Variety used/ Maize variety (Dummy variable, 1= if maize variety is SETO local, otherwise 0)/ Improved variety
Having access to the extension service
Number of extension visit per year
Extension contact (1 = Yes; 0 = No)
Labour
Labour dummy which takes the value '1' if the labour availability during harvesting was adequate and value '0', otherwise.
Type of labour used for harvesting (0=family labour, 1= hired labour)
Number of active labor force
Pre harvest working days (man days)
Harvest working days (man days)
Cooperation/farmer-based organization
Member of producer organization/Farmer based organization/Group membership
Membership to cooperative (1 = Yes, 0 = No);
Input cooperation (1 = Yes, 0 = No);
Output cooperation (1 = Yes, 0 = No);
Time of harvesting
Time of harvest after maturity (days)/Age of fruit at harvest (months)
Decision to harvest '1' if fruit harvested when mature and '0' otherwise;
Early harvest (=1 if crop is harvested before first week of September, otherwise 0)
Criteria to harvest (1 = maturity; 0 = other)
III) Post-harvest characteristics
Markets connection and sales
Area under commercial crops (ha or %)
Own consumption (%)
Frequency of sales
Experience in markets (Years)
Distance to nearest market (km)/Distance from the farm to the market (km)/Market access (km)
Time to nearest market
Sale price / Current prices
Ready market (1 = Yes, 0 = No);
Time between harvesting and selling of produce (days)/days fruit spend on the farm (days)
Days fruit spend in the market before getting to the consumer (days)
On-farm storage
Storage facility (Dummy: 1=Yes; 0=otherwise)
Storage structure
Type of storage
Storage dummy which takes the value '1' if the storage facility was adequate and value '0'
Storage period (month)/ days of storage
Specific for cold storage
Electricity supply (h)
Outside temperature (°C)
Relative humidity in the cool room (%)
Pre-cooling time (h)
Good bag used (%)
Capacity utilization (%)
Inversion of bag (No.)
Maturity of stored potato (%)

Storage period (month)
Bag per stack (No.)
Floor type (wood = 1, other = 0)
Age of cold storage (year)
Packaging, transportation, threshing
Packaging dummy which takes the value '1' if the packaging is suitable and value '0' otherwise
Packaging (yes/ no)
Transportation dummy which takes the value '1' if transport facility was adequate and value '0' otherwise
Ownership of transport
Number of livestock owned by HH (Number)
Threshing machine dummy which takes the value '1' if availability of threshing machine during harvesting was adequate, '0', otherwise.
IV) Weather/Climate
Weather conditions at harvesting
Weather condition (Dummy: 1=Good; 0= otherwise)
Rainfall dummy which takes the value '1' if rainfall occur during harvest and value '0' otherwise
Weather dummy which takes the value '1' if the weather during harvesting was favorable and value '0', otherwise.
Month of harvest and geographic region (control for differences in rainfall patterns during the time of harvest and general geographic characteristics)
Precipitation preharvest (rainfall prior to harvest should be indicative of overall production, proxy for humidity patterns)
Precipitation during harvest and post-harvest (direct cause of losses)
Agro-ecological conditions, geographical conditions
Agro-ecological conditions (proxied with AEZ indicator variables)
Altitude (=1 if <800 masl and 800-1500 masl = 0)
District (dummy Rakai=1, otherwise =0)
Location and accessibility
Use of weather information
Use of past weather experiences (Yes=1, No=0)
Use of weather information (Yes=1, No=0)

Annex III: Technical notes on statistical procedures

Brief description of exploratory factor procedures used

The Exploratory Factor Analysis was applied to identify environmental factors that appeared to be relevant for the level of post-harvest losses. To do this, the Principal Factor Method was used to analyse the correlation matrix. This method seeks to find the fewest factors that can account for the common variance (correlation) of a set of variables.

To select the number of factors that seem relevant, the so called “Kaiser criterion” was used, which selects the factors with an eigenvalue greater than one. Additionally, the cumulative proportion of the variance explained by the factors is used as a complimentary criterion. These are expressed as a percentage, with the proportion showing the factor´s contribution towards explaining the variance. To choose the factors to be considered, the criterion established a cutoff point at more than 80 percent of the cumulative proportion of the variance.

To make the results obtained from EFA more reliable, and to better understand the resulting factors, the Varimax rotation method was used.

Finally, estimated factor coefficients were used to generate the factor scores used as environmental variables in the GSEM model.

Annex IV: Additional country examples on the modelling approach tested for sub-sampling losses in the farm or household surveys

Models obtained GSARS Zimbabwe – maize:

The classification established as post-stratification to generate a percent loss estimate was used as a predictor variable in a Poisson model. The output from this model is shown in Table 3.

This is a parsimonious model under which only one classification variable is used as the predictor, but it includes three independent variables in the classification criteria. To test the use of a Poisson model (natural log as link function) and the specification of the model with respect to the independent classification variable, the corresponding linktest is shown in Table 19. This test indicates that the specified on-farm loss model shows very good functionality for estimating the mean percent losses, where the linear prediction L̂ presents a significant coefficient equal to 1, meaning a perfect correspondence (1:1) to the observed percent losses (p = 0.034), and the square predicted L̂^2 has no predictive power (p = 1), with an estimated coefficient equal to zero. This is the ideal situation for model-based predictions. The estimated percentage loss of maize using a Poisson model is also 8.66 percent, but with a smaller standard error of 0.429 percent (95 percent% CI: 7.8, 9.5). This improved variance can be attributed to the loss classifications identified in the post-stratification procedure of the CART method, and shows an efficiency increase of 30.3 percent from the sample-based standard error to the model-based standard error.

The sampling-based estimate for the harvest and post-harvest percent loss of maize in Zimbabwe using the GSARS farm loss survey gives a mean on-farm loss of 4.0 percent with a standard error of 0.404 percent

(95 percent CI: 3.2, 4.8). The same data-driven procedure was used to improve the mean estimate, where a regression tree was built to generate post-stratification criteria, as shown in Figure 18. The regression tree selected four cutting points on three variables, namely the quantity of maize produced (q_production), age of the household (b4), and whether the household received any assistance from the government (f3). This tree arrives at five terminal nodes used as stratification to generate a Poisson model; the output was omitted.

This model fits properly. It shows a good linear relationship for the predicted value L^ with a coefficient estimate equal to 1 (p = 0.018), as shown in Table 19. The square predicted L^^2 has no predictive power (p = 1), and the estimated coefficient is zero, so the model passes the linktest showed in Table 19 The estimated percent loss of maize using the model is 4.0 percent, but with a smaller standard error of 0.259 percent (95 percent CI: 3.5, 4.5). This represents a high efficiency increase of 59.1 percent.

Figure 18: CART classification Zimbabwe GSARS

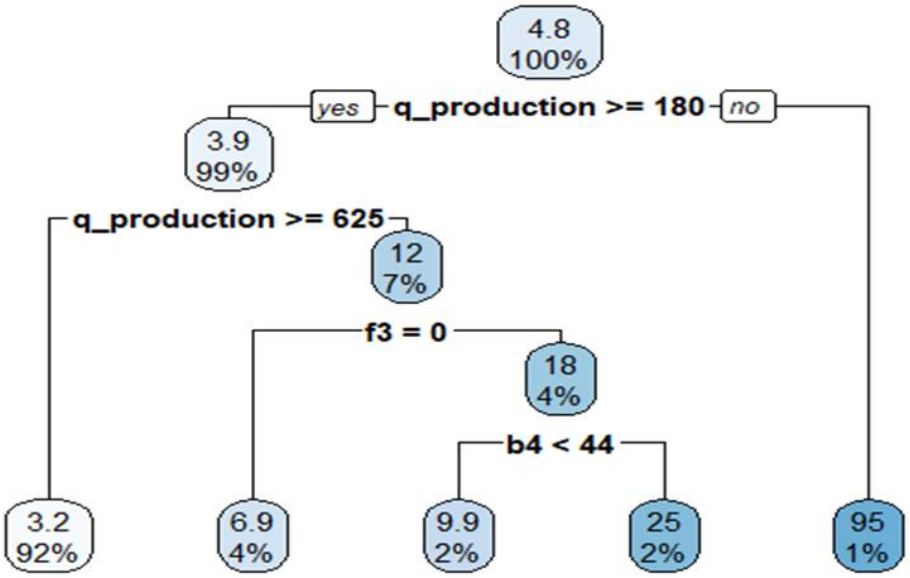


Table 18: Model specification tests, results from the linktests Zimbabwe GSARS

b) Zimbabwe GSARS			
Predictor	Coefficient	Std. Err.	P
$\hat{L}$	1.000	0.424	0.018
$\hat{L}^2$	0.000	0.073	1.000

Nigeria GHS 2015/16 –maize

The sampling-based estimate for post-harvest percentage losses of maize in Nigeria for the LSMS-ISA dataset provides a mean food loss of 10.1 percent with a standard error of 0.978 percent (95 percent CI: 8.2, 12.0). Data-driven procedures to improve the mean estimate generated a regression tree with seven cutting points on five variables: agro-ecological zone (hhgv_ssa_aez09), area planted (imp_area_planted), harvested quantity (imp_harv_qty), harvest length in days on average (harv_lenght), and plot elevation (plplotgv_elevation). This tree, shown in Figure 19, arrives at eight terminal nodes used as stratification to generate a Poisson model. The linktest for this model shows similar results (Table 20), with a correspondence of 1:1 between the observed percentage of post-harvest food loss and the linear prediction L^, but a coefficient not statistically different from zero (p = 0.119), and the square predicted L^^2 remains without predictive power, showing a zero coefficient (p = 1).

The estimated percentage loss of maize using this model is also 10.1 percent, but with a smaller standard error of 0.8.23 percent (95 percent CI: 8.5, 11.7). This represents an efficiency increase of 29.1 percent.

Figure 19: CART classification Nigeria GHS 2015/16

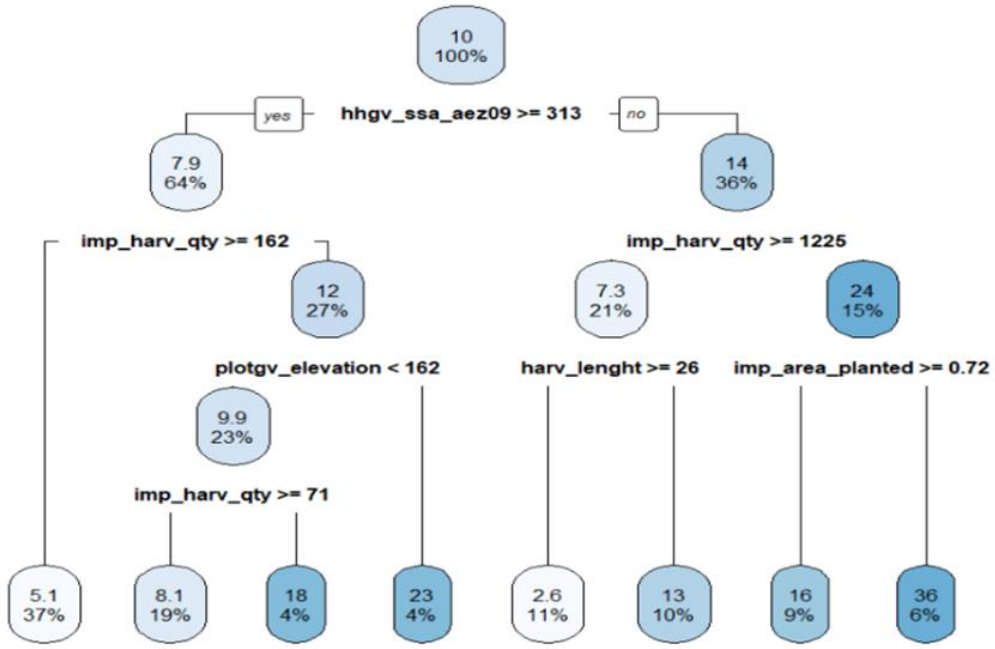


Table 19: Model specification tests, results from the linktests, Nigeria GHS 2015/16

d) Nigeria GHS 15/16			
Predictor	Coefficient	Std. Err.	P

$\hat{L}$	1.000	0.641	0.119
$\hat{L}^2$	0.000	0.134	1.000

Results obtained for the GSARS farm loss surveys in Zimbabwe

GSARS Zimbabwe – maize:

The same application of sample reduction on the food loss model was used for maize for the Zimbabwe survey dataset. In Table 21, a sample reduction of 50 percent achieved using model estimates, obtaining a relative efficiency of 9.4 percent with respect to the whole sample survey estimate.

Table 20: Estimates, standard errors and relative efficiencies for sub-sample – Zimbabwe GSARS

Sample reduction	Survey-based loss estimate		Post-stratification loss estimate (model-based)		Model relative efficiency
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_m$	$\hat{\sigma}_m$	
0%	3.98	0.404	3.98	0.259	59.1%
10%	4.09	0.450	4.09	0.277	53.1%
20%	3.85	0.429	3.85	0.298	45.7%
30%	3.80	0.447	3.80	0.310	41.1%
40%	3.87	0.508	3.87	0.347	26.3%
50%	3.99	0.572	3.99	0.384	9.4%

The use of imputed missing values in reduced samples as an option is presented in Table 22. The same concern is applicable here regarding standard error reductions with imputed values, which implies a greater risk of an increased probability of missing confidence interval estimates.

Table 21: Estimates, standard errors and relative efficiencies for imputed sub-samples – Zimbabwe GSARS

Sample reduction	Survey - based loss estimate		Estimates with model-based imputation	
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_i$	$\hat{\sigma}_i$
0%	3.98	0.404	3.98	0.404
10%	4.09	0.450	3.85	0.352

20%	3.85	0.429	3.96	0.356
30%	3.80	0.447	3.92	0.356
40%	3.87	0.508	3.81	0.323
50%	3.99	0.572	3.98	0.345

Results obtained for the Living Standard Measurement Studies in Nigeria

Nigeria GHS 2015/16 – maize

Sample-based and model-based estimates for post-harvest loss percentage for maize on the LSMS-ISA dataset is shown in Table 23.

Table 22 Estimates, standard errors and relative efficiencies for sub-sample – Nigeria GHS15/16

Sample reduction	Survey-based loss estimate		Post-stratification loss estimate (model-based)		Model relative efficiency
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_m$	$\hat{\sigma}_m$	
0%	10.10	0.978	10.10	0.823	29.1%
10%	9.85	1.024	9.85	0.866	21.6%
20%	10.17	1.122	10.17	0.959	3.8%
30%	9.76	1.111	9.76	0.954	4.7%
40%	8.75	0.997	8.75	0.869	21.0%
50%	9.36	1.154	9.36	1.018	-8.5%

It the case of Nigeria, compared to Malawi (previously shown), the gains in the standard error by using the food loss model are more relevant, as at least a similar relative efficiency with a sample reduction of 20-30 per cent is achieved. Table 24 shows survey and model-based estimates using imputed missing values on the reduced part of the sample.

Table 23: Estimates, standard errors and relative efficiencies for imputed sub-sample – Nigeria GHS15/16

Sample reduction	Survey-based loss estimate		Estimates with model-based imputation	
	$\hat{L}_s$	$\hat{\sigma}_s$	$\hat{L}_i$	$\hat{\sigma}_i$
0%	10.10	0.978	10.10	0.978
10%	9.85	1.024	9.39	0.877

20%	10.17	1.122	9.26	0.852
30%	9.76	1.111	9.58	0.835
40%	8.75	0.997	9.49	0.781
50%	9.36	1.154	10.12	0.773

Based on the results, model-based estimates using imputation procedures are not recommended for improving loss estimates because of the risk derived from artificial error reduction.

Annex V: General structural estimation model: estimation of the food loss model for Nigeria

Environmental factors can determine production and change from year to year by introducing external sources of variation not related to common determinants considered.

1.- Environmental Factors: Exploratory Factor Analysis was used to identify environmental factors among all observed environmental variables. The EFA first output for environmental variables recorded for Nigeria LSMS-ISA in the two Integrated Household Surveys is as follows.

Figure 20: Exploratory Factor Analysis, Environmental Factors, Nigeria

```
Factor analysis/correlation                      Number of obs    =      1,332
    Method: principal factors                    Retained factors =          4
    Rotation: (unrotated)                        Number of params =         54
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.72290	2.22536	0.3750	0.3750
Factor2	2.49755	0.34989	0.1983	0.5732
Factor3	2.14765	0.79119	0.1705	0.7438
Factor4	1.35647	0.43317	0.1077	0.8514
Factor5	0.92329	0.24541	0.0733	0.9247
Factor6	0.67788	0.26269	0.0538	0.9786
Factor7	0.41520	0.26845	0.0330	1.0115
Factor8	0.14675	0.12230	0.0117	1.0232
Factor9	0.02445	0.03285	0.0019	1.0251
Factor10	-0.00840	0.01204	-0.0007	1.0245
Factor11	-0.02044	0.02107	-0.0016	1.0228
Factor12	-0.04152	0.01038	-0.0033	1.0195
Factor13	-0.05190	0.01483	-0.0041	1.0154
Factor14	-0.06673	0.06069	-0.0053	1.0101
Factor15	-0.12741	.	-0.0101	1.0000

For this output, the Kaiser criterion indicates a selection of four common factors, representing a cumulative variance of 85.14 percent of the total variance present in the 15 variables included. Factor loadings and unique variances (uniqueness) for this EFA is, as follows:

Figure 21: Selection of main factors as by Kaiser criterion

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
hhgv_af_bi~1	0.1374	-0.5128	0.7122	0.3844	0.0632
hhgv_af_bi~8	0.0873	-0.5069	0.7348	0.3478	0.0745
hhgv_af_b~12	0.9311	-0.0109	-0.0667	0.1454	0.1073
hhgv_af_b~13	0.8423	0.1716	-0.2697	0.1114	0.1760
hhgv_af_b~16	0.8815	0.0894	-0.2493	0.2281	0.1008
hhgv_sq1	0.7187	0.1104	0.1705	-0.2751	0.3665
hhgv_sq2	0.6494	0.1519	0.2806	-0.3323	0.3660
hhgv_sq3	-0.3522	0.6780	-0.0895	0.4536	0.2024
hhgv_sq5	0.1722	0.6906	0.6040	-0.0534	0.1258
hhgv_sq6	0.1633	0.6909	0.6082	-0.0558	0.1230
hhgv_sq7	-0.3693	0.6651	-0.0700	0.4426	0.2204
hhgv_h_in_~t	0.8608	-0.0641	-0.1143	0.2170	0.1948
hhgv_h_in_~Q	-0.2405	-0.2061	-0.0196	0.4500	0.6968
hhgv_h_end~t	0.5330	-0.0369	-0.2043	0.2666	0.6017
hhgv_end_w~Q	0.0709	0.0215	-0.2251	0.2961	0.8562

Three environmental variables were excluded due to uniqueness > 0.5; these variables showed a poor contribution to the main environmental factors to be considered in the model. The second EFA obtained 12 environmental variables.

Figure 22: The second EFA obtained with reduced number of environmental variables

Factor analysis/correlation — Number of obs = 3,111
Method: principal factors — Retained factors = 4
Rotation: (unrotated) — Number of params = 42

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.37688	1.69162	0.4072	0.4072
Factor2	2.68526	0.60568	0.2498	0.6571
Factor3	2.07958	0.98975	0.1935	0.8505
Factor4	1.08983	0.51951	0.1014	0.9519
Factor5	0.57032	0.36235	0.0531	1.0050
Factor6	0.20797	0.21388	0.0193	1.0243
Factor7	-0.00591	0.01991	-0.0005	1.0238
Factor8	-0.02582	0.00863	-0.0024	1.0214
Factor9	-0.03445	0.01719	-0.0032	1.0182
Factor10	-0.05164	0.00723	-0.0048	1.0134
Factor11	-0.05886	0.02604	-0.0055	1.0079
Factor12	-0.08490	.	-0.0079	1.0000

For this output, the Kaiser criterion indicates a selection of four common factors, representing a cumulative variance of 95.19 percent of the total variance present in the 12 variables included.

The Varimax rotated eigenvalues representing the variance for each factors are, as follows:

Figure 23: Varimax rotated eigenvalues, GSEM Nigeria

```
Factor analysis/correlation                    Number of obs    =      3,111
    Method: principal factors                  Retained factors =          4
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         42
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	3.92563	1.48657	0.3652	0.3652
Factor2	2.43906	0.46954	0.2269	0.5922
Factor3	1.96953	0.07220	0.1832	0.7754
Factor4	1.89733	.	0.1765	0.9519

Rotated factor loadings for the Nigeria environmental factors are, as follows:

Table 24: Rotated factor loadings for the Nigeria environmental factors, Nigeria

Variable	f1	f2	f3	f4	Uniqueness
Annual mean temperature (degC * 10)	0.0344	0.0356	0.9669	-0.078	0.0566
Mean temperature of wettest quarter (degC * 10)	-0.0126	0.044	0.9583	-0.1034	0.0689
Annual precipitation (mm)	0.9288	0.024	0.1377	-0.1422	0.0976
Precipitation of wettest month (mm)	0.92	0.0155	-0.1172	-0.0905	0.1315
Precipitation of wettest quarter	0.9463	-0.0036	-0.0332	-0.0878	0.0956
Nutrient availability	0.6021	0.4908	0.1631	-0.2255	0.3191
Nutrient retention capacity	0.4754	0.6164	0.1797	-0.1886	0.3262
Rooting conditions	-0.1278	0.1671	-0.1163	0.9139	0.1069
Excess salts	-0.0316	0.9367	0.0158	0.2028	0.0802
Toxicity	-0.0418	0.9392	0.0157	0.2036	0.0744
Workability (constraining field management)	-0.1632	0.157	-0.0922	0.9115	0.1094
Avg 12-month total rainfall(mm) for Jan-Dec	0.8302	-0.0358	0.0374	-0.0897	0.3001

For the interpretation of latent factors, all loadings over 0.6 or below -0.6 are green. Factor 1 includes five observed variables:

- hhgv_af_bio_12 = Annual precipitation (mm)
- hhgv_af_bio_13 = Precipitation of wettest month (mm)
- hhgv_af_bio_16 = Precipitation of wettest quarter
- hhgv_sq1 = Nutrient availability
- hhgv_h_in_tot = Avg 12-month total rainfall(mm) for Jan-Dec

These variables are precipitation indicators descriptors of precipitation patterns including nutrients availability, so factor 1 is called “Precipitation”. Factor 2 includes four observed variables:

- hhgv_sq2 = Nutrient retention capacity
- hhgv_sq4 = Oxygen availability to roots
- hhgv_sq5 = Excess salts
- hhgv_sq6 = Toxicity

These variables are soil physical and chemical characteristics, so factor 2 is called “Soil salt & toxicity”. Factor 3 includes two variables:

- hhgv_af_bio_1 = Annual mean temperature (degC * 10)
- hhgv_af_bio_8 = Mean temperature of wettest quarter (degC * 10)

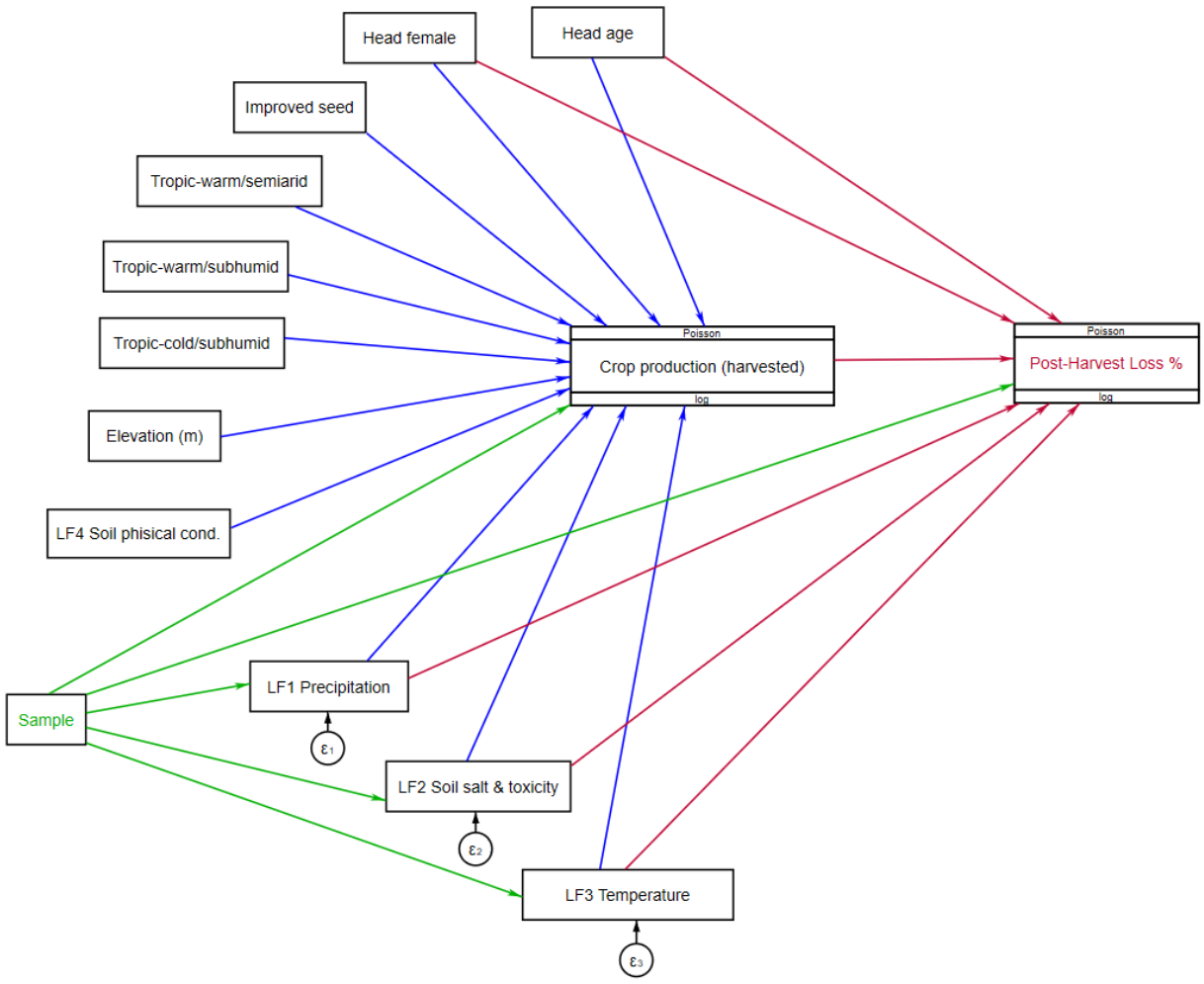
These are annual temperature variables, so factor 3 is called “Temperature”. Finally, factor 4 includes the last two variables:

- hhgv_sq3 = Rooting conditions
- hhgv_sq7 = Workability (constraining field management)

These variables represent rooting conditions of soil and the workability, so factor 4 is called “Soil physical conditions”.

3.– GSEM Estimation: A path diagramme for linear relationships between determinants of crop production and food losses for Nigeria is shown in Figure 24.

Figure 24 Path diagramme for linear relationships GSME Nigeria



In this representation, response variables are crop production and post-harvest losses, both modelled using a Poisson regression (loglink function and distributional family Poisson). In the bottom left of the diagramme, there is a box named "sample", which is a dummy variable and represents changes between survey rounds (year-to-year), and with linear effects on three of the four environmental latent factors (LF1 to LF3) and direct effects on the log of crop production and the log post-harvest losses in percent (green arrows). Environmental latent factors receive an arrow act like regression responses, each with their specific error term ε_i. At the top of the diagramme are two social variables, age of the head of the household and a dummy variable, indicating if their gender is female. Circling at the medium left are other determinants, including the use of improved seed, the agro-ecological zones, the field elevation and the latent variable related to soil physical conditions (LF4). Blue arrows represent the linear contribution of determinants on the log crop production, and red arrows represent linear contributions on the log post-harvest losses in percent. The model results are obtained using sampling weights for bias correction and standard error adjustment. The coefficient estimates for the full model, are presented in Table 26.

Table 25: Coefficient estimates for the full model, GSME Nigeria

Equation	Coef.	Robust Std. Err.	z	P	LB 95%	UB 95%
Crop production						
LF1 Precipitation	0.070	0.081	0.870	0.387	-0.089	0.229
LF2 Soil salt & toxicity	0.051	0.028	1.820	0.069	-0.004	0.107
LF3 Temperature	-0.415	0.118	-3.520	0.000	-0.646	-0.184
LF4 Soil physical conditions	-0.138	0.035	-3.960	0.000	-0.206	-0.069
age of household	0.005	0.003	1.980	0.048	0.000	0.010
Household female (yes)	0.114	0.186	0.610	0.542	-0.252	0.479
improved seed (yes)	-0.275	0.090	-3.040	0.002	-0.452	-0.098
Elevation (m)	-0.001	0.001	-2.660	0.008	-0.002	0.000
Tropic-warm/semiarid	-0.771	0.345	-2.230	0.026	-1.447	-0.094
Tropic-warm/subhumid	-1.056	0.298	-3.550	0.000	-1.639	-0.473
Tropic-cold/subhumid	-1.223	0.454	-2.690	0.007	-2.114	-0.332
sample (year)	-2.391	0.087	-27.490	0.000	-2.562	-2.221
_cons	8.545	0.342	24.970	0.000	7.875	9.216
ln(area planted)	1	(exposure)				
Post-harvest loss in percent						
Crop production	-0.0002	0.0001	-1.580	0.115	0.000	0.000
LF1 Precipitation	-0.030	0.131	-0.230	0.818	-0.286	0.226
LF2 Soil salt & toxicity	-0.540	0.233	-2.320	0.021	-0.996	-0.083
LF3 Temperature	0.095	0.085	1.120	0.264	-0.072	0.261
age of household	0.009	0.008	1.250	0.213	-0.005	0.024
Household female (yes)	0.449	0.287	1.560	0.118	-0.114	1.011
sample (year)	-1.114	0.318	-3.510	0.000	-1.737	-0.491
_cons	0.279	0.433	0.650	0.519	-0.569	1.127
LF1 Precipitation						
sample (year)	0.066	0.046	1.440	0.149	-0.024	0.155
_cons	-0.075	0.030	-2.500	0.013	-0.134	-0.016
LF2 Soil salt & toxicity						
sample (year)	0.003	0.044	0.070	0.943	-0.083	0.090
_cons	0.025	0.027	0.920	0.359	-0.029	0.079
LF3 Temperature						
sample (year)	-0.109	0.055	-1.990	0.046	-0.216	-0.002
_cons	0.000	0.029	-0.010	0.991	-0.057	0.056

var(e.f1)	1.003	0.033	0.940	1.070
var(e.f2)	1.176	0.414	0.589	2.345
var(e.f3)	1.157	0.077	1.016	1.318

The first equation is a Poisson regression to model crop production under which the size of the planted area is used as offset variable. Estimated coefficients show that precipitation had a positive effect and soil salt and toxicity and temperature had negative effects on crop production. Female household heads showed a negative effect on crop production. The use of improved seeds and lower terrain elevation are related to less crop production. There is a significant reduction in crop production for tropic-warm/semiarid, tropic-warm/subhumid and tropic-cold/subhumid agro-ecological zones with respect to the tropic-warm/humid agro-ecological zone.

The second equation is a Poisson regression to model post-harvest loss in percent in which a higher crop production is weakly related to a lower post-harvest loss. Soil salt and toxicity is related to less post-harvest losses. In the second survey (sampling round), there is a significant reduction in post-harvest losses. Only the fifth equation shows a reduction in temperature for the survey for the second round. At the bottom of coefficient estimation output, estimates of residual variance for linear regression equations representing the effects of sample year on the three environmental variables are given.

After the model reduction procedure, the crop production equation began with 12 paths, and in the final model, it contained only 10 paths. Precipitation and household gender were eliminated. The post-harvest losses in percent began with seven paths and ended with four paths; LF1 Precipitation, LF3 Temperature and the age of the household head were eliminated. Only the equation for the effect of sample year on LF3 Temperature remained in the model. The coefficient estimates for the final reduced model are presented in Table 27.

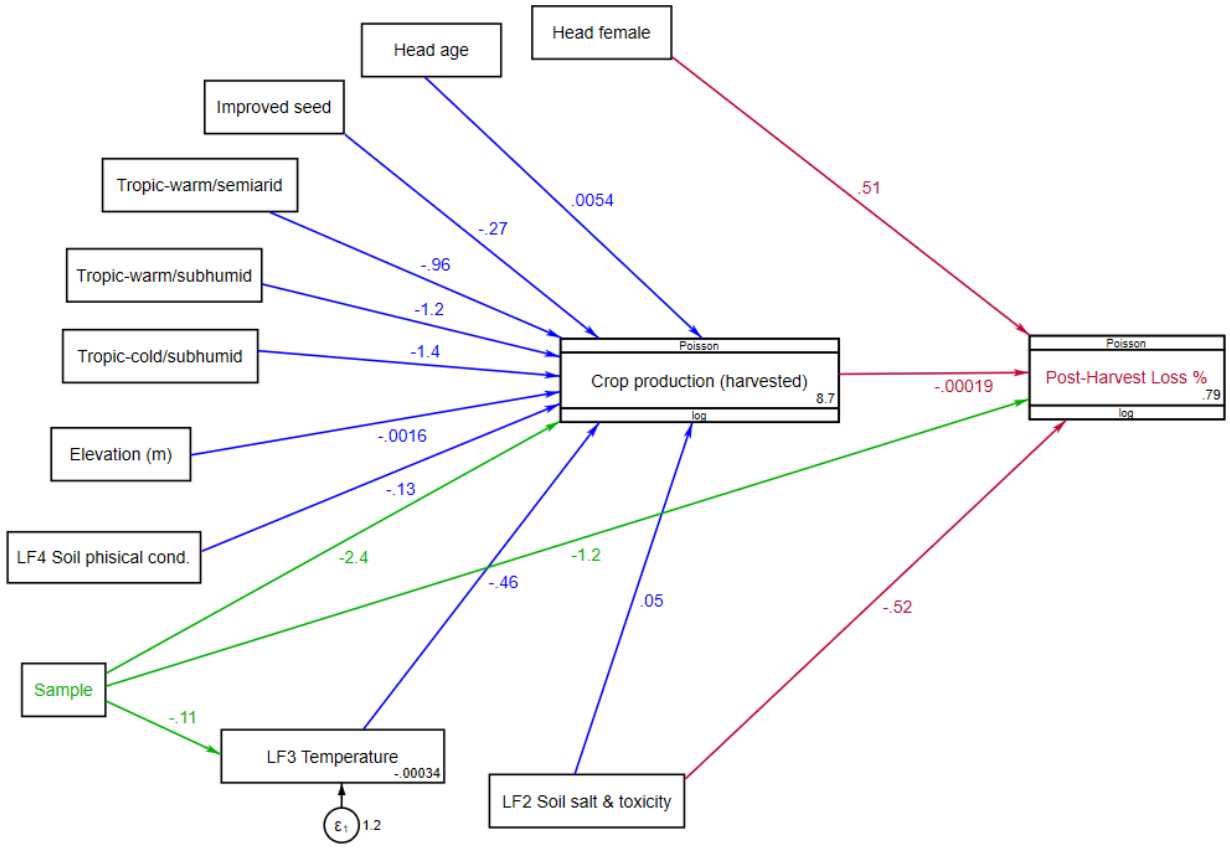
Table 26: The coefficient estimates for the final reduced model , GSME Nigeria

Equation	Coef.	Robust Std. Err.	z	P	LB 95%	UB 95%
Crop production						
LF2 Soil salt and toxicity	0.050	0.028	1.780	0.075	-0.005	0.104
LF3 Temperature	-0.460	0.111	-4.120	0.000	-0.678	-0.241
LF4 Soil physical conditions	-0.130	0.033	-3.990	0.000	-0.194	-0.066
age of household	0.005	0.003	1.990	0.047	0.000	0.011
improved seed (yes)	-0.273	0.091	-3.000	0.003	-0.451	-0.095
Elevation (m)	-0.002	0.001	-3.000	0.003	-0.003	-0.001
Tropic-warm/semiarid	-0.9598	0.2738	-3.510	0.000	-1.4965	-0.4231
Tropic-warm/subhumid	-1.223	0.251	-4.870	0.000	-1.715	-0.731
Tropic-cold/subhumid	-1.356	0.421	-3.220	0.001	-2.182	-0.530
sample (year)	-2.390	0.087	-27.380	0.000	-2.561	-2.219
_cons	8.746	0.279	31.330	0.000	8.198	9.293
ln(area planted)	1	(exposure)				
Post-harvest loss in percent						
Crop production	0.000	0.000	-1.760	0.078	0.000	0.000
LF2 Soil salt & toxicity	-0.518	0.231	-2.250	0.025	-0.970	-0.066
Household female (yes)	0.514	0.304	1.690	0.090	-0.081	1.110
sample (year)	-1.166	0.312	-3.740	0.000	-1.778	-0.555
_cons	0.794	0.173	4.600	0.000	0.456	1.133
LF3 Temperature						

sample (year)	-0.109	0.055	-1.990	0.046	-0.216	-0.002
_cons	0.000	0.029	-0.010	0.991	-0.057	0.056
var(e.f5)	1.157	0.077			1.016	1.318

This reduced model has the same interpretation of estimated direct effects on response variables obtained from the full model. The fitted path diagram for the reduced model, including coefficients for the direct effects, is presented in Figure 25.

Figure 25: Path diagramme for the reduced model, GSME Nigeria



There are no significant indirect effects on crop production or on post-harvest losses.

The mean estimate of post-harvest loss for the overall period, based on a model with two possibly distant survey rounds, is as follows:

Figure 26: Mean estimate of post-harvest loss for the overall period, Nigeria

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
_cons	**1.388514**	**.1635187**	**1.068023**	**1.709005**

An estimated 1.39 percent of maize crop production becomes post-harvest loss. The post-harvest loss estimates for each survey round, are as follows:

Figure 27: The post-harvest loss estimates for each survey round. Nigeria

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
_at				
1	2.327987	.3410498	1.659542	2.996432
2	.7253468	.1903545	.3522588	1.098435

The post-harvest losses were 2.33 percent in the first survey round, and it showed a significant reduction to 0.73 percent in the second survey round.

Finally, the linktest to validate the use of the GSEM model for estimation propose, showed the followings:

Figure 28: Linktest for the final GSEM model, Nigeria

Poisson regression — Number of obs = 3,110
Wald chi2(2) = 18.16
Log pseudolikelihood = -71083845 — Prob > chi2 = 0.0001

imp_per_phl	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	1.18944	.4565698	2.61	0.009	.2945797	2.0843
_hatsq	-.1395053	.0989441	-1.41	0.159	-.3334322	.0544216
_cons	-1.128435	.5116678	-2.21	0.027	-2.131285	-.1255845

The model passed this test correctly, as the squared prediction _hatsq has a non-significant coefficient and it is close to zero. The linearity in the prediction _hat is significantly positive, indicating that the model is a good instrument for post-harvest loss estimation.