Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operational Research

Estimating spillover effects on food insecurity in Uganda during the COVID-19 pandemic

Bram Baremans (578820bb)

Abstract

Food insecurity is a main issue in Uganda. Income losses and people getting fired due to the COVID-19 pandemic did not help either. But what are actually the driving factors of food concerns? This study investigates whether income loss and employment status are driving factors. Different from other studies, the regressors consist of the characteristics of other households in addition to their own, also known as spillover effects. By employing the data gathered by the High-Frequency Phone Survey distributed to households throughout Uganda, the structure and magnitudes of the interactions are recovered by the Double Pooled Lasso estimator. This method was first introduced by Manresa (2016). The structure of interactions shows that Central Uganda and the northern part of the country generate the most spillovers, whereas there is no region receiving substantially more externalities. In addition, the estimated effects of an income loss and being employed or not, are in line with existing literature; a positive effect for the former and a negative one for the latter characteristic.

Supervisor:	S.J. Koobs
Second assessor:	D.J.C Van Dijk
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Food insecurity is a severe problem around the world. Especially African countries struggle with this issue due to poverty, as Rose (2002) showed. Hence, poverty tends to be a driving factor of the food insecurity of a household.[1] While numerous studies have examined the underlying mechanisms of food insecurity, there is a growing recognition of possible interconnectedness between families. De Giorgi, Frederiksen and Pistaferri (2020) shed light on a similar problem; how the consumption of one is influenced by characteristics of others. Other examples of studies that investigated externalities are De Giorgi and Pellizzari (2014) (education), Liu, Patacchini, Zenou and Lee (2012) (crime) and Conley and Udry (2010) (technology adoption).[2] These externalities were estimated using an assumed structure of interactions. This included assumptions on who generates spillover effects on whom, only based on certain characteristics. This lacks theoretical justification which makes these results less reliable. Employing such a strategy to estimate the spillover effects on food insecurity is not preferable. Additionally, in recent years an upward trend occurred to estimate the structure of interactions through data-driven approaches. Therefore, a method is employed that estimates the structure of interactions based on the data, in which no pre-defined structures are included.

Based on the previous passage, we investigate the following research question: *Are social interactions driving factors of food insecurity in Uganda during the COVID-19 pandemic?*

The research question can be divided into sub-questions. Firstly, what is the structure of social interactions between different regions in Uganda?[3] Furthermore, this paper could show to which extent control variables drive the food insecurity score.

The relevance of this study extends in a multitude of ways. From a scientific perspective, this research fills an existing gap in the literature, namely that the Food Insecurity Experience Scale (FIES) score is probably driven by characteristics of other households. Not only knowledge on the social interaction of the income loss on food insecurity is gained, there will also be an understanding of the other driving factors of the FIES score in terms of included control variables; employment status in this application.

Besides its scientific relevance, this research also brings practical applications. Policy makers are highly eager to uncover what affects the food insecurity of households in order to reduce the insecurity accurately. Once the drivers are known, in terms of own characteristics as well as other's characteristics, policies can be applied in order to reduce the average food insecurity of the whole country. In general, policies consist of a treatment to a certain subgroup within the whole population in order to maximize the effectiveness of the policy.[4]

In order to answer the research question, the High-Frequency Phone Survey data for Uganda is employed. This survey was distributed to households in Uganda shortly after the COVID-19 pandemic started, with the purpose of gaining knowledge about the social and economic impact of this virus on families. As the survey has been distributed to households for a total of 11 rounds, the gathered data is actually panel data. Due to the lack of some survey rounds to

[1] "Households" can be replaced by "families", "house units" and "homes".

[2] Spillover effects are interchangeably used with externalities and relations.

[3] A region is similar to an area, a part and a zone in this study.

[4] An explicit example of how such an optimization problem looks like for this particular application is provided in Appendix E.

not satisfactory capture the results for the relevant questions, only six from all survey rounds are used. Based on answers to certain questions that occur in each round, dummy variables for income loss and being employed are constructed. A similar approach for the Food Insecurity Experience Scale (FIES) score is applied. After utilizing certain restrictions on the total sample, such as excluding families that moved between regions, the final sample consists of 535 households across Uganda.

The proposed methodology to recover the structure of interactions and the spillover effects is in accordance with the Double Pooled Lasso estimator introduced by Manresa (2016). To get to the Double Pooled Lasso estimator, the Post Pooled Lasso estimator has to be employed three times in total. This approach consists of applying a pooled Lasso to the data to recover the structure of the social network, after which a pooled Ordinary Least Squares (OLS) is performed to reduce the shrinkage bias. The foregoing steps can be applied to each house unit. The pooled Lasso step is improved compared to the original Lasso as it takes pair-specific weights into account. Besides, the optimal penalty term λ is based on theoretical justification rather than using k-fold cross validation. After executing the first two steps of the Double Pooled Lasso estimator, a pooled panel regression is performed to recover the coefficients of the control variables. The last step comes up with the final structure and magnitudes of the social interactions.

Gaining knowledge on what drives food insecurity is crucial as it has many detrimental consequences. For example, Tester, Rosas and Leung (2020) explain and use that chronic stress comes with food insecurity. Subsequently, chronic stress has a negative effect on people's physical health, especially children's as Shonkoff et al. (2012) found. An example of such an effect is the study of Jacobs and Bovasso (2000), which shows that breast cancer can be a consequence of chronic stress, which if not treated early enough/in the right way will lead to death. For Ethiopia, food insecurity was found to be the cause of malnutrition and had a high impact on early death from Noncommunicable Diseases (NCDs). This has been studied by Mosadeghrad, Gebru, Sari and Tafesse (2019) during the COVID-19 pandemic, the same period this research will focus on. Not only physical health suffers from stress, so does mental health. An explicit example is that food insecurity is associated with greater depression in rural Uganda, studied by Perkins et al. (2018). This yields for Lesotho too, as Marlow et al. (2022) found that food insecurity negatively affects people's mental health. Lesotho is like Uganda a sub-Saharan African country.

Considering the consequences of food insecurity, this research aims to obtain essential insights in how the FIES score of a house unit is influenced. From this, policy makers could try to reduce the food insecurity, which subsequently leads to less death due to chronic stress, malnutrition and depression. This thesis will take spillover effects into consideration by including characteristics of the other households as well to address whether externalities are part of the driving factors of the FIES score, which is the novelty of this study compared to previous research.

The results show that there are positive individual and social effects for the income loss in Uganda. This means that if a household loses (part of) their income, the food insecurity of that particular household increases. This yields for the social effects too: an income loss of households leads on average to an increase in the FIES scores of other families. Furthermore, the common effect of being employed is estimated to be negative, indicating that being employed leads on average to less food insecurity. Central Uganda seems to generate the most spillover effects, probably due to their huge contribution in food production mentioned by Leliveld et al. (2013). The northern region of Uganda exhibits a relatively diminished yet noteworthy spillover position. Other regions depend on this area as their position regarding pastoral livestock is strong, stated by Shively, Hao et al. (2012). Future policies are advised to focus on these parts of Uganda. The differences in receiving ability of spillovers are generally negligible.

Now the results for Uganda are known, what does this mean outside of Uganda? The structure of interactions is not applicable to other countries, but the found private and social effects possibly are. These might be similar for countries like Uganda. The same might yield for the effect of being employed.

The paper brings the following to discussion: Section 2 discusses previous research and the base theory behind the methodology, which is introduced in Section 3. The Monte Carlo simulation in Section 4 verifies the consistency of the method. The empirical application of this paper is executed in Section 6, employing the data mentioned in Section 5. Lastly, the research question is answered in Section 7 followed by a critical discussion.

2 Theoretical and scientific background

The existing literature on drivers of food insecurity is discussed in this section, together with the theory behind panel data and the Lasso estimator.

2.1 Existing literature

Previous research on the social impact effects on the food insecurity measure focuses on a predefined measure of social capital. An explanation of social capital is provided by Martin, Rogers, Cook and Joseph (2004), namely a measure of trust, reciprocity and social networks. The social networks part is of interest for this study. Martin et al. (2004) found that in the USA it is less likely to feel hungry with higher levels of social capital, using a logistic regression. A multinomial logistic regression is used in Malual and Mazur (2022) to find a strong positive link between social capital and household food security in a post-war area in Uganda. The higher the score on social capital, the more secure families feel about food. Sseguya, Mazur and Flora (2018) go a step further in defining the social capital measure. They used Principal Components Analysis to identify key factors of social capital, measured with both cognitive and structural indicators. Doing so, they found that being more socially active results in worrying less about food.[5]

A measure for social capital can be useful, but not in this study. This research aims to find the actual structure of interactions between households, with food insecurity as the dependent variable. Studies investigating on this matter are limited up till now. Despite limited investigation on spillover effects in the food insecurity setting, extensive research has been conducted on what drives the worries about food. Misselhorn (2005) examined the determinants of food insecurity in Africa. His findings revealed that poverty, environmental stressors and conflict emerged as significant drivers of being insecure about food. In addition, Semazzi and Kakungulu (2020) found a significant effect of having more land and a smaller household size on food insecurity.

[5] "Worries about food" and "food concerns" are substitutes of "food insecurity" in this paper.

Heading into a different direction, Dasgupta and Robinson (2021) and Agamile (2022) both found that the COVID-19 pandemic had detrimental consequences for the food insecurity in Uganda and other African countries. This study is not able to find the reaction of the food insecurity on the COVID outbreak as no data before the pandemic is recorded. However, deviations in the data could possibly be attributed to the effects of different periods during the crisis. For Nigeria, the distribution of income and the opportunities for education were found to be influencing food security, as well as the infrastructure and population growth, shown by Okpala, Manning and Baines (2021). Two other aspects that show to have a relation with food insecurity are income and having a job. Loopstra and Tarasuk (2013) perfectly show this relationship via their study on low-income households in Toronto, Canada. A raise in income or full-time employment both lead to a decrease in the food insecurity score. Simultaneously, these two drivers of food insecurity are highly correlated as income fully depends on whether someone is employed or not. We use variables similar to the ones of the latter research: a dummy variable for both income loss and employment status.

2.2 Panel data

To obtain the actual structure of interactions, panel data is employed. Panel data consists of data for each entity for multiple time periods. The data is a time series data for each entity or individual. Consequently, it captures both time varying changes per entity as well as cross-sectional heterogeneities at a specific point in time. Panel data comes with multiple advantages compared to other types of data sets. First, both within-unit and between-unit variation can be examined. Due to this feature, researchers can analyze the impact of individual-specific factors (within-unit) and factors that affect all units during a time period (between unit). An example for the latter can be the COVID-19 pandemic, that affected almost the whole world. Second, it facilitates the control of unobserved heterogeneity through a fixed or random effects model. This helps to address potential endogeneity issues. Panel data is especially attractive to policy makers. The outcomes can be compared before and after incorporating certain measures or treatments. By this, the impact of the policies can be captured and evaluated. Lastly, the estimation becomes more sharp and efficient as the sample size increases due to the multiple observations per unit. Overall, panel data is a valuable type of data that includes both cross-sectional and time series information.

2.3 The Lasso estimator

Generally known is that households or families do not have relations with everyone in their neighbourhood. Therefore, some spillover effects will be set to zero as no relation is present and the actual number of relations becomes sparse compared to the possible number of relations. To implement this wisely into the estimation, a panel version of the Least Absolute Shrinkage and Selection Operator (Lasso) regression will be employed. The original Lasso regression was first introduced by Tibshirani (1996). The difference with OLS estimation is that the Lasso approach includes a penalty term that constrains the sum of the absolute values of the regression coefficients. The main goal of Lasso regression is to strike a balance between model complexity and predictive accuracy. By penalizing the absolute values of the coefficients, Lasso regression encourages sparsity in the model, meaning it tends to set some coefficients to exactly zero. This property makes Lasso regression particularly useful for feature selection, especially in a high-dimensional sparse setting, whereas regular methods like OLS often fail. The Lasso estimator automatically identifies and selects the most relevant predictors while discarding the less important ones. Simultaneously, this leads to the prevention of overfitting due to the sparse structure. This is due to the shrinkage bias; the estimated coefficients are biased towards zero due to the penalty term. A drawback of the Lasso estimator regarding the empirical application is how it handles multicollinearity. When highly correlated regressors are present, Lasso selects only one of these while shrinking the others towards zero. From a different point of view this can be seen as an advantage as it automatically deals with multicollinearity. The objective function to be minimized by Lasso using panel data is similar to (3), which is discussed in more detail in Section 3. In (6), the Pooled Lasso estimator per individual is displayed, which comes even closer to the original Lasso estimator.

3 Methodology

In this section, we first introduce the model employed in this paper. Thereafter, we discuss the estimation method on how to recover the structure of interactions.

3.1 The model

Based on the research question, this study aims to employ the following panel data model with spillover effects:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + w_{it}'\theta + u_{it}. \tag{1}$$

In (1), y_{it} is the outcome variable of individual i at time t. Besides that its own characteristic has an effect on the outcome, captured by β_i, the same characteristic of others might have an effect too. This spillover effect is captured by γ_{ij}, which represents the effect of the characteristic of individual j on the outcome of individual i. Together with capturing the magnitude of spillover, those γ_{ij}'s, so called pair-specific parameters, represent the structure of interactions. If $\gamma_{ij} = 0$, it means that individual j has no spillover effect on individual i. Not necessarily it holds that $\gamma_{ij} = \gamma_{ji}$, for all i and j, which indicates that the externalities are modelled asymmetrically. In addition to the effects of the characteristics, the outcome depends on an individual specific intercept which is time-invariant (α_i). This term takes unobserved heterogeneity into account. Another factor driving y_{it} is a control variable, which is included in w_{it}.[6] The effect of this control variable is time-invariant and individual-invariant. Put differently, the effect of the control variable is the same for all individuals over time. The last element to influence the outcome, is the idiosyncratic error term u_{it}, which is assumed to be uncorrelated with both x_{it} and w_{it}, for $i = 1, ..., N$ and $t = 1, ..., T$.

There might be some minor issues with the proposed model. Especially when the number of individuals N exceeds the number of observations per individual T. In that case, the model

[6] Note that w_{it} does not generate spillover effects, whereas x_{it} does.

is not identified as the number of parameters exceeds the number of observations. To reduce the number of relevant parameters, the sparsity assumption is introduced which assumes that individuals only have a few relations with others. Until estimation it remains unknown which individual interacts with whom and what the corresponding magnitude is. Mathematically the sparsity assumption is represented by:

$$\sum_{j \neq i} \mathbb{1}\{\gamma_{ij} \neq 0\} = s_i << T. \tag{2}$$

This assumption states that the number of relations for each individual, denoted by s_i which is unknown, should be small relative to the total number of time periods. In some situations sparsity might not be the best assumption to apply, for instance if T is relatively low against a relatively high N. One way of relaxing the sparsity is to not consider individuals, but look at relations between certain clusters. It is expected that more relations within a cluster occur than between clusters.

3.1.1 Key metrics

By now, it should be clear that the individual effect is captured by β_i and that the spillover effects are captured by the γ_{ij}'s as well as the structure of interactions. A downside of the model is that there are many parameters estimated, which complicate interpretation. To summarize the estimation results for spillover models, the literature has developed two metrics: the private effect and the social effects. These measures are, in a somewhat different manner, extensively used for almost 50 years now. One of the first times they were employed, was in the study of Mansfield, Rapoport, Romeo, Wagner and Beardsley (1977).

The private effect explains itself sufficiently; this is the effect caused by an individuals own characteristic, which is captured by β_i. To summarize these individual specific effects, the average can be calculated: $P = \frac{1}{N}\sum_{i=1}^{N} \beta_i$. From this statistic it can be obtained to which extent own characteristics affect the FIES scores.

Social effects are slightly more complicated to find. It represents the average change in the outcome of others due to a change in the characteristics of individual j. The average spillover effect generated by individual j is calculated by $M_j = \frac{1}{N}\sum_{i \neq j}^{N} \gamma_{ij} + \frac{1}{N}\beta_j$. The first part represents the average spillover effect of j and the second part captures the fact that the average spillover effect is also affected by the individual coefficient of j. The M_j's are in fact the average marginal effects; it tells us how an outcome of others on average change due to a change in the characteristic of individual j.

The average of these spillover effects is often considered to be a policy parameter, that is $M = \frac{1}{N}\sum_{j=1}^{N} M_j$. They can be used to optimize the allocation of a certain treatment such that the biggest possible, positive utility change takes place. A concrete example with corresponding optimization problem based on the empirical application of this paper is provided in Appendix E.

3.2 Estimation

3.2.1 Post Pooled Lasso estimator

For simplicity, consider (1) with $\theta = 0$, that is, a model with no common effect. As we use panel data, this study aims to use the Pooled Lasso estimator. For the sake of the estimation, let Γ denote the $N \times N$ matrix containing the β_i's on the diagonal and the γ_{ij}'s off-diagonal. Then the Pooled Lasso estimator for Γ is equal to

$$\widehat{\Gamma} \in \underset{\Gamma}{\operatorname{argmin}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \sum_{j=1}^{N} \gamma_{ij} \tilde{x}_{jt} \right)^2 + \frac{\lambda}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_{ij} \left| \gamma_{ij} \right|. \tag{3}$$

The $\tilde{y}_{it}$ and $\tilde{x}_{jt}$ are constructed in a similar way via a within transformation, where $\tilde{y}_{it} = y_{it} - \frac{1}{T} \sum_{t=1}^{T} y_{it}$. Here, λ is the penalty parameter, that should be set by the researcher. The ϕ_{ij}'s are pair-specific weights depending on the data. A common choice for ϕ_{ij}^2 is an estimator of $\mathbb{V} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \tilde{u}_{it} \tilde{x}_{jt} \right)$. Using such weights leads to a sharper choice of the penalty parameter λ as the choice does not depend on the variability of the noise in estimation. The Pooled Lasso estimator is useful as it estimates the structure of interactions between individuals. Despite the relevance of each regressor is captured in the Lasso estimator, the actual estimates are not meaningful in terms of significance as no standard errors are provided. Besides, due to the penalty term, coefficients are biased towards zero which is called the shrinkage bias. To get rid of this shrinkage bias and to get statistically meaningful estimates, a second step will be performed: while maintaining the structure of the interactions found by Lasso, the model is estimated using a Pooled OLS regression. By this the magnitudes of the spillover effects are recovered via the following function

$$\widehat{\Gamma}^{P} \in \underset{(\gamma_{i1}, \ldots, \gamma_{iN}) : \gamma_{ij} = 0 \text{ if } j \notin \hat{T}_i}{\operatorname{argmin}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \sum_{j=1}^{N} \gamma_{ij} \tilde{x}_{jt} \right)^2, \tag{4}$$

which is called the Post Pooled Lasso estimator. Using the estimates that follow from this equation, an estimator for the social effects can be constructed as follows:

$$\widehat{M_j} = \frac{1}{N} \sum_{i \neq j}^{N} \widehat{\gamma}_{ij}^{P} + \frac{1}{N} \widehat{\beta}_{j}^{P} \tag{5}$$

Essentially, the estimation of (3) and (4) boils down to applying a pooled Lasso regression and pooled OLS to each individual time-series, respectively. The penalty term changes from $\frac{\lambda}{NT}$ to $\frac{\lambda}{T}$ as each individual is evaluated separately. This will lead to the Pooled Lasso estimator for each individual, which inherits the vital characteristics from the Lasso estimator. This indicates that the Pooled Lasso estimator is sparse by rows and therefore the estimator is well defined even when the number of regressors is larger than T. Manresa (2016) proved several theoretical properties such as consistency under a few assumptions. The objective to be minimized to get the Pooled Lasso estimator for each individual is provided in (6), which has a similar explanation as (3).

$$\hat{\gamma}_i \in \underset{(\gamma_{i1},...,\gamma_{iN})}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \sum_{j=1}^{N}\gamma_{ij}\tilde{x}_{jt}\right)^2 + \frac{\lambda}{T}\sum_{j=1}^{N}\phi_{ij}\left|\gamma_{ij}\right|. \tag{6}$$

3.2.2 Double Pooled Lasso estimator

For the case where control variables are present, i.e. $\theta \neq 0$, the Double Pooled Lasso estimator can be employed. The estimation procedure is the following. First, $w_{it}^d = \eta_i^d + \sum_{j=1}^N \lambda_{ij}^d x_{jt} + e_{it}^d$ and $y_{it} = \mu_i + \sum_{j=1}^N \nu_{ij} x_{jt} + v_{it}$ are estimated using the Post Pooled Lasso estimator with weights $\phi_{ij}^{d^2} = \widehat{\mathbb{V}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{x}_{jt}\tilde{e}_{it}^d\right)$ and $\phi_{ij}^2 = \widehat{\mathbb{V}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \tilde{x}_{jt}\tilde{v}_{it}\right)$, respectively. The former estimation contains $d \in \{1, ..., D\}$, where D is the number of control variables. Secondly, the common effect $\widehat{\theta}$ can be obtained by a pooled panel regression of $\tilde{y}_{it} - \widehat{\nu}_i \tilde{x}_t$ on $\tilde{w}_{it} - \widehat{\lambda}_i \tilde{x}_t$. This step can be justified by the Frisch-Waugh-Lovell theorem.[7] The purpose of this theorem is to find the common effect θ by regressing the estimated residuals of the latter regression on the estimated residuals of the former. Due to this, the coefficient represents the effect of the part of w_{it} uncorrelated with x_{it}. This forms the foundation for comprehending the individual contribution of each variable in a multivariate regression analysis. The last step is to estimate the structure of interactions and the spillover effects using the Post Pooled Lasso estimator with $\tilde{y}_{it} - \widehat{\theta}\tilde{w}_{it}$ as outcome variable and $\breve{\phi}_{ij}^2 = \widehat{\mathbb{V}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \left(\tilde{u}_{it} + \left(\widehat{\theta} - \theta^0\right)\tilde{w}_{it}\right)\tilde{x}_{jt}\right)$ as weights.

The main advantage of making use of the Double Pooled Lasso estimator is that it leads to the minimization of omitted variable bias arising from selection mistakes due to the double selection procedure adopted from Belloni, Chernozhukov and Hansen (2014). The double selection methodology provides a robust estimator for θ that achieves convergence at the optimal rate, even in situations where the prerequisites for flawless model selection are not met. The double selection constructs orthogonal projections for y and w on the regressors $x_{1t}, ..., x_{Nt}$ separately, due to which the omitted variable bias is minimized. Once these orthogonal components are constructed, the residuals of y are regressed on the residuals of w by a pooled regression to obtain a consistent estimate for θ. This procedure is in line with Chamberlain (1992), who proposed this method for high-dimensional, sparse random coefficients models. Therefore, this research assumes the model in (1) is a sparse, high-dimensional model with a random coefficients vector $\gamma_i = (\gamma_{i1}, ..., \gamma_{iN})$.

The actual calculation for λ is given by $\lambda^* = c2 \cdot \sqrt{T}\Phi^{-1}\left(1 - \nu/\left(2N\right)\right)$.[8] The denominator within the standardized Gaussian cumulative distribution function Φ can be replaced by $2N^2$. Both ways of calculating the penalty term are justified as both provide consistent estimates for each time-series Lasso estimation. For the application of Manresa (2016) they even gave qualitatively similar estimates. The simulation and empirical application focuses on the λ^* given here. In Appendix C.4 the difference between the two calculation methods is evaluated. The weights are constructed according to the HAC type estimator proposed by Newey and West (1986), which is robust to autocorrelation and heteroscedasticity of unknown structure. These weights are updated after each Post Pooled Lasso estimation. As the updated weights are based

[7] For a detailed explanation, see https://en.wikipedia.org/wiki/Frisch%E2%80%93Waugh%E2%80%93Lovell_theorem.

[8] The penalty term introduced by Manresa (2016) is marked with * as later on this penalty term will be divided by a number to account for multicollinearity.

on the estimated residuals, the initial weights look like this:

$$\phi_{ij}^{2(0)} = \frac{1}{T}\sum_{t=1}^{T} \tilde{x}_{jt}^2 \tilde{y}_{it}^2 + \frac{1}{T}\sum_{t=2}^{T} \tilde{x}_{jt}\tilde{x}_{jt-1}\tilde{y}_{it}\tilde{y}_{it-1}.$$

Subsequent iterations construct the weights similarly, except that $\tilde{y}_{it}$ is replaced by $\widehat{\tilde{u}}_{it} = \tilde{y}_{it} - \sum_{j=1}^{N} \widehat{\gamma}_{ij}^{(0)} \tilde{x}_{jt}$. These estimated residuals make use of the estimated coefficients found by the Lasso regression, $\widehat{\gamma}_{ij}^{(0)}$ estimated with weights constructed in the previous iteration. The iteration procedure continues until the largest difference between the current and previous weights is smaller than a certain threshold set by the researcher.

4 Replication part

Prior to applying the methodology discussed in Section 3 to the data set of this study, verification of its consistency is crucial. Therefore, the methodology is applied to simulated data to show how well it works under certain assumptions made during the data generation. First, the data generating process is introduced and discussed. Subsequently, the Double Pooled Lasso is employed on the simulated data to show the practical consistency of this method using simulated data for different time horizons and sample sizes.

4.1 Data generation

Data generating process We generate the data in this Monte Carlo simulation based on appropriate distributions and corresponding characteristics based on the real data set employed by Manresa (2016).[9] Contrary to the application of Manresa (2016), only one control variable is included here to generate the outcome variable. A justification for this is that the empirical analysis of this paper includes one control variable. The control variable we include in the simulation process is generated based on the characteristics of the number of employees (l) as a measure for labor. We base the variable that captures the spillover effects on the knowledge capital, represented by R&D expenditures by the companies (x) in the real data set. The dependent variable (y) is a measure for productivity, namely the number of sales. Based on this information, we construct the model for the dependent variable:

$$y_{it} = \alpha_i + \beta_i x_{it} + \sum_{j \neq i} \gamma_{ij} x_{jt} + l_{it}'\theta + u_{it}. \tag{7}$$

The model for the simulation analysis is similar to the model that is employed in the empirical application. For explanation of the parameters and the estimation procedure, see Section 3.

[9] The data set Manresa used in her paper can be found on `https://nbloom.people.stanford.edu/research`. Search for “Identifying Technology Spillovers and Product Market Rivalry” and choose for Main Data. Within that folder, the “spillovers.dta” file contains the actual data set.

Variable characteristics Manresa (2016) uses the natural logarithm of the original variables. Therefore, we base the DGP on the characteristics of the logarithmic values of the variables she uses. By evaluating these characteristics, it seems that both the log R&D expenditure and log number of employees follow a normal distribution approximately.[10] For the expenditure a mean of 2.969 with corresponding standard deviation of 1.918 is determined. The logarithmic number of employees shows a mean of 1.382 with a standard deviation of 1.761. We generate the idiosyncratic error term using a standard normal distribution.

To improve the performance of the method to recover the true parameters, we draw the self chosen parameters from random normal distributions with a considerable mean and standard deviation. For instance, the α_i's are drawn with a mean of 2.7 and a standard deviation of 0.3. The magnitudes of the β_i's and γ_{ij}'s are substantially larger, as they are drawn with means of 100 and 70 with standard deviations of 20 and 5, respectively. Not all coefficients are given a drawn number, only a pre-specified number of relations. Lastly, the common effect θ is drawn from a random normal distribution with mean 10 and standard deviation 1. Using the generated variables and coefficients, we construct the dependent variable by (7). Once these variables are constructed, a demeaning takes place. These components are used in the estimation procedure as they cancel out the unobserved heterogeneity.

4.2 Simulation results

Below, we provide the results of the simulations regarding methodological consistency.

We show methodological consistency under certain assumptions on how the data is generated, by increasing the number of time periods while keeping the other characteristics the same. To measure the performance of the Double Pooled Lasso estimator, we provide the Frobenius norm for the difference matrix between the real and estimated coefficients. This should converge to zero as the number of time periods increases to show consistency. Besides, the accuracy of estimating θ is evaluated by the difference between the real and estimated coefficient. How these measures are exactly constructed and calculated, can be found in Appendix B.1. For this simulation, $\lambda = \lambda^*$ is used. The number of iterations regarding the weights is defined by the difference in weights matrices before and after each iteration. If this difference is larger than a certain threshold, the loop will continue. As soon as none of the elements exceed the threshold, the loop stops. In this simulation the threshold is set to 0.000001. To avoid noise, a Monte Carlo simulation is performed by executing the simulation and estimation six times. A larger number would lead to even more reliable results, however performing it more than once already leads to the cancellation of noise. Therefore, six rounds is chosen to keep the run time acceptable.

4.2.1 Double Pooled Lasso

Different time horizons Below the results of a six unit Monte Carlo simulation for the Double Pooled Lasso estimator are shown. The number of relations is equal to one and the number of individuals equal to five, whereas the number of time periods will be varied.

[10] The evaluation of the characteristics of the logarithmic number of employees and R&D expenditure are described in Appendix C.1.

Table 1 shows the performance of the Double Pooled Lasso estimator for the case described above. The Frobenius norm shows methodological consistency in this particular setting; as the number of time periods increases, the estimates become closer to the real values. Consistency in estimation of the common effect θ is present as well, as the difference between the real common effect and the estimated value converges to zero as the number of time periods increases.

Table 1: The Frobenius norm and difference between the real and estimated common effect for λ calculated with N for different time horizons, with $N = 5$.

T	Fnorm	θ_diff
100	0.185	0.393
200	0.139	0.194
500	0.063	0.067
1000	0.050	0.040
5000	0.023	0.007

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

Different group sizes Increasing the number of time periods leads to consistency of the Double Pooled Lasso estimator for a sample size of five. To verify whether this conclusion holds for larger samples, $N = 10$ and $N = 20$ are compared to the results in the previous paragraph.

The findings for the different number of individuals against different time periods are demonstrated in Table 2. The Frobenius norm shows that the larger the number of individuals is, the worse the estimation performance gets. However, for all three cases the Frobenius norm still shows consistency in estimation as time increases. The estimation of θ provides us with similar information; the larger the number of time periods, the closer the estimate is to the real value of θ. The differences between the sample sizes are similar too, as the estimate of the common effect is further away from the real value for a larger sample.

Table 2: Frobenius norm and the difference between the real theta and the estimated theta for λ calculated with N for different time horizons and different number of individuals.

	Fnorm			θ_diff		
T	N=5	N=10	N=20	N=5	N=10	N=20
100	0.185	0.441	2.399	0.393	0.910	1.888
200	0.139	0.214	0.463	0.194	0.427	0.883
300	0.102	0.155	0.327	0.144	0.312	0.698
400	0.070	0.118	0.197	0.073	0.193	0.489
500	0.063	0.098	0.173	0.067	0.179	0.391

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

In addition to these results, Appendix C.2 shows the forecasting performance and percentage correct zero's, as well as the performance for a different number of relations. Furthermore, Appendix C.3 shows the Monte Carlo results for the Post Pooled Lasso estimator.

5 Data

This section discusses the data for the empirical application. First, we mention the data source and discuss how the data is collected. Second, the collected data is transformed to useful variables. Lastly, we explore the characteristics of the final sample.

5.1 Source

To investigate the research problem, this thesis aims to employ a panel data set. We find a satisfactory data set by means of the data collected by the High-Frequency Phone Survey (HFPS), which is part of the Living Standards Measurement Study (LSMS) conducted by the World Bank. The HFPS has been distributed to households in five sub-Saharan African countries: Ethiopia, Malawi, Nigeria, Tanzania and Uganda. Initially, the questionnaire was handed out shortly after the start of the COVID-19 pandemic to understand the impact of the virus on the economic and social aspects of households.

This research focuses on the data of only one country of five, namely the data on Uganda.[11] This choice is not necessarily based on specific criteria, although the relationship between food insecurity and social interactions has already been studied for rural Uganda by Perkins et al. (2018).

In Uganda, 11 rounds of the survey have been recorded in total, which ensures that the available data is actually panel data with $t = 1, ..., 11$. The first four rounds were distributed between June and November 2020, shortly after the first case of COVID-19 on the 22nd of March 2020. Rounds 5 and 6 were sent out between February and April 2021, followed by round 7 in September of that same year. The last set of survey rounds (8 to 11) was distributed between June 2022 and January 2023. Each round lasted approximately a month, with varying time between the surveys. From round to round, some questions and categories of questions differ. However, the questions of interest for this study are similar throughout the different rounds. Recurrent categories are for example the Food Insecurity Experience Scale, income loss, employment and agriculture. An example of a varying category is the behaviour or knowledge about COVID-19. The latter was only present in the baseline survey, whereas the former changed alongside the changing restrictions.[12]

5.2 Data transformation

In order to utilize the data to its full potential, a transformation is required. The dependent variable is the Food Insecurity Experience Scale (FIES) score, which is based on eight binary questions regarding food insecurity. These questions involved whether household members were worried about the food they had, whether it was (nutritious) enough and whether they skipped meals. The FIES score is a raw score consisting of the sum of affirmative answers to the survey, having a score of 1 if answered "yes" and 0 if "no". Therefore, the FIES score ranges from 0

[11] The data set for Uganda can be found on the World Bank website, `https://microdata.worldbank.org/index.php/catalog/3765` A description of the data collection, variables and exact time period can be found here too.

[12] The timeline of restrictions and COVID-19 cases in Uganda can be found via `https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Uganda`.

to 8, which is a similar construction procedure as Wambogo, Ghattas, Leonard and Sahyoun (2018). Using this way of construction, a score of 8 means the most food insecure, whereas a score of 0 means the least food insecure.

The explanatory variable will be a dummy variable representing whether a household experienced a loss in total household income. The original question gave the answer options "increased", "stayed the same", "decreased" and "total loss/no earnings". The latter two answers are considered to represent income loss. If someone answered the question with one of those two answers, their loss dummy is equal to one. Besides the regressors, a control variable is included as well in terms of an employment dummy. This variable is based on the employment status of the person that filled in the survey and has the common effect θ. To come to a dummy outcome, two questions were evaluated. The first one recorded whether someone actively tried to generate income the week before. The second question takes holidays, illness and unexpected events into account. It asks the representative, if the answer to the former question is "no", whether they expect to return to their income generating activities. If the answer is "yes" to this question, it can be assumed that the person is actually employed but missed a few working days due to the events described before.

Once the variables are constructed via the appropriate ways, it becomes clear that the total income loss question is not correctly captured by all survey rounds. Actually, only six rounds have captured it sufficiently to effectively use it. The specific rounds of the survey that lacked on this aspect are rounds 1, 5, 7, 8 and 9. Due to this inconvenience, only data of the six remaining rounds can be used. Hence, for this study, T is equal to 6.[13] The remaining rounds 2, 3, 4, 6, 10 and 11 are given a specific time period number, respectively from 1 till 6 (round 2 = 1, round 3 = 2,..., round 11 = 6).

5.3 Sample construction and characteristics

Regarding the sample that will be employed in this study, only the households that participated in all six rounds considered are included. The remaining sample after this type of selection consists of 1404 households. However, even when this is the case, there are some exceptional cases that will be left out. Such exceptional cases include families that moved between regions during the different rounds of the survey. These households may bias the results as they will represent two or more regions during the time periods, whereas the spillover effects are assumed to be time-invariant. Besides leaving the moved households out, households that report the same value for a variable in all time periods (the FIES score, loss dummy or employment status) are left out because the estimation strategy uses the demeaned variables. Households with the same outcome for each round will end up with the demeaned variable being equal to zero, which gives complications in the estimation process. Ultimately, the final sample consists of 535 house units spread over Uganda, which can be divided into five groups based on the regions in Uganda. One of the area's is the capital, Kampala. The other regions are based on cardinal points: Central, East, North and West.[14] The distribution of participating households across these regions is more or less equal, except Kampala. Only 25 households are located in the capital. The East

[13] The exact time periods that are included for the data collection can be found via `https://microdata.worldbank.org/index.php/catalog/3765#metadata-data_collection`.

[14] The division of Uganda into these regions is shown in Appendix D.1.

contains the most households with 145, closely followed by the North (134) and Central Uganda (124). The remaining house units are located in western Uganda (107 homes).

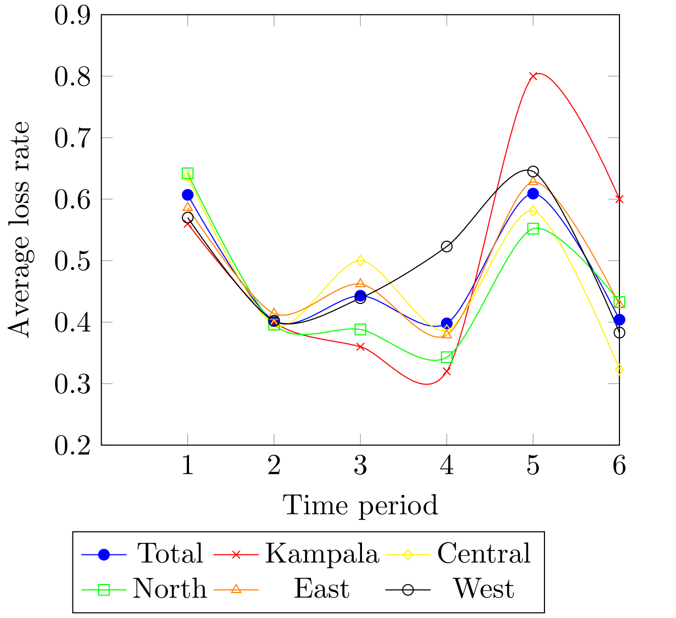


Figure 1: The percentage per sample that experiences an income loss during each time period.

Figure 1 shows the time series of the percentage of each region and the total sample that experiences an income loss during the survey rounds considered. All samples seem to start off with a loss rate ranging between 0.55 and 0.65, being slightly higher for Central Uganda and the North. At $t = 2$ the loss rates converge even more to around 40%. During round 4 of the survey ($t = 3$) the number of families that experience an income loss rises, whereas in Kampala it reduces even more. During this period, the number of new COVID-19 cases in Uganda rose to its maximum since the start, which explain the rises in loss rates. These small increases are followed by decreases in the next period, aside from the West which gain about 10% in the loss rate. For $t = 5$, the income loss rate increases for all samples to about 55-65%, excluding the 80% in Kampala. An explanation for this could be the Ebola outbreak during this survey round in the autumn of 2022, as the first Ebola case was close to Kampala.[15] At the end of the last survey round, the Ebola epidemic disappeared and less people experienced an income loss. Nonetheless the gap between Kampala and the other sample is still there.

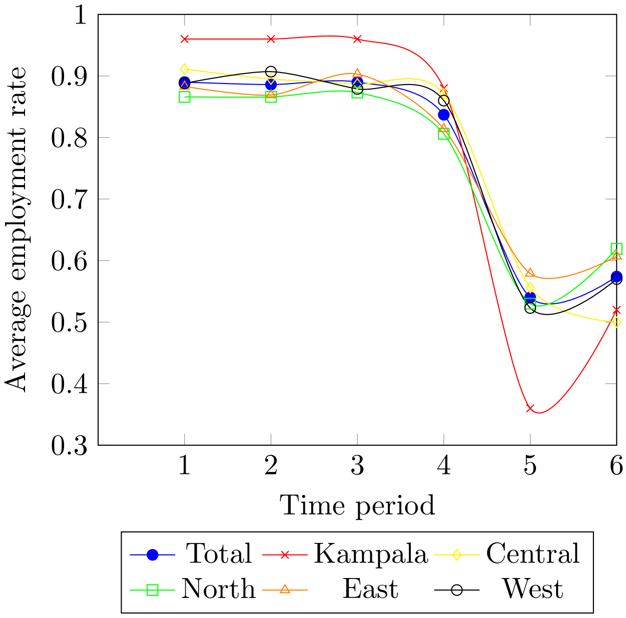


Figure 2: The percentage per sample that is employed at the moment of the survey each time period.

The development of labor participation for each survey round is presented in Figure 2. From this plot, some interesting features of the data are visible. All areas seem not to be affected much by the COVID-19 outbreak, lasting until round 4 ($t = 3$). Only some small fluctuations are present for Central Uganda, the West and their counterpart, the East. In time period 4, the descent starts. During the start of the Ebola crisis, a

[15] The first case was confirmed on 20 September 2022 in the Mubende district, which lies in the wake of Kampala. Later, districts in the West were mainly affected. The World Health Organization explains more about this Ebola outbreak in Uganda (see `https://www.who.int/emergencies/situations/ebola-uganda-2022`).

lot of households got unemployed, with around 55-60% working in each sample and only 36% in Kampala. Again, this can be explained by the fact that the virus was first discovered near Kampala, in the Mubende district in Central Uganda. What stands out is that the drop in Central Uganda is not as big as in Kampala. However, regarding the last survey round, all samples seem to recover except the area of the Mubende district which goes from 56% to 50%. This might indicate that this zone was too careful in letting the measures of the Ebola epidemic go, due to which even more people were fired.

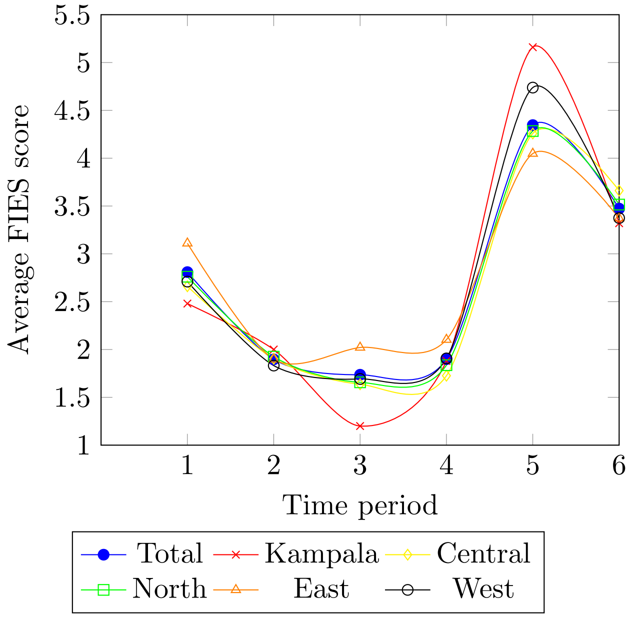


Figure 3: The average Food Insecurity Experience Scale score per sample during each time period.

Regarding the Food Insecurity Experience Scale score (FIES score), the averages per time period for each sample can be found in Figure 3. East Uganda starts with the highest food insecurity (3.11), followed by the total sample, North, West and Central Uganda all around 2.7. In Kampala, households are the least worried about food (2.5). During the quiet time periods 2, 3 and 4 all samples encounter an approximate U-shaped reduction in the food insecurity of their area, to face their peaks during the Ebola outbreak in September 2022. However, families in Kampala feel even more confident about food at $t = 3$. Looking at Figure 1, Kampala experiences a reduction in people that receive a pay cut, which might explain this drop in the average FIES score. During the Ebola era, tables have turned for Kampala: at $t = 5$, households are on average the most worried about getting enough food with a score of 5.2. The other main character in the Ebola scene in terms of areas, is the West, which is the second most worried about food (4.7). This relatively high food insecurity could also be due to the landslides caused by heavy rain around 7 September 2022 in the Kasese District in West Uganda which is about a month before $t = 5$.[16] All in all, it seems that the Ebola crisis had a considerable impact on the food insecurity score, loss rate and employment rate. At some points, it even seems to have more impact than the COVID-19 pandemic. However, considering the downward slope at $t = 1$ for the income loss rate and FIES score, it is within the realm of possibilities that the rates were already dropping before and came from an even higher number. The first survey round (not considered here due to missing data) was straight after the first COVID-19 case in Uganda, which could have caused even higher rates of income loss and food insecurity. In addition, the peaks of households in Kampala in the figures are questionable as the number of households in Kampala included (25) is relatively small compared to the other regions with around 125 households. Increasing the number of households located in Kampala would give a better representation of the population in Kampala.

[16] More information on this natural disaster can be found on this site: https://en.wikipedia.org/wiki/2022_Kasese_District_landslides.

6 Results

This section shows what the estimated structure of interactions between households in Uganda is, according to the Double Pooled Lasso estimator. In advance, we discuss the settings of the empirical application.

6.1 Settings empirical application

The settings of certain parameters are different in the empirical application compared to the simulation analysis. First of all, the threshold regarding the difference between two iterations of the weights is set to 0.1 in this application. Furthermore, the analysis is performed using multiple penalty terms. This is due to the fact that the bounds used for construction of λ^* become conservative when there is high correlation between the income loss dummy across households. Knowing that the number of observations is six and the number of households is respectively large, the probability of facing multicollinearity is high. To account for this multicollinearity, λ^* can be divided by an integer which is not too large.[17] To get even more reliable results, the order of the households is randomized. Originally, the households were clustered by region within the data. Regarding the correlation statistics of the regressors, it can be concluded that the data is highly collinear. Only seven loss dummy time series are not perfectly correlated (i.e. have a correlation of 1) with other time series, whereas the rest is identical with at least one other time series. There are even time series that are exactly the same as 42 other time series. On average, each loss time series is perfectly correlated with about 19 others.

Section 2 stated that in case of multicollinearity, Lasso picks one of the multicollinear predictors and puts the rest to zero. The results without randomizing the regressors for a penalty term of $\lambda = \lambda^*/3$ are shown in Appendix D.3. These suggest that the algorithm to select predictors, picks the first variable of the multicollinear columns and puts the rest to zero as Kampala has the highest amount of relations ánd is the first cluster to appear in the regressor matrix. This, together with this discussion about multicollinearity using Lasso, encourage randomizing the order to partly take away the bias.

The structure of interactions is displayed using the percentage of relations compared to the total number of possible relations rather than the actual number of relations.[18] This is due to Kampala having a relatively small number of households (25) compared to the other regions (around 125). Using the percentage makes comparison easier and more reliable, even though Kampala has substantially less households.

6.2 Double Pooled Lasso

Table 3 provides the estimated effects by the Double Pooled Lasso, using $\lambda = \lambda^*/3$.[19] The private effect is estimated to be 0.00862, which means that households that experience an income loss have on average a FIES score that is 0.009 higher compared to households with no income loss. Worth mentioning is the fact that only two individual effects (β_i's) are estimated to be non-zero. One of the households that affects itself is located in Central Uganda and the other in the

[17] This is the penalty term suggested by Manresa (2016). The calculation is provided by Section 3.2.2.

[18] The actual calculation of this percentage is described in Appendix D.2.

[19] In Appendix D.4 the same analysis is performed with different penalty terms.

eastern part of the country. Besides the private effect, such a household wage reduction leads on average to an increase of 0.0014 in the FIES scores of other households. This refers to the average social effect, which is based on 745 non-zero estimated externalities. The average social effect might not be direct network effects as it could be a correlated time shock as well. Take for example the Ebola crisis; a change in food insecurity is in this case probably not caused by other households income losses, but due to the Ebola outbreak. Nevertheless, both the private and average social effect are in line with Okpala et al. (2021), who states that a reduction in income leads to more worries about food. Lastly, the common effect θ, which represents the effect of being employed on the FIES score, is estimated to be -0.16. This indicates that a family of which the representative is employed, experiences less food insecurity as the score drops by 0.16 on average. Being employed has a highly significant effect on the FIES score as the p-value is equal to 0.005, which means that the estimate is significant on the 1% level. The estimated effects of both an income loss and being employed are in line with Loopstra and Tarasuk (2013), who found that an income reduction leads to more worries about food, whereas being full-time employed results in a decrease in concerns about food.

Table 3: The estimated private, social and common effect, using $\lambda = \lambda^*/3$. The standard error is in the parenthesis for the common effect.

	Estimate
Private effect	0.00862
Average social effect	0.00139
Common effect θ	-0.160 (0.0574)

Besides the estimated effects, we show the structure of interactions in Table 4. Looking at the percentages provided in the table, the sparse structure of interactions pops up immediately as none of the numbers comes near 1%.

What catches our attention is that all regions interact with each other. The extent to which this happens is another story to discuss. Households from Central Uganda are the most influencing group as 0.41% of the possible relations they can have is non-zero, followed by the North with 0.398%. Central Uganda contains the most agricultural production of food, which is further discussed by Leliveld et al. (2013). A phase marked by a decline in food production will result in more worries about food there, which will subsequently spill over to the other regions. The influence of the North can be attributed to its important role in pastoral livestock as discussed by Shively et al. (2012). On the one hand it brings the responsibility for meat, on the other hand harvesting crops using animals. If the pastoral livestock is not sufficient, the other regions will probably face the effects as well due to higher insecurity in the North. Especially the West is influenced by the North, not to mention that the interactions within the North are relatively high as well. Even higher percentages of within-interactions are found for Kampala (0.667%) and Central Uganda (0.439%).[20] However, the generated spillover effects on the other areas does not reach the levels of the North, except the influence of Central Uganda on the western part of the country (0.55%). The West seems to generate the least externalities

[20] Interactions percentage on the diagonal; households who interact with other households in the same region.

in percentage terms with only 0.11% of the total possible relations the West could have. The levels of both Kampala and the East are close to the West. This can be caused by Central and North Uganda, for taking a more important role in cultivating crops and pastoral livestock respectively. The other regions grow banana's, coffee and mixed food crops, which are nutritious and important as well. However, the importance of those compared to the role of the former two regions is smaller. Central and North Uganda seem to overtake a part of the spillover effects of the other regions due to their stronger position in food supply.

Looking at the receiving rate of spillover effects, the different regions do not differ much from each other, although Kampala seems to be less dependent than the other regions. We attribute this phenomenon to the urban farming introduced in Kampala during the 90's. Urban farming was found to be associated with higher food security among households by Maxwell (1995). Furthermore, Sabiiti and Katongole (2014) states that farming in the capital has become a vital process for the food supply in Kampala and therefore the city is less dependent on other regions. Households in the North and West of Uganda are the most depending on other households. The study of Funk et al. (2012) concluded that the West and North of Uganda are most affected by climate change. Especially rainfall declines threaten the future prospects of food production in these areas. Therefore, these regions depend on others in times of droughts.

Another interesting finding is the fact that Kampala and the West generate the same amount of spillover effects on each other (0.037%). This conclusion can be drawn for Central Uganda and the North as well with 0.403% of possible relations being estimated non-zero.

Table 4: Structure of interactions given as percentage of the total possible interactions between regions in Uganda, using $\lambda = \lambda^*/3$. The region by row is affected by the region by column.

region	Kampala	Central	East	North	West	Total
Kampala	0.667	0.226	0.221	0.299	0.037	0.225
Central	0.097	0.439	0.117	0.403	0.143	0.267
East	0.138	0.317	0.177	0.365	0.103	0.240
North	0.239	0.403	0.165	0.393	0.126	0.273
West	0.037	0.550	0.045	0.467	0.079	0.275
Total	0.157	0.409	0.136	0.398	0.110	

7 Conclusion

Food insecurity has been a major issue in sub-Saharan African countries throughout the years. The driving factors of these worries about food are of even more interest in order to be able to reduce it. By not only looking at household's own characteristics but taking other's characteristics into account as well, spillover effects can be identified. Both the structure and magnitudes of these spillover effects are of interest. Therefore, the main purpose of this study is to investigate *whether spillover effects are driving factors of the food insecurity score for Ugandan households during the COVID-19 crisis.*

To find the answer to this question, a survey within the Living Standards Measurement Study (LSMS) is employed. Eleven survey rounds were distributed to households throughout

Uganda, which makes the available data set a panel data set. After applying certain criterion, a final sample of 535 households is utilized in the actual analysis, together with six rounds of the survey. The Double Pooled Lasso estimator is employed to recover the structure of interactions in the Lasso step and subsequently the magnitude of these spillover effects is estimated via a pooled OLS regression. The advantage of this estimator is that the omitted variable bias arising from selection mistakes is minimized by creating orthogonal projections. This holds even under non-satisfied conditions for perfect model selection. However, the presence of omitted variable bias is still possible. The procedure uses demeaned variables to cancel out the individual fixed effects, or unobserved heterogeneity.

Before applying the methodology to the data, consistency of the Double Pooled Lasso estimator is verified using generated data under certain assumptions. Thereafter, the estimation procedure is applied to the data set of Uganda. The results are mostly in line with previous studies. A positive estimate is found for the private effect, meaning that an income loss leads to more worries about food. The same yields for the estimated social effects; an income loss of another household, related or not, leads on average to more worries about food. At the 1% significance level, the estimated effect of being employed indicates a statistically significant relationship, revealing a negative impact. Regarding the structure of the relations, the North and Central Uganda are the big game players in generating spillovers. This can be explained by their strong positions regarding pastoral livestock (Shively et al., 2012) and food production (Leliveld et al., 2013), respectively. The different regions do not show large differences in receiving spillover effects.

Taking the found results into account, the research question can be answered. Spillover effects regarding income loss are drivers of the food insecurity in Uganda during the COVID-19 pandemic. Additionally, the control variable in terms of the employment status is driving the food insecurity significantly.

7.1 Discussion

This project contributes to the existing knowledge on the drivers of the Food Insecurity Experience Scale score in Uganda, while facing the COVID-19 pandemic, by taking characteristics of other households into account. Other families tend to influence people's food insecurity based on income loss. Similar to Loopstra and Tarasuk (2013) and Okpala et al. (2021), this study finds that food concerns are driven by income loss as well as being employed. Besides the new knowledge gained, this is one of the first papers that takes dummy variables as regressors employing the Double Pooled Lasso estimator. Reason for this might be the fact that demeaned variables are used. Once a household has the same dummy outcome throughout time, the demeaned value becomes zero which is not preferable and will not work out.

Other countries for which the results may be relevant, are like Uganda sub-Saharan African countries that are facing moderate poverty. Examples of such countries are Ethiopia, Malawi, Nigeria and Tanzania. To verify whether the results also hold for such countries, a similar analysis can be performed as to the aforementioned countries the same survey is distributed.

Despite the strong aspects and contributions of this study, improvements can be made. First of all, this study assumes the FIES score to be a continuous variable as this makes it possible to employ the pooled OLS in the estimation procedure. However, this FIES score is actually a discrete variable ranging from 0 to 8, being there nine classes. Therefore, an extension of this study would be one that drops the pooled OLS and employs an ordered multinomial logistic model with nine classes. This would add to the existing literature as it adjusts the Double Pooled Lasso estimator in terms of estimating the coefficients.

Building on the previous points, it would be wise to extend the simulation analysis by including dummy variables as regressors. This is useful to investigate whether the Double Pooled Lasso estimator is still consistent when dummies are used as regressors. Now it is assumed that it works but actual validation would invigorate the use of dummy variables in this method. In addition, such a simulation can help in determining the best penalty parameter. The empirical application now tries different values for λ by dividing it by 2, 3 and 4. This shows that multicollinearity is tempered but there is no theoretical justification for which value is best to use. An extra simulation step would help in determining the optimal value to deal with predictor interdependence. Another option would be using cross-validation.

The multicollinearity issue encountered in this study is that huge, that the less strict penalty term is not sufficient to solve this problem. Now the columns are randomized to get even more reliable results. However, the randomization is done once, whereas a different order would have led to different results probably. Therefore, employing a data set with less multicollinearity gives more reliable results as the multicollinearity bias has a smaller impact. One way to temper multicollinearity is to employ data that contain more observations per household, i.e. the time period is larger. A second option is to not employ dummy variables as the chance of having the same sequence using only 0's and 1's is higher compared to using more than two values.

Additionally, this study does not verify or check whether the found average spillover effects (M_j's) are statistically significant different from zero, something Manresa (2016) did. She regressed the estimated demeaned residuals on the demeaned regressors $\tilde{x}_{1t}, ..., \tilde{x}_{Nt}$. The coefficients represent estimates for the M_j's.

7.2 Different pathways

Besides improvements on this study, deviations regarding methodology and variables can be made. This thesis only considered income loss and employment status are as drivers of the FIES score. Omitted variable bias is likely to be present in this study as the literature review in Section 2 shows other proven driving factors of food insecurity. Therefore, further analysis should focus on including more control variables in order to recover an even more reliable structure of interactions. For example, Misselhorn (2005) suggests to include poverty, conflicts and environmental stressors. Including lagged outcome variables could be an option as well, something Manresa (2016) discussed in the Supplementary Appendix of her paper.

This research focused on using Lasso regression in recovering the structure of interactions. Soloveva (2020) wrote her Bachelor thesis about social interactions, but used other methods to do so. For example, she used a different version of the Lasso estimator, Elastic Net and Smoothly Clipped Absolute Deviation (SCAD). The latter method was first proposed by Fan and Li (2001). Further research could investigate how these proposed methods work for the FIES scores and whether it is applicable in this case.

References

Agamile, P. (2022). Covid-19 lockdown and exposure of households to food insecurity in uganda: Insights from a national high frequency phone survey. *The European Journal of Development Research*, *34*(6), 3050–3075.

Balestra, P. & Krishnakumar, J. (2008). Fixed effects models and fixed coefficients models. *The econometrics of panel data: fundamentals and recent developments in theory and practice*, 23–48.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, 567–596.

Conley, T. G. & Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American economic review*, *100*(1), 35–69.

Dasgupta, S. & Robinson, E. J. (2021). Food insecurity, safety nets, and coping strategies during the covid-19 pandemic: Multi-country evidence from sub-saharan africa. *International Journal of Environmental Research and Public Health*, *18*(19), 9997.

De Giorgi, G., Frederiksen, A. & Pistaferri, L. (2020). Consumption network effects. *The Review of Economic Studies*, *87*(1), 130–163.

De Giorgi, G. & Pellizzari, M. (2014). Understanding social interactions: Evidence from the classroom. *The Economic Journal*, *124*(579), 917–953.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Funk, C., Rowland, J., Eilerts, G., White, L., Martin, T., Maron, J. et al. (2012). A climate trend analysis of uganda. *US Geological Survey Fact Sheet*, *3062*(4).

Jacobs, J. R. & Bovasso, G. B. (2000). Early and chronic stress and their relation to breast cancer. *Psychological medicine*, *30*(3), 669–678.

Leliveld, A., Dietz, A., Foeken, D., Klaver, W., Akinyoade, A., Smits, H., … van't Wout, M. (2013). Agricultural dynamics and food security trends in uganda. *Agricultural dynamics and food security trends in Uganda*(2).

Liu, X., Patacchini, E., Zenou, Y. & Lee, L.-F. (2012). Criminal networks: Who is the key player? *FEEM Working Paper No. 39.2012*.

Loopstra, R. & Tarasuk, V. (2013). Severity of household food insecurity is sensitive to change in household income and employment status among low-income families. *The Journal of nutrition*, *143*(8), 1316–1323.

Malual, J. D. & Mazur, R. E. (2022). Social capital and food security in post-conflict rural lira district, northern uganda. *Disasters*, *46*(1), 80–94.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. *MIT Sloan Working Paper*.

Mansfield, E., Rapoport, J., Romeo, A., Wagner, S. & Beardsley, G. (1977). Social and private rates of return from industrial innovations. *The quarterly Journal of economics*, *91*(2), 221–240.

Marlow, M., Skeen, S., Hunt, X., Sundin, P., Weiss, R. E., Mofokeng, S., … Tomlinson, M. (2022). Depression, anxiety, and psychological distress among caregivers of young children in rural lesotho: Associations with food insecurity, household death and parenting stress. *SSM-Mental Health*, *2*, 100167.

Martin, K. S., Rogers, B. L., Cook, J. T. & Joseph, H. M. (2004). Social capital is associated with decreased risk of hunger. *Social science & medicine*, *58*(12), 2645–2654.

Maxwell, D. G. (1995). Alternative food security strategy: A household analysis of urban agriculture in kampala. *World Development*, *23*(10), 1669–1681.

Misselhorn, A. A. (2005). What drives food insecurity in southern africa? a meta-analysis of household economy studies. *Global environmental change*, *15*(1), 33–43.

Mosadeghrad, A. M., Gebru, A. A., Sari, A. A. & Tafesse, T. B. (2019). Impact of food insecurity and malnutrition on the burden of non-communicable diseases and death in ethiopia: A situational analysis. *Human Antibodies*, *27*(4), 213–220.

Newey, W. K. & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *NBER Working Paper No. 55.*.

Okpala, E., Manning, L. & Baines, R. (2021). Socio-economic drivers of poverty and food insecurity: Nigeria a case study. *Food Reviews International*, 1–11.

Perkins, J. M., Nyakato, V. N., Kakuhikire, B., Tsai, A. C., Subramanian, S., Bangsberg, D. R. & Christakis, N. A. (2018). Food insecurity, social networks and symptoms of depression among men and women in rural uganda: a cross-sectional, population-based study. *Public health nutrition*, *21*(5), 838–848.

Rose, D. (2002). Quantitative indicators from a food expenditure survey can be used to target the food insecure in south africa. *The Journal of nutrition*.

Sabiiti, E. & Katongole, C. (2014). Urban agriculture: a response to the food supply crisis in kampala city, uganda. *The Security of Water, Food, Energy and Liveability of Cities: Challenges and Opportunities for Peri-Urban Futures*, 233–242.

Semazzi, J. B. & Kakungulu, M. (2020). Household determinants of food security in rural central uganda. *African Journal of Agricultural Research*, *16*(9), 1245–1252.

Shively, G., Hao, J. et al. (2012). A review of agriculture, food security and human nutrition issues in uganda. *AgEcon Search: research in agricultural & applied economics*(12-3).

Shonkoff, J. P., Garner, A. S., Siegel, B. S., Dobbins, M. I., Earls, M. F., McGuinn, L., … Wood, D. L. (2012). The lifelong effects of early childhood adversity and toxic stress. *Pediatrics*, *129*(1), e232–e246.

Soloveva, V. S. (2020). *Comparison of adaptive lasso, adaptive elastic net and scad in the context of recovering the network structure from panel data* (Bachelor's thesis, Erasmus University Rotterdam. Erasmus University Thesis Repository.) Retrieved from `http://hdl.handle.net/2105/53440`

Sseguya, H., Mazur, R. E. & Flora, C. B. (2018). Social capital dimensions in household food security interventions: Implications for rural uganda. *Agriculture and human values*, *35*(1), 117–129.

Tester, J. M., Rosas, L. G. & Leung, C. W. (2020). Food insecurity and pediatric obesity: a double whammy in the era of covid-19. *Current obesity reports*, *9*, 442–450.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wambogo, E. A., Ghattas, H., Leonard, K. L. & Sahyoun, N. R. (2018). Validity of the food insecurity experience scale for use in sub-saharan africa and characteristics of food-insecure individuals. *Current developments in nutrition*, *2*(9), nzy062.

Appendices

A Code description

Both the Post Pooled Lasso and Double Pooled Lasso estimator are implemented using Rstudio. The knowledge on the characteristics of the data set used by Manresa (2016) is gained using STATA.

The data set of Manresa (2016) is found on the website of Nicholas Bloom, which is provided in a zip-file containing STATA do-files. Therefore, the analysis on the data set of the original paper is performed using this program.

The results for both estimators are obtained using Rstudio, like the first paragraph stated. The generation of the data for the Monte Carlo simulation is done based on the characteristics of the data set of Manresa (2016), which is similar for both estimators. After the data has been simulated, either the Post Pooled or Double Pooled Lasso estimator can be calculated using the code. The Post Pooled Lasso performs Lasso using the "**glmnet**" function in R, where after a pooled OLS is performed. This is done for each individual separately. The results are stored in lists and matrices. This procedure is repeated until a certain threshold is not satisfied anymore, as each iteration the weights are updated. After that, the performance is evaluated by comparing the estimated coefficients with the real coefficients using the Frobenius norm. Other additional measures are calculated as well, which can all be found in Appendix B.

The Double Pooled Lasso estimator works a little different compared to the Post Pooled version. Namely, the Double Pooled makes use of the Post Pooled estimator three times. In the first step, the control variable (employment status) and the outcome variable (FIES score) are regressed on the income loss dummies of all households in two separate regressions. Subsequently, a pooled panel OLS is performed by regressing the residuals of the latter regression in the first step on the residuals of the former. The last step consists of regressing the FIES score minus the control variable times its estimated common effect on all household's income loss dummies.

The simulation results are similar in terms of measures to the Post Pooled Lasso estimator. The results regarding the empirical application to the Ugandan data set, are gathered differently. The private and average social effect are calculated using the estimates by the Double Pooled Lasso estimator. The formulas are discussed in Section 3. Furthermore, a binary matrix is created with 1's for non-zero estimated spillover effects and 0's when they are estimated to be zero. The diagonal elements are put to zero to not take individual effects into account in the relation matrix. After that, a relation matrix containing percentages is constructed by dividing the actual number of relations by the total possible relations. Here again, the individual effects are not considered to belong to the total possible relations.

B Performance measures simulation

In this Section of the Appendix, the performance measures used during the simulations are explained.

B.1 Frobenius norm and θ difference

Frobenius norm The Frobenius norm is calculated for the difference matrix between the real coefficients and the estimated values. Therefore, the estimated coefficient matrix is element wise subtracted from the real coefficient matrix. Once this difference matrix D has been constructed, the Frobenius norm can be calculated via its formula:

$$\|D\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|d_{ij}|^2}. \tag{8}$$

This measure shows how much the whole matrix deviates from the zero matrix, as the best scenario would be that all elements in the difference matrix D are zero. In that case all coefficients are estimated perfectly. However, that is not likely to happen due to noise present in the model. Therefore, the Frobenius norm is a good measure to pick.

θ difference Measuring the performance of the Double Pooled Lasso estimator regarding the estimation of θ, the difference between the real θ and the estimated θ is used. The estimated value is subtracted from the real value. A positive value means that the real coefficient is larger than the estimated one, so θ is underestimated in this case. Similarly, when the difference is negative, the common parameter for the control variable is overestimated, as the estimated value is larger than the real value. The mathematical equation corresponds to $\theta - \hat{\theta}$.

B.2 Mean Squared Error (MSE) and % correctly to zero

MSE intercept The real intercepts are stored in a vector, being there one for each individual. To estimate the intercepts, the following property of the fixed effects model is used: $\hat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\beta}_{FE,i}(-\bar{w}_i\hat{\theta})$. The coefficients of each individual are captured by $\beta_{FE,i}$ and includes β_i as well as the γ_{ij}'s for $j \neq i$ in the correct order. The part between brackets is added when considering the Double Pooled Lasso estimator. The estimator for the individual intercepts is consistent for $T \to \infty$ and independent of the number of individuals. More intuition behind this estimation of the intercepts can be found in Balestra and Krishnakumar (2008, Chapter 2.3.3).

Once the intercept is estimated, the Mean Squared Error can be calculated. This is done by squaring the difference between the real and estimated intercepts and taking the average over the individuals subsequently. In mathematical format:

$$\text{MSE}_{int} = \frac{1}{N}\sum_{i=1}^{N}(\alpha_i - \hat{\alpha}_i)^2 \tag{9}$$

MSE y To get a prediction of the outcome variable, first the intercept should be calculated via the procedure provided in the former paragraph. The actual calculation of the outcome variable is the following:

$$\hat{y}_{it} = \hat{\alpha}_i + \hat{\beta}_i x_{it} + \sum_{j \neq i}^{N} \hat{\gamma}_{ij} x_{jt} (+\hat{\theta} w_{it}). \tag{10}$$

All estimated coefficients are used, as well as the estimated intercept. For the regressors the original variables are used. The latter part is again only included when the Double Pooled Lasso estimator is evaluated.

The Mean Squared Error for the outcome variable is calculated via two steps. Step 1:

$$\text{MSE}_i = \frac{1}{T} \sum_{t=1}^{T} (y_{it} - \hat{y}_{it})^2, \tag{11}$$

which calculates the MSE for each person by evaluating the estimation of each time series. As the MSE is preferably summarized to one number, the average over the individual MSE's is taken:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \text{MSE}_i. \tag{12}$$

Percentage correct zero's The percentage correctly estimated to be zero is calculated in the following way: keep track of the real number of zero's and the estimated number of zero's. Subsequently, divide the estimated number of zero's by the real number of zero's and multiply by 100%. This can be written as the following math equation:

$$\% \text{ correct zero's} = \frac{\# \text{ estimated zero's}}{\# \text{ real zero's}} * 100\%, \tag{13}$$

where # is used as the number sign.

C Detailed simulation results

This section provides additional explanation and results of the simulation analysis in Section 4.2 of the main text. The results are explaining the estimation performance of the intercept and the outcome variable, as well as the percentage correctly put to zero. How these measures are constructed, is explained in Appendix B.2.

C.1 Evaluation characteristics real data

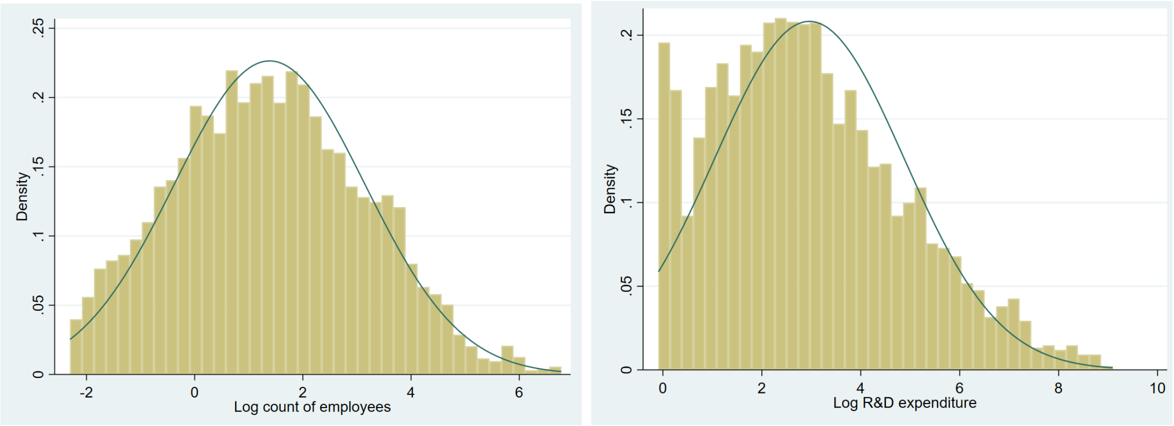
Section 4.1 discussed the characteristics and distributions for the generated variables. Here the evaluation process will be discussed.

Table 5 provides the normality measures in terms of the skewness and kurtosis for both variables to be generated via a distribution. As the skewness and kurtosis of a normal distribution are equal to 0 and 3, the values in the table can be compared. For both logarithmic variables the skewness and kurtosis come close to the assumed values.

Table 5: The skewness and kurtosis provided for the logarithmic number of employees and R&D expenditure.

log	Skewness	Kurtosis
Employees	0.169	2.529
R&D expenditure	0.553	2.799

Regarding the graph of the log number of employees in Figure 4a, it can be seen that it approximately follows a normal distribution. The distribution of the log R&D expenditure is a little less obvious, it seems to get close to a normal distribution except the left tail that is cut off. However for the simulation a normal distribution is assumed for simplicity. The simulation process is after all about showing the estimation performance of the method, not about recovering the results of Manresa (2016).



(a) Distribution of the log number of employees.

(b) Distribution of the log R&D expenditure.

Figure 4: Distributions of the logarithmic variables in the original data set.

C.2 Additional Double Pooled results

C.2.1 Different time horizons

Table 6 gives extra information on the prediction performance of the Double Pooled Lasso estimator and corresponds to Table 1. This case consists of five individuals with one relation each. The table shows that all real zero's are correctly put to zero by the Lasso regression. Furthermore, the prediction of the dependent variable starts of relatively bad and becomes better as T increases. The Mean Squared Error seems to converge towards one. Like the forecasting of the outcome variable, the prediction accuracy of the intercept becomes better as T increases.

Table 6: MSE of the intercept and the outcome variable, together with the percentage of zero's correctly put to zero for different time horizons, with $N = 5$.

T	MSE y	MSE int	% correct zero
100	1.481	0.358	100
200	1.135	0.114	100
500	1.037	0.012	100
1000	1.015	0.007	100
5000	1.003	0.001	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.2.2 Different group sizes

The results in Table 7 correspond to Table 2. The evaluation is extended to $N = 10$ and $N = 20$. For $N = 20$ the MSE's show convergence patterns, as well as the estimations of y and the intercept for $N = 10$. The values of the MSE for $N = 5$ regarding the outcome y seems to start fluctuating after $T = 400$. The relatively high MSE's in the upper right triangle of the sub tables can be attributed to the fact that not all zero's are correctly put to zero by the Lasso regression. A larger amount of wrongly estimated zero's leads to higher values for the MSE for both the intercept and outcome variable.

Table 7: MSE of the intercept and the outcome variable, together with the percentage of zero's correctly put to zero for different time horizons and different group sizes.

	MSE y			MSE int			% correct zero		
T	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
100	1.481	3.526	11.269	0.358	1.874	10.026	100	99.59	92.87
200	1.135	1.566	3.420	0.114	0.426	1.574	100	100	99.86
300	1.066	1.297	2.505	0.052	0.236	1.015	100	100	100
400	0.996	1.127	1.729	0.016	0.072	0.463	100	100	100
500	1.037	1.110	1.473	0.012	0.067	0.303	100	100	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.2.3 Different number of relations

The last variation in this simulation analysis will be assigned to the number of relations an individual has. Previously, the number of relations was set to one per individual. Now this number will be changed to see what the effect of having more relations is on the estimation performance of the Double Pooled Lasso estimator. The analysis on this matter includes $N = 5$ and $N = 10$. For $N = 10$, both $T = 100$ and $T = 200$ are considered, whereas for $N = 5$ only the former time horizon is evaluated.

In Table 8 the results for $N = 5$ can be obtained. The Frobenius norms for the different number of relations show that the more relations someone has, the more difficult it is to recover the real coefficients. This is in line with what is expected as the number of non-zero coefficients increases as the number of relations increase. Hence, more coefficients need to be estimated which brings more deviation from the real effects. In addition, the estimated common effect is approximately the same across the number of relations having only a little deviation. Therefore, it can be stated that the common effect is estimated consistently, regardless of the number of relations an individual has. Besides this, the MSE of the dependent variable improves as the number of relations increases, whereas the accuracy level of the intercept shows a fluctuation pattern first, after which larger MSE's occur. The zero's are again correctly estimated to be zero.

Table 8: The relevant statistics for N=5 and T=100 with varying number of relations.

relations	Fnorm	θ_diff	MSE y	MSE int	% correct zero
1	0.185	0.393	1.481	0.358	100
2	0.203	0.394	1.479	0.403	100
3	0.243	0.396	1.474	0.398	100
4	0.289	0.395	1.456	0.413	100
5	0.317	0.398	1.450	0.481	-

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

Table 9 provides the results for $N = 10$ with two different time periods. $T = 200$ is included, due to the % correct to zero results for $T = 100$. Not all zero's are correctly put to zero for $T = 100$; a relatively large amount of relations leads to worse estimations of the zero's. However, increasing T leads to cancelling out this phenomenon for all number of relations. Consequently, the values of the other measures shifted downwards after using $T = 200$. Hence, an increasing T does not only ensure consistency for one relation, it does for multiple as well. Regarding consistency, the Frobenius norm and θ_diff show that more observations lead to better estimates, whereas having more relations leads to less accurate estimates.

Table 9: The relevant statistics for N=10 using T=100 and T=200 with varying number of relations.

	Fnorm		θ_diff		MSE y		MSE int		% correct zero	
relations	T=100	T=200	T=100	T=200	T=100	T=200	T=100	T=200	T=100	T=200
1	0.441	0.214	0.910	0.427	3.526	1.566	1.874	0.426	99.59	100
3	0.631	0.273	0.915	0.428	3.487	1.558	1.983	0.471	98.65	100
5	0.797	0.340	0.916	0.429	3.413	1.546	2.299	0.486	97.69	100
7	0.907	0.385	0.920	0.432	3.378	1.541	2.908	0.545	96.12	100
9	1.001	0.429	0.924	0.433	3.340	1.533	2.994	0.636	96.25	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.3 Post Pooled results

A similar analysis as for the Double Pooled Lasso estimator is performed for the Post Pooled Lasso estimator. In fact, the same cases are evaluated except that there is no θ present for the Post Pooled Lasso estimator.

C.3.1 Different time horizons

The first case consists of one relation per individual, with the group size being equal to five.

Table 10 shows consistency of the Post Pooled estimator as the Frobenius norm decreases when the number of time periods increases. This means, with more available observations the differences between the real coefficient matrix and the estimated one, converges to zero. The prediction ability of the Post Pooled estimator regarding the intercept follows a similar structure as the Frobenius norm; more data leads to a better prediction. The dependent variable is predicted equally good regardless of the number of time periods. The Mean Squared Error of the outcome variable fluctuates between 0.99 and 1.006. The percentage correctly put to zero is 100 for all time periods.

Table 10: The Frobenius norm, MSE of the intercept and the outcome variable, and the percentage of zero's correctly put to zero for different time horizons, with $N = 5$.

T	Fnorm	MSE y	MSE int	% correct zero
100	0.154	0.991	0.047	100
200	0.137	0.991	0.028	100
500	0.065	1.006	0.008	100
1000	0.046	1.001	0.004	100
5000	0.023	1.004	0.001	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.3.2 Different group sizes

The next step is to compare the estimation performance for different number of individuals. The number of relations per person is still equal to one. For all number of individuals considered in this simulation analysis, the Post Pooled Lasso estimator shows convergence to zero and thus consistency. The Frobenius norm increases in the number of individuals for all T as Table 11 shows. In line with the results in Table 10, the estimation performance for the outcome y is approximately the same for all time periods. In addition, not much difference can be found between a different sample size in predicting the outcome variable. A similar conclusion can be drawn for the MSE of the intercept; over time, the estimation performance increases whereas the group sizes do not show large deviations from each other. Again, all actual zero's are estimated to be zero.

Table 11: The Frobenius norm, MSE of the intercepts and the outcome variables and the percentage of zero's correctly put to zero for different time horizons and different group sizes.

	Fnorm			MSE y			MSE int			% correct zero		
T	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20	N=5	N=10	N=20
100	0.154	0.220	0.315	0.991	0.950	0.987	0.047	0.038	0.041	100	100	100
200	0.137	0.158	0.230	0.991	0.990	0.995	0.028	0.024	0.021	100	100	100
300	0.096	0.114	0.199	0.994	1.013	0.996	0.014	0.011	0.018	100	100	100
400	0.075	0.107	0.147	0.991	1.006	0.994	0.010	0.011	0.009	100	100	100
500	0.065	0.091	0.144	1.006	0.999	0.988	0.008	0.006	0.009	100	100	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.3.3 Different number of relations

Lastly, the number of relations per individual will be changed. The explanation on this variant is similar as the one for the Double Pooled Lasso estimator. N is equal to either five or ten, while the number of relations differ per N. For the former situation, T is equal to 100, whereas for the latter $T = 200$ is considered as well.

Table 12 reveals that the estimation accuracy becomes weaker in the number of relations an individual has. This can be explained by the fact that more coefficients are non-zero. Estimation will not likely recover the true parameters. Therefore, the more estimates, the more deviation present from the real coefficients. Contrary to the estimation performance, is the fact that the validity of prediction of the outcome variable increases as the number of relations increases. The reliability of the estimated intercept (MSE int) decreases as more observations are included. Similar to the previous cases of the Post Pooled Lasso estimator, all zero's are correctly estimated to be zero.

Table 12: The relevant statistics for N=5 and T=100 with varying number of relations.

relations	Fnorm	MSE y	MSE int	% correct zero
1	0.154	0.991	0.047	100
2	0.191	0.983	0.067	100
3	0.219	0.975	0.081	100
4	0.233	0.970	0.105	100
5	0.254	0.964	0.121	-

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

The results for $N = 10$ are displayed in Table 13. The reason to consider two T's for this case can be found in the last two columns. $T = 100$ and a higher number of relations causes that the estimator includes relations that are not really there or leaves out relations that are present. This problem is thereafter solved by taking T higher, due to which the percentage of correct estimated zero's goes back to 100.

The Frobenius norm and the MSE's lead to the same conclusions as for $N = 5$. More relations lead to less accurate estimates and predictions of the intercept but are more precise in predicting the outcome variable. As stated in the passage above, as the number of relations increases, the ability to recover the true zero's becomes worse. However, considering more information in terms of data points solves this issue.

Table 13: The relevant statistics for N=10 using T=100 and T=200 with varying number of relations.

	Fnorm		MSE y		MSE int		% correct zero	
relations	T=100	T=200	T=100	T=200	T=100	T=200	T=100	T=200
1	0.220	0.158	0.950	0.990	0.038	0.024	100	100
3	0.304	0.218	0.935	0.982	0.057	0.038	100	100
5	0.394	0.255	0.914	0.976	0.106	0.064	99.61	100
7	0.450	0.300	0.899	0.967	0.137	0.072	98.10	100
9	0.503	0.342	0.885	0.958	0.164	0.093	97.92	100

The estimation is executed via 6 Monte Carlo simulations to avoid noise.

C.4 Robustness check

In the foregoing subsections, the penalty term λ is calculated using N rather than N^2, which is also proven to be accepted (Manresa (2016)). Here, a small discussion of the robustness of the latter way to calculate the penalty term λ.

The Post Pooled Lasso estimator does not discriminate between the two calculation methods of λ, i.e. calculation of the penalty parameter with either N or N^2 do not differ in estimation performance. The same yields for the Double Pooled Lasso estimator; the usage of the two different methods to calculate the penalty parameter does lead to the same results.

D Additional application results

Here the calculation procedure of the percentage in the relation matrices is discussed, where after the results are provided for λ's different from the ones in the main text, to show how more/less strict parameters influence the results.

D.1 Map of Uganda

Figure D.1 shows how the regions in Uganda are located within the country.[21] The North and West are relatively large compared to Central Uganda and the East. Noteworthy is that Kampala is located within Central Uganda. However, the data set considered this region to be standing alone. Therefore it will be considered as separate region. Besides, considering Kampala as separate region will give insight on how the capital city affects other regions in Uganda.

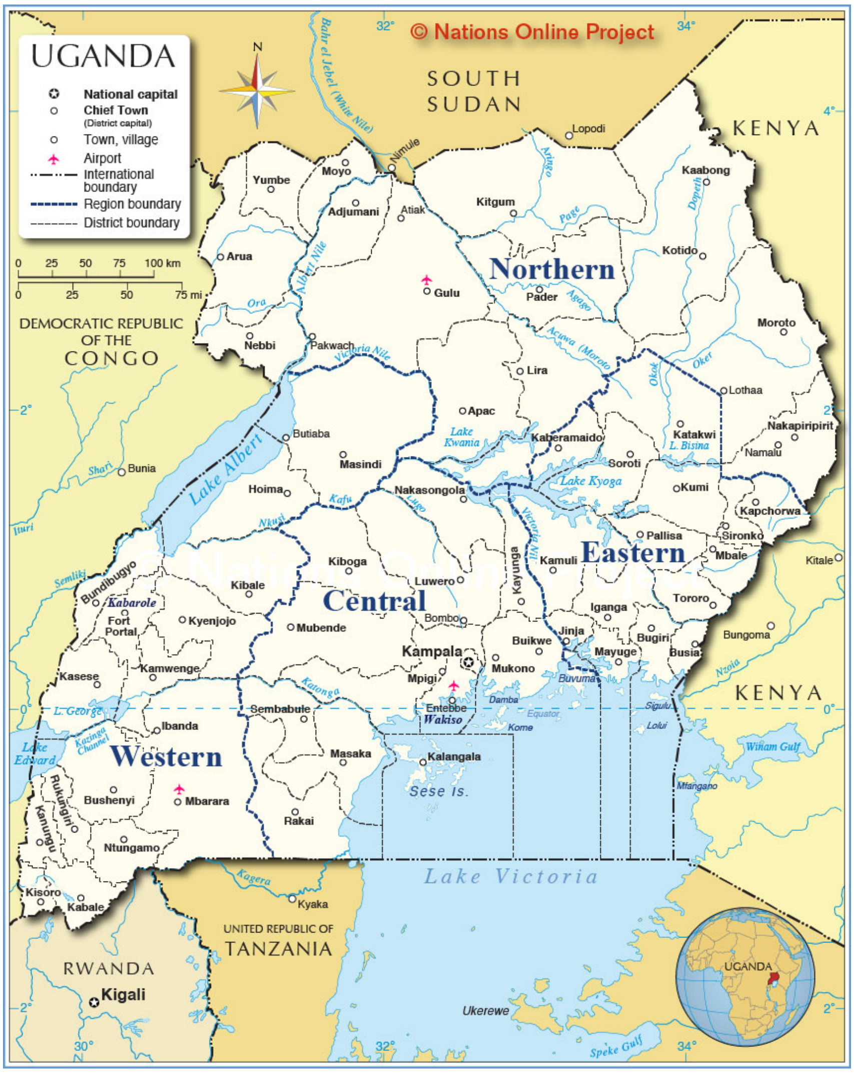


Figure 5: Caption

[21] The picture shown here is taken from this website: `https://www.nationsonline.org/oneworld/map/uganda-administrative-map.htm`.

D.2 Calculation percentage

To make the comparison between Kampala and the other regions more easy and interpretable, the actual number of relations is replaced by the percentage estimated relations of the total possible relations. For the off diagonal percentages, this is achieved by dividing the number of estimated relations by the total number of relations possible between the two regions. For the diagonal percentages, it is a little more complicated, as the individual effects need to be taken out. To achieve this, the number of households in that region is subtracted from the total number of possible relations. Subsequently, the number of estimated relations is divided by this. To give a mathematically visualisation of the calculation, see (14).

$$\text{percentage} = \frac{\#\text{estimated relations}}{\text{total possible relations}(-\text{number of households in the region})} * 100\%. \tag{14}$$

For the total percentages per row and column of each region, a similar calculation as (14) can be performed. For each row or column, the number of estimated relations is summed. Similarly, the total possible relations per row or column are summed to get the denominator. Again, for the diagonal element of the relation matrix, the individual effect is taken out by subtracting the number of households in that particular region. Please be advised that per row or column, taking the average will not result in the required percentage: taking the average will result in the average percentage of relations that are estimated to be non-zero for one region compared to all regions. This number is not the same when dividing the sum of estimated non-zero's by the sum of possible relations.

Here an explicit example for taking out the individual effects: for Kampala, the total number of relations is $25 \times 25 = 625$. However, the individual effects, β_i's are still included as they are the same as γ_{ii}. Therefore the total number of relations in Kampala without the individual effects is $625 - 25 = 600$.

To give a more clear overview of how the total percentages are calculated, the total number of relations without considering the individual effects are provided by Table 14.

Table 14: The total number of possible relations without considering the individual effects for each region, i.e. the denominator in (14).

Region	Total number of possible relations
Kampala	13350
Central	66216
East	77430
North	71556
West	57138

D.3 Non-randomized application results

As mentioned in Section 6, the order of the columns is randomized to account for the high multicollinearity that is present in the data. Below the results without the randomization are shown.

D.3.1 $\lambda = \lambda^*/3$

Compared to the results obtained in Section 6, the private and average social effect switched sign as they are estimated to be negative. This implies that if a family experiences an household income loss, on average their food insecurity score goes down by 0.0012. This private effect is based on three non-zero individual effects, whereas there are just two with the randomization. Similarly for the average social effects; an income loss for a family results on average in a decrease in worries about food of 0.004. Both these effects are against previous findings of Okpala et al. (2021) and Loopstra and Tarasuk (2013). The estimated common effect did not change due to randomizing the columns; it is still highly significant with an effect of -0.16. In addition to the common effect, the estimated number of relations remained the same; 745. This shows that randomizing the columns does not result in a different number of relations in this case.

Table 15: The estimated private, social and common effect, using $\lambda = \lambda^*/3$. The standard error is in the parenthesis for the common effect.

	Estimate
Private effect	-0.00121
Average social effect	-0.00396
Common effect θ	-0.160 (0.0574)

Unlike the number of estimated relations, the structure of interactions changed drastically. Kampala shows a huge influence on the other regions compared to the rest, followed by Central Uganda. The East has only an influence on households in Central and North Uganda, whereas the North and West do not have any influence at all. Regarding the small discussion in Section 6, this is due to the high multicollinearity among the regressors. The variables of Kampala show up first and are selected first, whereas the similar variables of other regions are dropped. Therefore, Kampala shows high influence and the other regions little to none. In addition, the results here are not in line with statements made by other studies. For example, as already discussed using the results with the randomization, the North and Central Uganda are expected to have the most influence (Shively et al. (2012) and Leliveld et al. (2013)), whereas Kampala has an extraordinary influence here. An explanation for this ability to generate spillovers cannot be found, let alone a justification for why the North and West do not have any influence at all.

Table 16: Structure of interactions given as percentage of the total possible interactions between regions in Uganda, using $\lambda = \lambda^*/3$. The region by row is affected by the region by column.

region	Kampala	Central	East	North	West	Total
Kampala	3.333	0.258	0	0	0	0.210
Central	3.355	0.479	0.006	0	0	0.269
East	2.566	0.517	0	0	0	0.240
North	3.582	0.451	0.005	0	0	0.274
West	3.738	0.430	0	0	0	0.275
Total	3.273	0.462	0.003	0	0	

D.4 Different penalty terms

Results for penalty terms different from those discussed in Section 6 are provided and discussed here.

D.4.1 $\lambda = \lambda^*$

The results for using $\lambda = \lambda^*$ as penalty parameter to divide by T show that this penalty term is too restrictive. The private effect is estimated to be zero as no β_i's are selected. This also holds for the γ_{ij}'s, which are all put to zero. Consequently, the average social effect is estimated to be zero. Using the original penalty term results in no selection of any variable by Lasso regression. In contrast, an estimate for the common effect θ is provided. In this case, the estimated common effect is equal to -1.47. This means that if the representative of the household is employed, the food insecurity score goes down on average with 1.5. Due to the standard error of 0.12, the estimate of θ is highly significant. The reliability of this conclusion could be questioned due to the controversial penalty term and estimates of the income loss dummies.

D.4.2 $\lambda = \lambda^*/2$

Using a slightly stricter penalty parameter than in Section 6, $\lambda = \lambda^*/2$, a private effect of approximately zero is estimated. This means that if a household experiences an income loss, its FIES score does not change on average. This conclusion might be questionable as only one individual effect is estimated to be non-zero, which might be due to the stricter penalty term. The average social effect is -0.00027 which is based on 140 non-zero spillover effects. Therefore, if a household experiences a reduction of their income, the food insecurity of other households decreases on average by 0.0003. This is contradictory to the existing literature, as Okpala et al. (2021) states that an income loss leads on average to feeling more food insecure. The effect of being employed is estimated to be -1.088, which means that having a job leads on average to a 1.09 reduction in the food insecurity score. This result is again in line with the results found by Loopstra and Tarasuk (2013). Similar to $\lambda^*/3$, the estimated common effect θ is highly significant in this case with a p-value of 0.000.

Table 17: The estimated private, social and common effect, using $\lambda = \lambda^*/2$. The standard error is in the parenthesis for the common effect.

	Estimate
Private effect	≈ 0
Average social effect	-0.000269
Common effect θ	-1.088 (0.106)

The actual structure is similar to the case were $\lambda = \lambda^*/3$ holds, aside from the substantially shrinking magnitudes. This is according to the expectation when a more penalizing parameter is used, as the shrinkage bias acts harder. The most influencing parts of Uganda are Central and northern Uganda, which compared to the previous case, changed positions in who generates relatively the most spillovers. In addition, Kampala lost its ability to generate spillovers on itself, Central Uganda and the West, whereas the East and West lost it for the West and Kampala respectively. This could partially be due to the stricter penalty term, which might indicate that this λ is too restrictive and that the multicollinearity issue is not sufficiently solved by dividing λ^* by 2. Due to the shrinkage of the magnitudes, the differences between receiving percentages become relatively larger as the West and Central Uganda receive twice as much spillover effects compared to the East.

Table 18: Structure of interactions given as percentage of the total possible interactions between regions in Uganda, using $\lambda = \lambda^*/2$. The region by row is affected by the region by column.

region	Kampala	Central	East	North	West	Total
Kampala	0	0.065	0.055	0.030	0	0.037
Central	0	0.105	0.017	0.090	0.038	0.059
East	0.028	0.050	0.019	0.046	0.013	0.032
North	0.030	0.060	0.031	0.101	0.007	0.050
West	0	0.121	0	0.119	0.018	0.061
Total	0.015	0.080	0.019	0.084	0.018	

D.4.3 $\lambda = \lambda^*/4$

Dividing λ^* by 4 makes the penalty term even less strict compared to dividing by 3. This relaxed penalty term leads to similar conclusions on the estimated effects, although the magnitudes for the private and common effect are smaller. A larger impact is estimated for the average social effect compared to using $\lambda^*/3$. This might be attributed to the fact that more relations are identified (1138) in this case as well as non-zero individual effects (4), whereas for $\lambda^*/3$ the numbers are equal to 745 and 2, respectively. Lastly, the estimated common effect is not statistically significant for this penalty parameter, as it has a p-value of 0.521. All the conclusions are still in line with both Loopstra and Tarasuk (2013) and Okpala et al. (2021).

Table 19: The estimated private, social and common effect, using $\lambda = \lambda^*/4$. The standard error is in the parenthesis for the common effect.

	Estimate
Private effect	0.00304
Average social effect	0.00210
Common effect θ	-0.0219 (0.0342)

Compared to the structure of interactions in Section 6, the magnitudes are larger. The central part of Uganda is still the most influencing region, followed by the North. Different from the conclusions for $\lambda^*/3$ is that Central and North Uganda do not generate the same percentage of spillovers on each other anymore, with the North generating 0.006% more externalities on Central Uganda. In addition, Kampala left its status of receiving the least spillovers. Nevertheless, the differences in being influenced remain small.

Table 20: Structure of interactions given as percentage of the total possible interactions between regions in Uganda, using $\lambda = \lambda^*/4$. The region by row is affected by the region by column.

region	Kampala	Central	East	North	West	Total
Kampala	0.667	0.355	0.359	0.657	0.112	0.397
Central	0.161	0.656	0.206	0.566	0.173	0.391
East	0.331	0.495	0.302	0.556	0.193	0.390
North	0.448	0.560	0.273	0.623	0.167	0.414
West	0.112	0.784	0.129	0.586	0.150	0.399
Total	0.292	0.600	0.240	0.586	0.170	

E Policy model

Section 3.1.1 mentions that the average of the spillover effects can be used as policy parameter. An example of how this is applied in practice is given here.

Suppose that policy makers are looking into food insecurity reducing programs in Uganda. The objective to minimize in this case is the average FIES score of the population, as a higher score means being more insecure about food. In this case it would not be necessarily the best option to allocate the specific treatment to the households with the highest individual effect, β_i. This is due to possible spillover effects which is accounted for in this paper. A household with a relatively low individual effect but large spillover effects on others might result in an even lower mean FIES score than households with a large individual effect does. Hence, the change in the mean FIES score due to treatment is not only depending on the individual effects, but on spillover effects as well. Based on this information, the change in the mean FIES score when individual k takes the treatment can be written mathematically:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i(d_k = 1) - y_i(d_k = 0)) = \frac{1}{N}\beta_k + \frac{1}{N}\sum_{i=1}^{N}\gamma_{ik}. \tag{15}$$

The left side of (15) shows the difference in the outcome variable (FIES score in this case) of household i if household k receives the treatment ($d_k = 1$) or not ($d_k = 0$). The effect of the treatment on the FIES score of household k (β_k) as well as the capacity of household k to generate spillover effects on the FIES scores of other households (γ_{ik}) determine the change in the mean FIES score. Using the provided information, the minimization problem can be constructed as follows:

$$\min_{(D_1,\dots,D_N)\in\{0,1\}^N} \frac{1}{N}\sum_{k=1}^{N}\left(\beta_k + \sum_{i=1}^{N}\gamma_{ik}\right)\cdot D_k$$

$$\text{subject to } \sum_{k=1}^{N} D_k = C.$$

In this optimization problem $D_1, ..., D_N$ are the treatment dummies for each household. The total costs available is captured by C, which simultaneously determines how many household can receive treatment as the costs per household are constant and unitary.