Networked Interactions, Graphical Models and Econometrics
Perspectives in Data Analysis

by
Jean-Baptiste Seby

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2020
© Massachusetts Institute of Technology 2020. All rights reserved.
Networked Interactions, Graphical Models and Econometrics

Perspectives in Data Analysis

by

Jean-Baptiste Seby

Submitted to the Institute for Data, Systems, and Society
on August 24, 2020, in partial fulfillment of the
requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science

Abstract

This thesis is composed of two independent parts.

In Part I, we study higher-order interactions in both graphical models and networks, i.e., interactions between more than two nodes. In the graphical model setting, we do not assume that interactions are known and our goal is to recover the structure of the graph. Our main contribution is an algebraic criterion that enables us to determine whether a set of observed variables have a single cause or multiple causes. We also prove that this criterion holds in the presence of confounders, i.e., when the causes are hidden. In the network setting, we assume that the structure of the graph is known. Our objective is then to identify what kind of information about data can be learned from the analysis of higher-order interactions. More precisely, using the generalization of the normalized Laplacian and random walks on graphs to simplicial complexes, we study a simplicial notion of PageRank centrality as defined in [Schaub et al., 2018]. Conducting numerical experiments on both synthetic and true data, we find evidence that the so-called edge PageRank is related to the concepts of local and global bridges in networks.

In Part II, we analyze the determinants of yield gaps in Semi-Arid Tropics (SAT) regions in India. Analyzing a panel data of households within 30 villages over 6 years in India, we apply a fixed effects estimation method and a quantile regression with fixed effects to identify the most significant explanatory variables of yield gaps for 5 different crops. Using a correlated random effects estimator for unbalanced panel data, we can also estimate coefficients for time-invariant variables. We find that yield gaps determinants are crop specific. In addition to that, soil characteristics show the most significant effects on output rate. When statistically significant, correlations with the type of soil are negative. This result might suggest that the choice of cropping pattern is not necessarily appropriate. Finally, results suggest that unobservable heterogeneity of households
is critical in explaining farm productivity. Time-invariant variables hardly explain this heterogeneity for which more research is needed.

Thesis Supervisor: Ali Jadbabaie  
Title: JR East Professor of Engineering, Institute for Data, Systems, and Society  
Department of Civil and Environmental Engineering

Thesis Supervisor: Chintan Vaishnav  
Title: Senior Lecturer,  
Sloan School of Management

Thesis Reader: John Tsitsiklis  
Title: Clarence J. Lebel Professor of Electrical Engineering and Computer Science
Acknowledgments

I would like to thank my advisors, Professor Ali Jadbabaie and Professor John Tsitsiklis. This work would not have been possible without their help and insightful guidance. It was very exciting to work on algebraic topology and its applications to data analysis and I thank Professor Jadbabaie for proposing this research topic. I also thank him for his time and his comments during our weekly meetings, that always suggested a new path when I was stuck. I was very privileged to be part of his research group for the last year and I also learnt a lot from the other people – students, post doc, guest speakers – in the group. In particular, I thank Arnab Sarker, with whom I worked on the edge PageRank project. All the opportunities that Professor Jadbabaie proposed me, such as teaching and presentations made my research a very rich experience for which I am very grateful to him.

I thank a lot Professor Tsitsiklis for accepting to read my thesis, for his time and for his always relevant comments throughout this year. Climbing the stairs up to the 7th floor of LIDS – I almost never take elevators – was always both an exciting and daunting exercise, but I always left our meeting with ideas to explore and valuable comments to pursue my work. Thanks a lot for your guidance throughout this year and your careful review of my thesis.

I would like to thank Professor Michael Schaub that gave me a lot of advice, guidance and comments on this work and many more topics! Thank you very much for your time Michael, your patience and explanations. Chapters IV, V and VI are mostly an extension of your work and I owe you a lot on this matter. Thank you for being a great mentor to me and for being a source of inspiration when doing research.

I would like to thank Professor Elina Robeva by whom I was advised for the Chapter III of this thesis. It was really great to work with you as part of the course Graphical Models. This chapter would not have been possible without your dedication (you spent so much time on Skype with me trying to solve this problem), insightful comments and advice. I also really enjoyed to present this work as part of a conference in Germany and I thank you a lot for this opportunity.

I am very grateful to Professor Chintan Vaishnav by whom I was advised for the project on yield gaps in Semi-Arid Tropics in India. I will never be enough thankful to Chintan for his advice, leadership and guidance not only on research but also on various topics as well. It was very exciting to work with you on this topic that is closed to my heart, namely farming and in particular farming in developing countries. I also thank Professor Chandra Madramootoo, people at the Tata Center and ICRISAT (International Crop Research Institute for Semi-Arid Tropic) for giving me this great research and human opportunity.

Nothing would have been possible without the precious support of the Technology and Policy Program. In particular, I thank Professor Frank Field for his patience, advice and attention. I thank Barb and Ed whose doors are always open and that actively help to
make TPP a warm community. I thank also Barb and Ed for our discussion, their advice and their willingness to help. I thank Professor Noelle Selin for her attentive leadership.

I enjoyed a lot my time at MIT and I would need pages and pages to thank all people and friends that I met, discussed, laughed and shared. I will not give up to this temptation, but those that I am talking about know that I am thinking about them.
Contents

1 Analysis of higher-order interactions in graphs: a graphical model and network science perspective 19

1 Introduction 21
   1.1 Motivation ................................................. 21
   1.2 Methodology and contributions ......................... 24
   1.3 Organization ............................................ 24

2 Introduction to graphical models 27
   2.1 Definition of a graphical model ......................... 27
   2.2 Conditional independence and $d$-separation ........ 28
   2.3 Markov property and faithfulness ...................... 29
      2.3.1 Undirected case .................................. 29
      2.3.2 Directed case .................................. 30
   2.4 Results on Gaussian graphical models ............... 32
   2.5 Conclusion and statement of the problem ............ 35

3 Multi-trek Separation in Linear Structural Equation Models 37
   3.1 Introduction ............................................. 37
   3.2 Background ............................................. 40
      3.2.1 Linear structural equation models .............. 40
      3.2.2 Cumulants of linear structural equation models 42
   3.3 Multi-trek separation for directed acyclic graphs .... 44
      3.3.1 Multi-treks and the multi-trek rule ............. 44
3.3.2 Multi-trek systems and determinants of higher-order cumulants . . 46
3.3.3 Main result . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
3.4 Hidden variables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
3.5 Determinants of higher-order moments and multi-trek systems . . . . . . 53
3.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

4 Simplicial complexes as a modeling paradigm for networks 59
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
4.2 Simplicial complexes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
4.3 Spaces of chains and co-chains . . . . . . . . . . . . . . . . . . . . . . . . . . 61
4.4 Boundary and co-boundary operators . . . . . . . . . . . . . . . . . . . . . . 61
4.5 Graph Laplacian and Hodge Laplacian . . . . . . . . . . . . . . . . . . . . . 63
  4.5.1 Graph Laplacian . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
  4.5.2 Hodge Laplacians . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
4.6 Hodge decomposition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66

5 Random walks on simplicial complexes and edge PageRank 67
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
5.2 Diffusion processes on graphs . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
5.3 Diffusion processes on simplicial complexes . . . . . . . . . . . . . . . . . . 69
  5.3.1 Lifting of an edge-flow in a higher-dimensional space . . . . . . . . . 69
  5.3.2 Lifting of a linear matrix operator . . . . . . . . . . . . . . . . . . . . . . 70
  5.3.3 Lifting of Hodge Laplacians . . . . . . . . . . . . . . . . . . . . . . . . . 71
  5.3.4 Normalized Hodge decomposition . . . . . . . . . . . . . . . . . . . . . . 74
5.4 Edge PageRank . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
  5.4.1 Node PageRank on graphs . . . . . . . . . . . . . . . . . . . . . . . . . . 76
  5.4.2 PageRank for k-simplices . . . . . . . . . . . . . . . . . . . . . . . . . . . 77

6 Applications and empirical results 79
6.1 Application of edge PageRank to synthetic data . . . . . . . . . . . . . . . . 79
6.2 Application of edge PageRank to real datasets . . . . . . . . . . . . . . . . . 87
10 Data description

10.1 Introduction .................................................. 133
10.2 Demographics .................................................. 134
  10.2.1 Occupation .............................................. 134
  10.2.2 Education ................................................ 137
10.3 Farming production ......................................... 139
  10.3.1 Land area .................................................. 139
  10.3.2 Number of plots ........................................ 140
  10.3.3 Percentage irrigation .................................. 141
  10.3.4 Output rate for all population – Kharif Cotton . 142
  10.3.5 Output rate by landholding category ............... 143
10.4 Revenue ....................................................... 144
  10.4.1 Farm profit ............................................... 144
  10.4.2 Total revenue .......................................... 145
10.5 Finance ....................................................... 145
  10.5.1 Debt for all population ............................... 145
  10.5.2 Debt by category of landholding .................... 146

11 Econometric analysis of yield gap in Semi-Arid Tropics India 149

11.1 Introduction .................................................. 149
11.2 Methodology .................................................. 150
11.3 Hypotheses .................................................... 157
11.4 Data description ............................................. 158
11.5 Results ........................................................ 159
11.6 Discussion .................................................... 165

12 The dynamics of farm productivity 185

12.1 Introduction .................................................. 185
12.2 Literature review ............................................ 185
12.3 Methodology .................................................. 188
  12.3.1 Summary statistics of output rate and change in output rate . . . . 190
12.3.2 Including land rental value as fixed assets ........................................ 191
12.3.3 Without fixed assets ................................................................. 195
12.4 Results ......................................................................................... 199
  12.4.1 Change in output rate .............................................................. 199
  12.4.2 Growth in TFP - Fixed assets included in the productivity function . 202
  12.4.3 Growth in TFP - Without fixed assets ....................................... 214

13 Conclusion ..................................................................................... 225
  13.1 Conclusions from the second part ................................................. 225
  13.2 High-level interpretation of the challenges around agriculture in India . 228
  13.3 Limitations ................................................................................. 229
  13.4 Directions for further research ..................................................... 230
List of Figures

1-1 Graph representation of two models of co-authorship ........................................ 22
1-2 Multiple confounding factors Shaded nodes are hidden variables .................... 22
1-3 Single confounding cause Shaded node is a hidden variable ............................. 23
1-4 Mixed graph .................................................................................................. 23

2-1 \(d\)-separation ................................................................................................. 32
2-2 Graph corresponding to Example 20 .................................................................. 34
2-3 Two models of causal structure to explain the observed phenotypes \(color,\) \(height,\) \(leaves\ texture\) .............................................................. 36
2-4 Mixed graph corresponding to both Model 1 (Figure 2-3a) and Model 2 (Figure 2-3b) .................................................................................................................. 36

3-2 A model with a single cause (Figure 3-2a) vs. a model with multiple causes (Figure 3-2b) .................................................................................................................. 40
3-3 Example of a 4-trek (Figure 3-3a) and a DAG (Figure 3-3b) ......................... 45
3-4 Trek-system with a sided intersection ............................................................ 47
3-5 Figure illustrating Example 32 ......................................................................... 48
3-6 Model with a single cause (Figure 3-6a) vs. model with multiple causes (Figure 3-6b) .................................................................................................................. 49
3-7 Example showing that the reverse implication of Corollary 39 does not hold 51
3-8 A multidirected edge between \(i_1, i_2, i_3, i_4\) of order 4. .................................... 51
3-9 Figures for Example 43 .................................................................................. 52
3-10 A model with a single hidden cause (Figure 3-10a) vs. a model with multiple hidden causes (Figure 3-10b) ................................................................. 53
Part I

Analysis of higher-order interactions in graphs: a graphical model and network science perspective
Chapter 1

Introduction

1.1 Motivation

Most models to date have focused on networks that involve only pairwise interactions among agents in a system. However, in a number of cases there are arguably interactions that take place between multiple agents (groups). This is the case for instance for the co-authorship network, the congressional co-sponsoring network, or protein interaction networks.

Many methods have been developed in graph theory and network science in order to analyze pairwise interactions. Yet, these methods usually do not account for higher-order interactions.

One example illustrating why this is a concern is the following. Let $A$, $B$ and $C$ be three authors in the co-authorship network and consider two cases. In the first one, $A$ writes a paper with $B$, $B$ writes a paper with $C$ and $C$ writes a paper with $A$. In the second case, $A$, $B$ and $C$ write a paper all together. In graph theory, both cases would be represented by the same graph depicted in Figure 1-1.

By representing two different models by the same graph, we are running the risk to lose some information. This observation is the initial point of departure of this work and raises the following high-level question:

Can we learn additional information from higher-order interactions?
In order to address this question, we adopt two different perspectives.

First, we consider the scenario where relationships between entities in the system are not known. In this setting, the dependencies (or by contrast conditional independencies) between entities can be represented by a graphical model and the objective is to learn the structure of this graph based on observations, i.e., data on the nodes. Causal structure discovery methods have also to account for variables in the graph that are not observed. In this case, we talk about latent variables.

So far in the graphical models literature [Drton and Maathuis, 2017a], algorithms for learning graph structure in the presence of latent variables only consider pairwise interactions. In other words, a latent variable is the cause of two observed variables as in the graph in Figure 1-2 where \(g_1\) causes 1 and 2, \(g_2\) causes 1 and 3, \(g_3\) causes 2 and 3.

However, we may also consider a model where all variables have a single hidden cause as in Figure 1-3. We could interpret this model as a higher-order interaction between 1, 2 and 3 due to the presence of the hidden variable \(g\).
In the graphical model framework, both models (multiple causes vs. one single cause) would be represented by the same *mixed* graph, represented in Figure 1-4.

![Figure 1-3: Single confounding cause](image)

**Figure 1-3: Single confounding cause**  
Shaded node is a hidden variable

![Figure 1-4: Mixed graph](image)

**Figure 1-4: Mixed graph**

Based on this observation, the question that we address in this context is:

**Question 1:**

*Can we distinguish graph structures with a single cause from graph structures with multiple causes?*

Second, we consider the case where the relationships between entities in the network are known. In order to model higher-order interactions between entities, we use simplicial complexes. Simplicial complexes are a type of hypergraphs that represent polyadic interactions by adding faces in the graph. We analyze this type of models using some results from algebraic topology. The question that we address is the following:

**Question 2:**

*Does the modeling of networks by simplicial complexes provide insights about data that cannot be obtained by modeling a network as a graph?*
1.2 Methodology and contributions

Both Question 1 and Question 2 have the same initial motivation (i.e. higher-order interactions) but differ by their point of view. The first one asks whether we can discover higher-order interactions simply based on node data and by the analysis of conditional independence statements. The second one asks whether or not higher-order interactions provide additional information, assuming you know the interactions.

They also differ by the approach that we adopt. In the first case, our contribution is purely theoretical. In the second case, our approach is mostly empirical.

As for Question 1, our main contribution is an algebraic criterion that enables us to distinguish a model with a single cause from a model with multiple causes. More precisely, the vanishing of subdeterminants of the tensor of higher cumulants indicate the presence of a common cause in the graph. We prove that this criterion also holds in the case of hidden variables, i.e., when the common cause of multiple variables is not observed.

As for Question 2, we model higher-order interactions using simplicial complexes. We focus on the concept of diffusion on simplicial complexes using the framework developed in the seminal paper by [Schaub et al., 2018]. In particular, Schaub et al. propose a generalization of the node PageRank centrality measure to edges. Conducting many numerical experiments on synthetic and true data, we find evidence that the edge PageRank centrality measure is closely related to the concepts of global and local bridges, thus enabling us to connect this notion of centrality to well-defined sociological concepts.

1.3 Organization

The content of this part is divided as follows. Chapters 2 and 3 focus on the graphical model perspective (Question 1). Chapters 4, 5, and 6 adopt a network science perspective (Question 2).

In Chapter 2, we give an introduction to graphical models and provide the main concepts that support the results in Chapter 3. In Chapter 3, we introduce a criterion that
enables us to distinguish a model with multiple interactions from a model with pairwise interactions only. We show that this result also applies to the case of latent variables. In Chapter 4, we introduce simplicial complex models. In Chapter 5, we state the main results from algebraic topology that we use to analyze data with higher-order interactions. In Chapter 6, we present the empirical results that we obtained by using the methods presented in Chapter 5. In Chapter 7, we conclude and provide directions for future research.
Chapter 2

Introduction to graphical models

Graphical models provide a mathematical formalism to define conditional independence relationships between random variables. These random variables are represented as nodes in a graph and two random variables that are connected by an edge in the graph are not independent. More refined relationships between random variables are encoded via conditional independence statements. The objective of the field of graphical models is to recover (or infer) the structure that connects these random variables. In this chapter, we present the necessary background on graphical models in order to understand the results presented in Chapter 3. We start with the definition of a graphical model. We then present the notion of conditional independence and its translation into the graph structure via $d$-separations. We then introduce the concept of Markov property and faithfulness and we conclude with results on Gaussian graphical models.

In this chapter, we provide the reader with the building blocks that are used in Chapter 3. We also note that this chapter is inspired from the lecture notes of the class 6.244 taught by Prof. Caroline Uhler and Prof. Elina Robeva during the Spring 2019.

2.1 Definition of a graphical model

A graph $G = (V, E)$ consists of a set of vertices $V$ and a set of edges $E$. A graph can be directed or undirected. In an undirected graph, $i$ is adjacent to $j$ if $(i, j) \in E$. In a directed graph, $j$ is said to be a parent of $i$ and $i$ is said to a child of $j$ if there is a directed edge $j \rightarrow i$. 
A node \(j\) is said to be an ancestor of \(i\) if there exists a path \(j \rightarrow \cdots \rightarrow i\). A directed graph is called a DAG (Directed Acyclic Graph) if there is no cycle. The skeleton of a directed graph \(G\) is the corresponding undirected graph once we have removed the direction of the edges.

### 2.2 Conditional independence and \(d\)-separation

We associate to each node \(i\) in the graph a random variable \(X_i\), taking values in a set \(\mathcal{X}_i\). Let \(X\) be the random vector whose \(i\)-th entry is \(X_i\). Then \(X\) takes values in \(\mathcal{X} = \prod_{i \in V} \mathcal{X}_i\).

Let \(f(x)\) be a continuous density with respect to a product measure \(\mu\) on \(\mathcal{X}\).

We introduce the marginal probability distribution of a set of vertices.

**Definition 1.** Let \(A \subset V\), let \(x_A \in \mathcal{X}_A\) and \(x_{V/A} \in \mathcal{X}_{V/A}\). Then the marginal probability distribution of the set of vertices \(A\) is

\[
f_A(x_A) = \int f(x_A, x_{V \setminus A}) d\mu(x_{V \setminus A})
\]

**Remark 2.** \(x_A \in \mathcal{X}_A\) (resp. \(x_{V/A} \in \mathcal{X}_{V/A}\)) is a random vector whose coordinates are random variables associated with each node \(i\) in the set of vertices \(A\) (resp. \(V/A\)).

We introduce the conditional probability distribution of a set of vertices \(A\) given another set of vertices \(B\).

**Definition 3.** Let \(A, B \subset V\) be two disjoint sets of vertices. Let \(x_B \in \mathcal{X}_B\) with \(f(x_B) > 0\). Then

\[
f_{A|B}(x_A|x_B) = \frac{f_{A\cup B}(x_A, x_B)}{f_B(x_B)}
\]

We define the notion of conditional independence.

**Definition 4.** Let \(A, B, C \subset V\) be pairwise disjoint sets of vertices. If for all \(x_A \in \mathcal{X}_A\), \(x_B \in \mathcal{X}_B\) and \(x_C \in \mathcal{X}_C\) with \(f(x_C) > 0\), we have

\[
f_{A\cup B|C}(x_A, x_B|x_C) = f_{A|C}(x_A|x_C)f_{B|C}(x_B|x_C)
\]
then the conditional statement $X_A \perp X_B | X_C$ holds.

"$X_A$ is independent of $X_B$ given $X_C$" means that if we know $X_C$, then we do not learn more information on $X_A$ by knowing $X_B$. From this interpretation, it results that

$$f_{A|B\cup C}(x_A|x_B, x_C) = \frac{f_{A\cup B|C}(x_A, x_B|x_C)}{f_{B|C}(x_B|x_C)} = f_{A|C}(x_A|x_C).$$

(2.4)

Conditional independence is a notion that refers to the probability distribution of the random variables. If we consider a graph, we can also define some separation statements on the vertices. The next section develop the notion of separation in the graph and makes the connection with the conditional independence statements.

### 2.3 Markov property and faithfulness

#### 2.3.1 Undirected case

For an undirected graph, we say that $C \subset V$ separates $A \subset V$ from $B \subset V$ if every path from a node $i \in A$ to a node $j \in B$ contains a node $k \in C$. In particular, if there is no edge between two nodes $i$ and $j$, then $i$ and $j$ are separated by $V \setminus \{i, j\}$.

The Markov property makes the connection between the separations in the graph and the conditional independences.

**Definition 5.** Let $G = (V, E)$ be an undirected graph. We say that a probability distribution on $V$ satisfies the global Markov property with respect to $G$ if for all disjoint sets $A, B, C \subset V$ such that $C$ separated $A$ from $B$, we have $X_A \perp X_B | X_C$.

**Definition 6.** Let $G = (V, E)$ be an undirected graph. We say that a probability distribution on $V$ satisfies the pairwise Markov property with respect to $G$ if for all pairs of nodes $i, j \in V$ such that $i$ and $j$ are separated by $V \setminus \{i, j\}$, we have $X_i \perp X_j | X_{V\setminus{i,j}}$.

**Remark 7.** Notice that the global Markov property implies the local Markov property, but the reverse does not hold.
The Markov property guarantees that separations in the graph are translated into conditional independences in the distribution. We are also interested in knowing whether all conditional independences in the distribution are reflected in the graph. This property is called faithfulness.

**Definition 8.** We say that a probability distribution is faithful to a graph if every conditional independence \( X_A \indep X_B \mid X_C \) implies that \( C \) separates \( A \) and \( B \) in the graph.

We now extend the notion of separation and the Markov properties to the case of directed graphs.

### 2.3.2 Directed case

The notion of separation in a directed graph is less straightforward. We first need to introduce the concept of the moral graph.

**Definition 9.** For a DAG \( G = (V, E) \), the corresponding moral graph is the undirected graph \( G^m = (V, E^m) \) such that

- there is an edge between \( u \) and \( v \) in \( G^m \) if \( u \to v \) or \( v \to u \)

- there is an edge between \( u \) and \( v \) if there exists a node \( w \) such that \( u \to w \) and \( v \to w \)

**Remark 10.** The word "moral" comes from the fact that we "marry" parent nodes that have a child node in common, i.e., we remove all the "immoralities".

**Definition 11.** An ancestral set \( A \subset V \) is a set that contains all its ancestors.

We are now ready to define the Markov properties in the case of directed graphs.

**Definition 12.** Let \( G = (V, E) \) be a DAG. We say that a distribution on \( V \) satisfies the (directed) global Markov property if for any ancestral set \( A \subset V \), it satisfies the global Markov property with respect to the moral graph \( G^m_A \), where \( G_A \) is the graph induced by the nodes in \( A \).
**Definition 13.** Let $G = (V, E)$ be a DAG. Let us denote by $nd(i)$ the non-descendants of $i$ and $pa(i)$ the parents of $i$. We say that a distribution on $V$ satisfies the (directed) local Markov property for all $i \in V$ if the conditional independence statements $X_i \perp X_{nd(i)} \mid X_{pa(i)}$ are satisfied.

We can also define the notion of separation in an directed graph without referring to the undirected case (i.e., without going through the moral graph). It was introduced by Judea Pearl and Tom Verma ([Pearl, 1988a, Pearl, 1988b, Verma and Pearl, 1988, Verma and Pearl, 1990]).

**Definition 14.** Let $G = (V, E)$ be a DAG. Two nodes $u, v \in V$ are $d$–separated in $G$ given a subset $S \subseteq V \setminus \{u, v\}$ if no undirected path between $u$ and $v$ is active. An undirected path between $u$ and $v$ is active if very triple $(u,v,w)$ of consecutive nodes in the path satisfies

- $u \rightarrow v \rightarrow w$ and $v \notin S$, or
- $u \leftarrow v \leftarrow w$ and $v \notin S$, or
- $u \leftarrow v \rightarrow w$ and $v \notin S$, or
- $u \rightarrow v \leftarrow w$ and $v \in S$ or one of $v$’s descendants is in $S$

We call a collider a node that is the children of two other nodes (i.e. $v$ is a collider if $u \rightarrow v \leftarrow w$). We can now reformulate the notion of $d$-separation as follows: $S \subseteq V \setminus \{u, v\}$ $d$-separates $u$ from $v$ if every undirected path between $u$ and $v$ either

- has a non-collider of $u$ and $v$ in $S$, or
- has a collider that is not in $S \cup an(S)$.

We depict in Figure 2-1 the definition of $d$-separation.

**Remark 15.** The global Markov property provides a unique characterization of an undirected graph. However, this is no longer the case for directed graphs: two directed graphs can have the same Markov properties. We say in this case that they share the same Markov equivalence class. This implies that we cannot distinguish them only using conditional independence statements.
The following lemma gives a criterion according to which two DAGs are Markov equivalent.

**Lemma 16.** Two DAGs are Markov equivalent if and only if they have the same skeleton (i.e., the same underlying undirected graph) and the same unshielded colliders (i.e., graph structure of the type $u \rightarrow v \leftarrow w$).

**Remark 17.** A collider is a subgraph of the form $u \rightarrow v \leftarrow w$. It is said to be **unshielded** if there is no edge between $u$ and $w$.

To conclude this chapter, we introduce the main results that enables us to learn the structure of a graph when random variables are Gaussian. Chapter 3 will provide a method to learn the structure of a graph in the non-Gaussian setting.

### 2.4 Results on Gaussian graphical models

We start by summarizing some results about Gaussian distributions. Let $X \sim \mathcal{N}_m(\mu, \Sigma)$ and let $A, B \subset \{1, \ldots, m\}$ be disjoint sets of vertices.

- The marginal distribution of a Gaussian distribution is a Gaussian distribution, and

  $$X_A \sim \mathcal{N}_{|A|}(\mu_A, \Sigma_{AA})$$  \hspace{1cm} (2.5)

- The conditional of a Gaussian distribution is a Gaussian distribution, and

  $$X_A | X_B = x_B \sim \mathcal{N}_{|A|}(\mu_A + \Sigma_{AB} \Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})$$  \hspace{1cm} (2.6)
There is an equivalence between the independence of two Gaussian random variables and the vanishing of their covariance, i.e.,

\[ X_A \perp X_B \iff \Sigma_{AB} = 0 \quad (2.7) \]

A conditional independence in the distribution is equivalent to a condition on the rank of a submatrix of \( \Sigma \). A more formal statement is given in Proposition 18.

**Proposition 18.** Let \( X \sim \mathcal{N}_m(\mu, \Sigma) \), \( A, B, C \subset \{1, \ldots, m\} \) pairwise disjoint. Assuming that the covariance matrix \( \Sigma \) is non-singular, then

\[ X_A \perp X_B \mid X_C \iff \text{rank}(\Sigma_{A\cup C, B\cup C}) = |C| \quad (2.8) \]

In particular, if \( A = \{i\} \), \( B = \{j\} \), then a subdeterminant of \( \Sigma \) vanishes

\[ i \perp j \mid C \iff \text{rank}(\Sigma_{iC, jC}) = |C| \iff \text{det}(\Sigma_{iC, jC}) = 0 \quad (2.9) \]

**Remark 19.** The notation \( \Sigma_{iC, jC} \) means that we consider the submatrix of \( \Sigma \) whose rows are indexed by \( i \) and \( C \) and whose columns are indexed by \( j \) and \( C \).

We present the proof in a similar way as in [Lauritzen, 1996] and in the 6.244 Lecture notes.

**Proof.** The conditional independence \( X_A \perp X_B \mid X_C \) is equivalent to the vanishing of the submatrix indexed by \( (A, B) \) of the conditional covariance \( \Sigma_{A\cup B\mid C} \), i.e.,

\[ X_A \perp X_B \mid X_C \iff \text{Cov}(X_{A\cup B\mid X_C})_{A, B} = 0 \quad (2.10) \]

\[ \iff (\Sigma_{A\cup B, A\cup B} - \Sigma_{A\cup B, C}\Sigma_{C, C}^{-1}\Sigma_{C, A\cup B})_{A, B} = 0 \quad (2.11) \]

\[ \iff \Sigma_{A, B} - \Sigma_{A, C}\Sigma_{C, C}^{-1}\Sigma_{C, B} = 0 \quad (2.12) \]

The term \( \Sigma_{A, B} - \Sigma_{A, C}\Sigma_{C, C}^{-1}\Sigma_{C, B} \) is the Schur complement of the block \( \Sigma_{C, C} \) of the matrix
Using Guttman’s rank additivity formula, we have

\[
\text{rank}(\Sigma_{A \cup C,B \cup C}) = \text{rank}(\Sigma_{C,C}) + \text{rank}(\Sigma_{A,B} - \Sigma_{A,C}\Sigma_{C,C}^{-1}\Sigma_{C,B}).
\]

(2.13)

The covariance matrix \(\Sigma\) is non-singular and the matrix \(\Sigma_{C,C}\) has full rank. From this we deduce that

\[
X_A \perp X_B | X_C \iff \Sigma_{A,B} - \Sigma_{A,C}\Sigma_{C,C}^{-1}\Sigma_{C,B} = 0
\]

(2.14)

\[
\iff \text{rank}(\Sigma_{A,B} - \Sigma_{A,C}\Sigma_{C,C}^{-1}\Sigma_{C,B}) = 0
\]

(2.15)

\[
\iff \text{rank}(\Sigma_{A \cup C,B \cup C}) = |C|
\]

(2.16)

We illustrate Proposition 18 with the following example.

**Example 20.** Let us suppose that we observe data generated by an unknown model that we want to recover. Let us assume that the unknown model is given by the following equations.

\[
X_1 = \varepsilon_1 \\
X_2 = 3X_1 + \varepsilon_2 \\
X_3 = 4X_1 + \varepsilon_3
\]

(*)

where \(\varepsilon_1, \varepsilon_2\) and \(\varepsilon_3\) are i.i.d Gaussian random variables with mean 0 and variance 1. We can see that this model induces the conditional independence statement \(X_2 \perp X_3 | X_1\). This corresponds to the \(d\)-separation "1 \(d\)-separates 2 from 3" and this corresponds to the following graph

\[\begin{align*}
X_1 & \\
X_2 & \\
X_3 & 
\end{align*}\]

Figure 2-2: Graph corresponding to Example 20
If we were only given observations, we could actually apply Proposition 18 and learn the conditional independence statements induced by the model. In order to illustrate this, we first generate 10,000 samples from the model. We get the following covariance matrix (rounded to the closest integer):

\[
\Sigma = \begin{bmatrix}
1 & 3 & 4 \\
3 & 10 & 12 \\
4 & 12 & 17
\end{bmatrix}
\]

As expected from Proposition 18, we get

\[
|\Sigma_{12,13}| = 1 \times 12 - 4 \times 3 = 0
\]

Notice that the precision matrix \( K = \Sigma^{-1} \) is equal to

\[
K = \begin{bmatrix}
26 & -3 & -4 \\
-3 & 1 & 0 \\
-4 & 0 & 1
\end{bmatrix}
\]

The conditional independence \( X_1 \perp X_2 | X_1 \) also translates into \( K_{23} = K_{32} = 0 \).

### 2.5 Conclusion and statement of the problem

In this chapter, we have introduced the formalism used in graphical models. In this framework, we aim at representing a probability distribution by a graph. The nodes of this graph are therefore associated to random variables. We use conditional independence statements between these random variables in order to recover the structure of the graph. Such representation is very appealing when the graph is directed and without cycles (DAG) because it encodes causal relationships between random variables. When the random variables are Gaussian, the vanishing of some subdeterminants of the covariance
matrix enables us to learn the causal structure of a DAG. However, existing algorithms for causal structure discovery do not go beyond pairwise interactions. To understand this, let us consider the following example. Suppose that we observe three characteristics of a plant: its color, its height and the texture of its leaves. We consider two models. In Model 1, the observed phenotypes result from the expression of a single gene (Figure 2-3a). In Model 2, the observed phenotypes stem from the expression of three genes, each controlling for a pair of observed variables. (Figure 2-3b).

Figure 2-3: Two models of causal structure to explain the observed phenotypes color, height, leaves texture.
Shaded nodes represent confounders (i.e., latent variables).

In the graphical models literature, both models would be represented by the graph in Figure 2-4

One natural question would be the following: is it possible to distinguish these two models? The objective of Chapter 3 is to develop a criterion that enables us to learn whether there is a single cause (i.e. higher order interactions in Model 1) or multiple causes (Model 2). We propose a criterion that succeeds in distinguishing Model 1 and Model 2 when random variables are non-Gaussian. Our result relies on higher-order cumulants and does not provide more insights into the case where the random variables are Gaussian. We are not aware of any work addressing the Gaussian setting.
Chapter 3

Multi-trek Separation in Linear Structural Equation Models

This chapter is based on a joint paper with Elina Robeva (ArXiv: 2001.10426).

3.1 Introduction

Although randomized experiments are the most commonly used method for causal inference, they are sometimes not feasible for practical or ethical reasons. Because of these constraints, scientists often need to learn the structure of the graph underlying the relationships between variables based on purely observational data. Suppose that $G = (V, \mathcal{D})$ is a directed acyclic graph (or DAG) with vertex set $V = \{1, \ldots, p\}$ and edge set $\mathcal{D} \subseteq V \times V$. The graph $G$ gives rise to a linear structural equation model (LSEM), which consists of the joint distribution of a random vector $X = (X_1, \ldots, X_p)$ in which the variable $X_i$ associated to vertex $i \in V$ is a linear function of $X_j$, where $j$ varies over the parent set $\text{pa}(i)$ of $i$ (i.e., all vertices $j \in V$ such that $j \rightarrow i \in \mathcal{D}$), and a noise term $\epsilon_i$,

$$X_i = \sum_{j \in \text{pa}(i)} \lambda_{ji}X_j + \epsilon_i, \ i \in V. \quad (3.1)$$

If no hidden variables are present, we assume that the noise terms $\epsilon_i$ are mutually independent. To encode the presence of hidden variables, we allow dependencies between
the $\epsilon_i$ variables, and graphically we depict this via multi-directed edges (see Figure 3-1a and Definition 40). These encode a hidden common cause of a few of the observed variables. We represent this more complicated hidden structure via a mixed graph $G = (V, D, H)$, where $H$ is the set of multi-directed (hyper)edges (see Section 3.4).

When the noise terms $\epsilon_i$ are Gaussian, then so are the $X_i$ variables. In this setting, the linear structural equation model given by a graph $G$ corresponds to the set $\mathcal{M}^{(2)}(G)$ of covariance matrices of a Gaussian distribution consistent with the graph $G$ [Lauritzen, 1996]. This is precisely the set of covariance matrices that possess a certain parametrization arising from the structure of $G$. Furthermore, bidirected edges suffice to parametrize the model $\mathcal{M}^{(2)}(G)$ in the case of hidden variables. For example, the mixed graphs $G_1$ and $G_2$ in the figure above give rise to the same model in Zariski closure $\overline{\mathcal{M}^{(2)}(G_1)} = \overline{\mathcal{M}^{(2)}(G_2)}$.

When the variables are non-Gaussian, we can depict the model using covariances as well as higher-order moments/cumulants of the random vector $X$. We denote by $\mathcal{M}^{(k)}(G)$ the set of cumulants up to order $k$ consistent with the graph $G$. This set can also be parametrized using the graph (Definition 24), and provides a more refined description of the graph structure. For instance, the two graphs above give rise to different models $\overline{\mathcal{M}^{(3)}(G_1)} \neq \overline{\mathcal{M}^{(3)}(G_2)}$.

A parametrization of the model, however, may not always be sufficient. Statistical problems like model selection, model equivalence, and constraint based statistical inference often require an implicit description of the model in terms of (polynomial) equations which can be read off from the graph $G$, e.g., via a combinatorial criterion.

When $G$ is a DAG and the variables are Gaussian, the implicit description of the model $\mathcal{M}^{(2)}(G)$ is given by the vanishing of specific subdeterminants of the covariance matrix which can be read off from the graph via $d$-separation and the more general trek-separation criteria [Sullivant et al., 2010]. In fact, trek-separation helps describe the van-
ishing of all subdeterminants of the covariance matrix in any (not necessarily Gaussian) LSEM. It turns out that covariance information is only sufficient to identify the graph up to Markov equivalence. That is, if two graphs $G$ and $G'$ give rise to the same conditional independence relations, then they produce the same sets of covariance matrices $\mathcal{M}^{(2)}(G) = \mathcal{M}^{(2)}(G')$. Therefore, when the graph $G$ is a DAG and the variables are Gaussian, we can only recover $G$ up to Markov equivalence given observational data. Finding an implicit description of $\mathcal{M}^{(2)}(G)$ in the presence of hidden variables is more challenging, although there has been promising recent progress. In particular, [Yao and Evans, 2019] prove that the minimal generators for the vanishing ideal $\mathcal{I}(G)$ containing all the constraints for a Gaussian Acyclic Directed Mixed Graph $G$ are in one-to-one correspondence with the pairs of non-adjacent vertices in the graph, and provide an algorithm to find all these generators.

When the variables are non-Gaussian, the graph $G$ can be recovered uniquely from observational data. In particular, [Shimizu et al., 2006a] use Independent Component Analysis (ICA) [Comon, 1994] to estimate the graph structure via the Linear non-Gaussian Acyclic Model (LiNGAM). This framework and its derived versions DirectLiNGAM and PairwiseLiNGAM [Shimizu et al., 2006b] make it possible to distinguish graphs within Markov equivalence classes. Furthermore, [Wang and Drton, 2019] provide an algorithm that extends causal discovery of the causal structure in the high-dimensional setting based on higher-order moments, under a maximum in-degree condition.

In this chapter, we also work under the framework of a non-Gaussian LSEM. Building on the trek rule [Sullivant et al., 2010], we define the multi-trek rule (Proposition 27) which gives a polynomial parametrization of the higher-order moments/cumulants, and enables us to study LSEMs (3.1) via their higher-order cumulant representation $\mathcal{M}^{(k)}(G)$ from the perspective of algebraic statistics [Sullivant, 2018] which has so far only been used for Gaussian and discrete graphical models. By analogy with the vanishing of subdeterminants in the covariance matrix, we give a necessary and sufficient combinatorial criterion, called multi-trek separation, for the vanishing of subdeterminants of the tensor $\mathcal{C}^{(k)}$ of $k$-th order cumulants (Theorem 35), which extends to the hidden variable case (Theorem 45). Our multi-trek separation criterion, for example, enables us to identify the
presence of a hidden common cause of multiple observed variables.

(a) Single cause
\( C^{(3)}_{123} \neq 0 \)

(b) Multiple causes
\( C^{(3)}_{123} = 0 \)

Figure 3-2: A model with a single cause (Figure 3-2a) vs. a model with multiple causes (Figure 3-2b)

The rest of the chapter is organized as follows. In Section 3.2, we define linear structural equation models (LSEMs) and their cumulant tensors. In Section 3.3, we introduce the notion of a multi-trek and we state our main theorem for DAG models that establishes a combinatorial criterion for the vanishing of subdeterminants of the \( k^{th} \)-order cumulant tensor. In Section 3.4, we consider the case of hidden variables. Graphically we encode the presence of such variables via multi-directed edges, and we show that our results generalize to this case. In Section 3.5, we conjecture that our multi-trek criterion is also equivalent to the vanishing of subdeterminants of higher-order moment (rather than cumulant) tensors. In section 3.6, we conclude and discuss directions for further research.

3.2 Background

In this section we provide the necessary background on linear structural equation models, and their higher-order cumulants.

3.2.1 Linear structural equation models

Let \( G = (V, D) \) be a directed acyclic graph (DAG) with finite vertex set \( V = \{1, \ldots, p\} \) and edge set \( D \subseteq V \times V \). Here acyclic means that there are no directed cycles, i.e., no sequences of the form \( i_0 \rightarrow i_1 \rightarrow \cdots \rightarrow i_s = i_0 \), where \( i_j \rightarrow i_{j+1} \in D \). The edge set is always assumed to be free of self-loops, so \( (i, i) \notin D \) for all \( i \in V \). For each vertex \( i \), define its set of parents as \( \text{pa}(i) = \{ j \in V : (j, i) \in D \} \). The graph \( G \) induces a statistical model,
called a *linear structural equation model*, for the joint distribution of a collection of random
variables \((X_i, i \in V)\), indexed by the graph’s vertices. The model hypothesizes that each
variable is a linear function of the parent variables and a noise term \(\varepsilon_i\):

\[
X_i = \lambda_{0i} + \sum_{j \in \text{pa}(i)} \lambda_{ji} X_j + \varepsilon_i, \ i \in V. \tag{3.2}
\]

The \(\varepsilon_i\) variables for \(i \in V\), are independent and centered. The coefficients \(\lambda_{0i}\) and \(\lambda_{ji}\)
are unknown real parameters that are assumed to be such that the system (3.2) admits
a unique solution \(X = (X_i : i \in V)\). Typically termed a system of structural equations,
(3.2) specifies cause-effect relations whose straightforward interpretability explains the
wide-spread use of the models [Spirtes et al., 2000, Pearl, 2009a].

The random vector \(X\) that solves the system (3.2) may have an arbitrary mean depending
on the choice of parameters \(\lambda_{0i}\). Since the mean can easily be learned from data, and
we are mainly concerned with learning the underlying graph structure, we disregard the
offsets \(\lambda_{0i}\), and the system (3.2) becomes

\[
X = \Lambda^T X + \varepsilon. \tag{3.3}
\]

Here \(\Lambda = (\lambda_{ij}) \in \mathbb{R}^D\), and \(\mathbb{R}^D\) is the set of \(V \times V\) matrices \(\Lambda\) with support \(\mathcal{D}\),

\[
\mathbb{R}^D = \{ \Lambda \in \mathbb{R}^{V \times V} : \lambda_{ij} = 0 \text{ if } i \rightarrow j \notin \mathcal{D} \}.
\]

Since \(G\) is acyclic, the matrix \(I - \Lambda\) is always invertible and the solution of the system
(3.3) is:

\[
X = (I - \Lambda)^{-T} \varepsilon. \tag{3.4}
\]
### 3.2.2 Cumulants of linear structural equation models

Recall that the \( k \)-th cumulant tensor of a random vector \( Z = (Z_1, \ldots, Z_p) \) is the \( p \times \cdots \times p \) (\( k \) times) table with entry at position \( (i_1, \ldots, i_k) \) given by

\[
\text{cum}(Z_{i_1}, \ldots, Z_{i_k}) = \sum_{(A_1, \ldots, A_L)} (-1)^{L-1}(L-1)! \ E \left[ \prod_{j \in A_1} Z_j \right] E \left[ \prod_{j \in A_2} Z_j \right] \cdots E \left[ \prod_{j \in A_L} Z_j \right],
\]

where the sum is taken over all partitions \( (A_1, \ldots, A_L) \) of the set \( \{i_1, \ldots, i_k\} \). If each of the variables \( Z_i \) is centered, i.e., has mean 0, then we can restrict to summing over partitions for which each \( A_i \) has size at least 2. For example, the first four cumulants are given as follows:

\[
\text{cum}(Z_i) = E[Z_i] = 0, \quad \text{cum}(Z_{i_1}, Z_{i_2}) = E[Z_{i_1}Z_{i_2}], \quad \text{cum}(Z_{i_1}, Z_{i_2}, Z_{i_3}) = E[Z_{i_1}Z_{i_2}Z_{i_3}],
\]

\[
\text{cum}(Z_{i_1}, Z_{i_2}, Z_{i_3}, Z_{i_4}) = E[Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}] - E[Z_{i_1}Z_{i_2}]E[Z_{i_3}Z_{i_4}] - E[Z_{i_1}Z_{i_3}]E[Z_{i_2}Z_{i_4}]
\]

\[
- E[Z_{i_1}Z_{i_4}]E[Z_{i_2}Z_{i_3}]
\]

Now, let \( k \geq 2 \), and let \( \mathcal{E}^{(k)} \) and \( \mathcal{C}^{(k)} \) be the \( k \)-th order cumulant tensors of the random vectors \( \epsilon \) and \( X \), respectively. The linear structural equation model, and, particularly, the expression (3.4) yield the following relationship between \( \mathcal{C}^{(k)} \) and \( \mathcal{E}^{(k)} \).

**Lemma 21** ([Comon and Jutten, 2010, Chapter 5]). The tensor \( \mathcal{C}^{(k)} \) of \( k \)-th order cumulants of \( X \) equals

\[
\mathcal{C}^{(k)} = \mathcal{E}^{(k)} \bullet (I - \Lambda)^{-k}, \quad (3.5)
\]

where \( \mathcal{E}^{(k)} \bullet (I - \Lambda)^{-k} \) denotes the Tucker product of the order-\( k \) tensor \( \mathcal{E}^{(k)} \) and the matrix \( (I - \Lambda)^{-1} \) along each of its \( k \) dimensions. In other words,

\[
\left( \mathcal{E}^{(k)} \bullet (I - \Lambda)^{-k} \right)_{i_1, \ldots, i_k} = \sum_{j_1, \ldots, j_k} \mathcal{E}^{(k)}_{j_1, \ldots, j_k} ((I - \Lambda)^{-1})_{j_1, i_1} \cdots ((I - \Lambda)^{-1})_{j_k, i_k}.
\]

One of the reasons this factorization is extremely useful is that when the entries of the
noise vector $\epsilon$ are mutually independent, its cumulants $\mathcal{E}^{(k)}$ are diagonal tensors.

**Lemma 22** ([Comon and Jutten, 2010, Chapter 5]). *If the variables $Z_1, \ldots, Z_p$ are independent, then the $k$-th order cumulant tensor of $Z = (Z_1, \ldots, Z_p)$ is diagonal, i.e., the entry at position $(i_1, \ldots, i_k)$ is 0 unless $i_1, \ldots, i_k$ are all equal.*

**Remark 23.** (a) We originally considered the moments of $X$ and $\epsilon$, instead of their cumulants. Lemma 21 also applies to the factorization of the moments, however, the moment tensors of $\epsilon$ are no longer diagonal, i.e., Lemma 22 does not apply. We give further details of this study in Section 3.6.

(b) Lemma 22 is the reason cumulants are widely used for Independent Component Analysis (ICA) [Comon and Jutten, 2010, de Lathauwer, 2006]. Indeed, finding the CP-decomposition of the cumulant tensor $\mathcal{C}^{(k)}$ can recover the matrix $(I - \Lambda)^{-1}$. For an extended study of cumulants in ICA, we refer the reader to [Comon and Jutten, 2010].

Lemmas 21 and 22 provide a means to parametrize the set of all cumulant tensors of the distributions in a given graphical model.

**Definition 24.** Let $G = (V, D)$ be a DAG, and let $k \geq 2$ be an integer. The set

$$
\mathcal{M}^{(2)}(G) = \{(I - \Lambda)^{-T} \mathcal{E}^{(2)} (I - \Lambda)^{-1} : \Lambda \in \mathbb{R}^D, \mathcal{E}^{(2)} \succeq 0 \text{ diagonal}\}
$$

consists of all covariance matrices of distributions in the graphical model given by $G$. For $k \geq 3$, define

$$
\mathcal{M}^{(k)}(G) = \{\mathcal{E}^{(2)} \bullet (I - \Lambda)^{-2}, \ldots, \mathcal{E}^{(k)} \bullet (I - \Lambda)^{-k} : \Lambda \in \mathbb{R}^d, \mathcal{E}^{(i)} \text{ diagonal for } 2 \leq i \leq k, \mathcal{E}^{(2)} \succeq 0\}.
$$

The set $\mathcal{M}^{(k)}(G)$ contains all cumulants up to order $k$ of a distribution in the graphical model given by $G$.

When the error terms $\epsilon_i$ are Gaussian, the random vector $X$ also follows a Gaussian distribution and all of its cumulants of order $k \geq 3$ equal 0. Therefore, the model equals $\mathcal{M}^{(2)}(G)$. In this case, different DAGs $G_1$ and $G_2$ that lie in the same Markov equivalence class can give rise to the same models $\mathcal{M}^{(2)}(G_1) = \mathcal{M}^{(2)}(G_2)$. An implicit description of
\( \mathcal{M}^{(2)}(G) \) is well-known when \( G \) is a DAG and is given by conditional independences [Harri Kiiveri, 1984; Luis David Garcia, 2005]. These correspond to \( d \)-separations in the graph \( G \) [Pearl, 1986].

When the noise terms \( \varepsilon_i \) are not Gaussian, the higher-order moments will not necessarily vanish, and we can obtain more information about the distribution from them. It turns out that \( \mathcal{M}^{(k)}(G) \) uniquely identifies the DAG \( G \) [Shimizu et al., 2006a; Shimizu et al., 2006b; Wang and Drton, 2019]. An implicit description of \( \mathcal{M}^{(k)}(G) \) is not known completely, although [Wang and Drton, 2019] discover enough of the defining equations to identify the DAG \( G \).

### 3.3 Multi-trek separation for directed acyclic graphs

In this section we present our main result, a particular type of constraint on the model \( \mathcal{M}^{(k)}(G) \) that corresponds to a combinatorial criterion in the graph \( G \). Given data, one can check whether the constraint holds for the sample cumulant tensors in order to obtain information about the structure of the unknown DAG \( G \). Our result generalizes to the case of hidden variables as shown in Section 3.4.

#### 3.3.1 Multi-treks and the multi-trek rule

We begin by generalizing the notion of a trek [Sullivant et al., 2010].

**Definition 25.** A \( k \)-trek in a DAG \( G \) between \( k \) nodes \( v_1, \ldots, v_k \) is an ordered collection of \( k \) directed paths \( (P_1, \ldots, P_k) \), where \( P_i \) has sink \( v_i \), and \( P_1, \ldots, P_k \) have the same source vertex, called the top of the \( k \)-trek and denoted by \( \text{top}(P_1, \ldots, P_k) \).

Note that a 2-trek is exactly the same as the usual notion of a trek [Sullivant et al., 2010].

**Example 26.** In the directed acyclic graph from Figure 3-3b, (2 → 6, 2 → 8) is a 2-trek between 6 and 8 with top 2;
We now generalize the trek rule, which originates in the work of [Wright, 1921, Wright, 1934] and relies on the observation that $(I - \Lambda)^{-1} = I + \Lambda + \Lambda^2 + \cdots$. Since the graph $G$ is acyclic, this formal sum is finite, and each of the entries of $(I - \Lambda)^{-1}$ can be expressed as

$$
(I - \Lambda)^{-1}_{ji} = \sum_{P \in P(j,i)} \lambda_P
$$

(3.6)

where $P(j,i)$ is the set of all directed paths from $j$ to $i$, and $\lambda_P$ is the product of the coefficients $\lambda_{uv}$ along the edges $u \rightarrow v$ on a directed path $P$ (e.g., if $P$ is the path $1 \rightarrow 2 \rightarrow 4$, then $\lambda_P$ is $\lambda_{12}\lambda_{24}$).

**Proposition 27** (The multi-trek rule). When the entries of the noise vector $\varepsilon$ are independent, the entries of the $k$-th order cumulant tensor $\mathcal{C}^{(k)}$ of $X$ can be expressed as a sum over $k$-trek monomials,

$$
\mathcal{C}^{(k)}_{i_1,\ldots,i_k} = \sum_{(P_1,\ldots,P_k) \in T(i_1,\ldots,i_k)} \mathcal{E}^{(k)}_{\top(P_1,\ldots,P_k),\ldots,\top(P_1,\ldots,P_k)} \lambda_{P_1} \cdots \lambda_{P_k},
$$

(3.7)

where $T(i_1,\ldots,i_k)$ is the set of all $k$-treks between $i_1,\ldots,i_k$.

**Proof.** By Lemma 21, we can express the $k$-th cumulant tensor of $X$ via the Tucker decomposition $\mathcal{C}^{(k)} = \mathcal{E}^{(k)} \cdot (I - \Lambda)^{-k}$. Furthermore, by Lemma 22, the $k$-th cumulant tensor $\mathcal{E}^{(k)}$
of \( \epsilon \) is diagonal. Therefore,

\[
C^{(k)}_{i_1, \ldots, i_k} = \sum_{j_1, \ldots, j_k} E^{(k)}_{j_1, \ldots, j_k} ((I - \Lambda)^{-1})_{j_1, i_1} \cdots ((I - \Lambda)^{-1})_{j_k, i_k}
\]

\[
= \sum_{j} E^{(k)}_{j, \ldots, j} ((I - \Lambda)^{-1})_{j, i_1} \cdots ((I - \Lambda)^{-1})_{j, i_k}.
\]

Using equation (3.6), we obtain the result. \( \square \)

**Example 28.** Consider the DAG from Figure 3-3bb. According to the multi-trek rule, we have, for instance, the following relationships between the cumulants \( C^{(k)} \) and \( E^{(k)} \) of \( X \) and \( \epsilon \).

\[
C^{(2)}_{4, 5} = E^{(2)}_{4, 4} \lambda_{45} + E^{(2)}_{1, 1} \lambda_{14} \lambda_{45}, \quad C^{(3)}_{5, 6, 7} = E^{(3)}_{1, 1, 1} \lambda_{14} \lambda_{45} \lambda_{16} \lambda_{17}, \quad C^{(3)}_{5, 6, 8} = 0.
\]

These follow because there are two 2-treks between 4 and 5: \((4, 4 \rightarrow 5)\) and \((1 \rightarrow 4, 1 \rightarrow 4 \rightarrow 5)\). There is one 3-trek between 5, 6, and 7: \((1 \rightarrow 4 \rightarrow 5, 1 \rightarrow 6, 1 \rightarrow 7)\). There are no tri-treks between 5, 6, and 8.

### 3.3.2 Multi-trek systems and determinants of higher-order cumulants

We now generalize the multi-trek rule, giving an expression of the determinants of sub-tensors of \( C^{(k)} \) in terms of multi-trek systems. We extend the notion of a matrix determinant to the case of tensors.

**Definition 29.** Let \( T \) by an order-\( k \) \( n \times \ldots \times n \) tensor. Then, its determinant is

\[
\det(T) = \sum_{\sigma_1, \ldots, \sigma_{k-1} \in \mathfrak{S}(n)} \text{sign}(\sigma_1) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} T_{i, \sigma_1(i), \ldots, \sigma_{k-1}(i)},
\]

where \( \mathfrak{S}(n) \) is the set of permutations of the set \( \{1, \ldots, n\} \).

Next, we introduce the notion of a multi-trek system.

**Definition 30.** Given a collection of \( k \) sets of nodes \( S_1, \ldots, S_k \subseteq V \) such that \( \#S_1 = \ldots = \#S_k = n \), a \( k \)-trek system \( T \) is a collection of \( n \) \( k \)-treks between \( S_1, \ldots, S_k \). We define the
top of this k-trek system, top(T), to be the union of the tops of the k-treks. We allow repeated elements in top(T). A k-trek system T has sided intersection if there exist two k-treks (P_1, \ldots, P_k) and (Q_1, \ldots, Q_k) in T and a number 1 \leq i \leq k so that the directed paths P_i and Q_i have a common vertex. We denote by \tilde{\mathcal{T}}(S_1, \ldots, S_k) the set of k-trek systems between S_1, \ldots, S_k that have no sided intersection.

**Example 31.** In Figure 4 below, the two 3-treks between S_1, S_2, S_3 have a sided intersection along the paths leading to the set S_1.

![Figure 3-4: Trek-system with a sided intersection](image)

Given a collection of k sets of nodes S_1, \ldots, S_k \subseteq V such that \#S_1 = \ldots = \#S_k = n, and an ordering of the nodes in each set S_i, a k-trek system T gives rise to a permutation of the nodes in each of S_2, \ldots, S_k (if we keep the ordering of S_1 fixed). The sign of T is the product of the signs of those (k − 1) permutations.

**Example 32.** In Figure 3-5 below, S_1 = \{v_1, v_2\}, S_2 = \{v_3, v_4\}, S_3 = \{v_5, v_6\}. Assume that the initial ordering is (v_1, v_2), (v_3, v_4), (v_5, v_6). The trek system in Figure 3-5a has two 3-treks, one between v_1, v_4, v_5 and one between v_2, v_3, v_6. Therefore, \text{sign}(T) = -1. The trek system in Figure 3-5b has two 3-treks, one between v_1, v_4, v_6 and one between v_2, v_3, v_5. Therefore, \text{sign}(T) = +1.

The determinant of the subtensor of k^{th} order cumulants indexed by the sets S_1, \ldots, S_k can be expressed in terms of the trek systems involving S_1, \ldots, S_k.

**Proposition 33.** Let S_1, \ldots, S_k \subseteq V be k sets of nodes such that \#S_1 = \ldots = \#S_k = n. Then,

\[
\det C_{S_1,\ldots,S_k}^{(k)} = \sum_{T \in \tilde{\mathcal{T}}(S_1,\ldots,S_k)} \text{sign}(T) m_T,
\]

(*

47
where $\mathcal{T}(S_1, \ldots, S_k)$ is the set of $k$-trek systems between $S_1, \ldots, S_k$, and $m_T$ is the trek-system monomial of the trek system $T = \{(P^{(1)}_1, \ldots, P^{(1)}_k), \ldots, (P^{(n)}_1, \ldots, P^{(n)}_k)\}$, defined as

$$m_T = \prod_{i=1}^{n} \mathcal{E}^{(k)}_{\text{top}(P^{(i)}_1, \ldots, P^{(i)}_k), \ldots, \text{top}(P^{(i)}_1, \ldots, P^{(i)}_k)} \prod_{j=1}^{k} \lambda^{p^{(j)}_i}.$$ 

In fact, the sum in $\star$ can be taken over treks $T \in \mathcal{T}(S_1, \ldots, S_k)$ without sided intersections, i.e.,

$$\det C^{(k)}_{S_1, \ldots, S_k} = \sum_{T \in \mathcal{T}(S_1, \ldots, S_k)} \text{sign}(T) m_T.$$  \hspace{1cm} (3.9)

**Example 34.** Consider, once again, the DAG from Figure 3-3bb. Then, we have that

$$\det C^{(2)}_{46,78} = \mathcal{E}_{1,1}^{(2)} \mathcal{E}_{2,2}^{(2)} \lambda_{14} \lambda_{17} \lambda_{26} \lambda_{28},$$

since there is only one 2-trek system between 46 and 78 without sided intersection, namely \{(1 → 4, 1 → 7), (2 → 6, 2 → 8)\}. Similarly, we have

$$\det C^{(3)}_{46,58,78} = \mathcal{E}_{1,1,1}^{(3)} \mathcal{E}_{2,2,2}^{(3)} \lambda_{14}^2 \lambda_{15} \lambda_{17} \lambda_{26} \lambda_{28}^2,$$

since there is only one 3-trek system between 46, 58, and 78, namely \{(1 → 4, 1 → 4 → 5, 1 → 7), (2 → 6, 2 → 8, 2 → 8)\}. 

48
3.3.3 Main result

Our main theorem shows that the non-existence of a trek system without sided intersection between $k$ sets of vertices is equivalent to the vanishing of a specific subdeterminant of the $k$-th cumulant tensor $C^{(k)}$.

**Theorem 35.** Let $G = (V, \mathcal{D})$ be a DAG, and let $S_1, ..., S_k$ be subsets of $V$ with $\# S_1 = ... = \# S_k$. Then,

$$\det C_{S_1, ..., S_k}^{(k)} = 0$$

for every $C^{(k)}$ from $\mathcal{M}^{(k)}(G)$ if and only if every system of $k$-treks between $S_1, ..., S_k$ has a sided intersection.

The proof of Theorem 35 can be found in Appendix 7.2. The proof relies on an application of the Cauchy-Binet formula for tensors. This formula is of independent mathematical interest and is proven in Appendix 7.2.

**Example 36.** Theorem 35 enables us to determine whether random variables have a common cause. Consider the graphs in Figures 3-6a and 3-6b below. Let $A = \{1\}$, $B = \{2\}$, and $C = \{3\}$. In Figure 3-6a there is one tri-trek joining $A$, $B$ and $C$, thus, $\det(C_{ABC}^{(3)}) = C_{123}^{(3)} \neq 0$. In Figure 3-6b there is no tri-trek joining $A$, $B$, and $C$ and, a fortiori, no tri-trek without sided intersection. Therefore, $\det(C_{ABC}^{(3)}) = 0$.

The seminal paper [Sullivant et al., 2010] shows that the vanishing of determinants of the covariance matrix of $X$ is equivalent to a 2-trek separation criterion in the graph $G$. In the rest of this section, we illustrate that a generalization of this criterion to the case $k > 2$ only works in one direction.
**Definition 37.** The collection of sets \((A_1, \ldots, A_k)\) \(k\)-trek-separates \(S_1, \ldots, S_k\) if for every \(k\)-trek with paths \((P_1, \ldots, P_k)\) between \(S_1, \ldots, S_k\), there exists \(j \in \{1, \ldots, k\}\) such that \(P_j\) contains a vertex from \(A_j\).

**Theorem 38 (Sullivant et al., 2010, Theorem 2.8).** The submatrix \(\Sigma_{A,B}\) has rank less than or equal to \(r\) for all covariance matrices consistent with the graph \(G\) if and only if there exist subsets \(C_A, C_B \subset V\) with \(#C_A + #C_B \leq r\) such that \((C_A, C_B)\) 2-trek-separates \(A\) from \(B\). Consequently,

\[
\text{rk}(\Sigma_{A,B}) \leq \min\{#C_A + #C_B : (C_A, C_B) 2\text{-trek-separates } A \text{ from } B\}
\]

and equality holds for generic covariance matrices in the model \(\mathcal{M}^{(2)}(G)\).

In Corollary 39, we show that when \(k \geq 3\), \(k\)-trek-separation implies the vanishing of the corresponding cumulant tensor determinant (but not necessarily vice-versa).

**Corollary 39.** Consider \(k\) sets of vertices \(S_1, \ldots, S_k\) with \(#S_1 = \ldots = #S_k = n\). For all tensors \(C^{(k)}\) of \(k\)th-order cumulants consistent with the graph \(G\), the subtensor \(C^{(k)}_{S_1, \ldots, S_k}\) has a null determinant if there exist subsets \(A_1, \ldots, A_k \subset V\) with \(#A_1 + \ldots + #A_k < n\) such that \((A_1, \ldots, A_k)\) \(k\)-trek-separates \(S_1, \ldots, S_k\).

**Proof.** Let us suppose that there exist \(A_1, \ldots, A_k\) such that \(#A_1 + \ldots + #A_k < n\) and \((A_1, \ldots, A_k)\) \(k\)-trek-separates \(S_1, \ldots, S_k\). Suppose there exists a \(k\)-trek system with no sided intersection \(T = (T_1, \ldots, T_n)\) between \(S_1, \ldots, S_k\). Then, for every \(i = 1, \ldots, n\), there exists \(m_i\) such that the \(m_i\)-th component of \(T_i\) intersects \(A_{m_i}\). Since \(#A_1 + \ldots + #A_k < n\), by the Pigeon-Hole Principle, there exist \(i \neq j\) such that \(m = m_i = m_j\), and the \(m\)-th components of \(T_i\) and \(T_j\) go through the same element \(s \in A_m\). Therefore, \(T\) has a sided intersection. Contradiction! Thus, by Theorem 35, \(\det C^{(k)}_{S_1, \ldots, S_k} = 0\).

When \(k \geq 3\), the reverse implication of Corollary 39 is not true, i.e., \(\det (C^{(k)}_{S_1, \ldots, S_k}) = 0\) does not imply that there exists \((A_1, \ldots, A_k)\) such that \(#A_1 + \ldots + #A_k < n\) and \((A_1, \ldots, A_k)\) \(k\)-trek-separates \((S_1, \ldots, S_k)\). Consider the graph in Figure 3-7 below. In this graph, \(#S_1 = #S_2 = #S_3 = 2\). There is no system of two 3-treks between \(S_1, S_2, S_3\) without sided intersection. However, it is not possible to find three sets \(A_1, A_2, A_3\) such that \(#A_1 + #A_2 + #A_3 <
2 = \#S_t, i.e., there is no such set that 3-trek-separates \(S_1, S_2, S_3\). Intuitively, this happens because "max flow" and "min cut" are no longer equal, while the reason the statement holds when \(k = 2\) is precisely because of the min-cut max flow (Menger’s) Theorem.

Figure 3-7: Example showing that the reverse implication of Corollary [39] does not hold

### 3.4 Hidden variables

We now consider the case where the structural equation model (3.3) involves some hidden (i.e., unobserved) variables. Alternatively, this is equivalent to allowing the noise variables \(\varepsilon_i\)'s to be correlated. We represent such a case with a mixed graph \(G = (V, \mathcal{D}, \mathcal{H})\), where \(V\) is the set of vertices, \(\mathcal{D} \subseteq V \times V\) is the set of directed edges, and \(\mathcal{H}\) is the set of multidirected edges (see Definition 40 below), which reflect the dependencies between the \(\varepsilon\) variables. We assume that \(G\) does not contain any cycle nor loop.

**Definition 40.** A multidirected edge between nodes \(i_1, \ldots, i_k\) is the union of \(k\) directed edges with the same source and with sinks \(i_1, \ldots, i_k\). The \(k\)-directed edges are merged at their source without an additional node. We call \(k\) the order of the multidirected edge.

Figure 3-8: A multidirected edge between \(i_1, i_2, i_3, i_4\) of order 4.

**Remark 41.** Note that this is a generalization of the notion of bidirected edges widely used in the literature.
As before, define $\mathcal{E}^{(k)}$ to be the tensor of $k$-th order cumulants of the variables $\varepsilon_i$'s. Since we do not assume the independence of the variables $\varepsilon_i$'s, then any entry of the tensor $\mathcal{E}^{(k)}$ may be non-zero. Specifically, $\mathcal{E}^{(k)}$ is a $V \times V \times \ldots \times V$ tensor for which the entry $w_{i_1,\ldots,i_k}$ is non-zero if there exists a multidirected edge between $a_1,\ldots,a_l$ such that $\{i_1,\ldots,i_k\} \subseteq \{a_1,\ldots,a_l\}$. Note that some of the elements $i_1,\ldots,i_k$ could be equal.

We now define the notion of a $k$-trek in this setting.

**Definition 42.** A $k$-trek between vertices $i_1,\ldots,i_k$ in a mixed graph $G$ is composed of $k$ directed paths $(P_1,\ldots,P_k)$ where $P_s$ goes from $j_s$ to $i_s$, and the tops $j_1,\ldots,j_k$ are connected in one of the following ways.

(a) Either the vertices $j_1,\ldots,j_k$ coincide to form the top of the trek; or

(b) There exist an $l$-directed edge between $a_1,\ldots,a_l$ such that $\{j_1,\ldots,j_k\} \subseteq \{a_1,\ldots,a_l\}$.

**Example 43.** In the mixed graph from Figure 3-9a, $\langle 1 \to 7, 1 \to 6, 1 \to 4 \to 5 \rangle$ is a 3-trek between 7, 6 and 5. The tops of the paths going to the nodes 7, 6 and 5 coincide with node 1. In the mixed graph from Figure 3-9b, $\langle 7, 6, 4 \to 5 \rangle$ is a 3-trek between 7, 6 and 5. The tops of the paths are connected by a 3-directed edge between 4, 6, and 7.

The tensor of $k$-th order cumulants of a mixed graphical model is then given by the equation (21), $\mathcal{C}^{(k)} = \mathcal{E}^{(k)}(I - \Lambda)^{-k}$. In a similar way to Proposition 27, every entry of the tensor $\mathcal{C}^{(k)}$ can be expressed as a sum of $k$-trek monomials.

**Corollary 44.** For a noise vector $\varepsilon$ whose entries are dependent, the entries of the $k$-th order...
cumulant tensor $C^{(k)}$ of $X$ can be expressed as a sum over $k$-trek monomials,

$$C_{i_1,...,i_k}^{(k)} = \sum_{(P_1,...,P_k) \in T(i_1,...,i_k)} \mathcal{E}_{\text{top}(P_1),...,\text{top}(P_k)}^{(k)} \lambda^{P_1} ... \lambda^{P_k},$$

where $T(i_1,...,i_k)$ is the set of all $k$-treks between $i_1,...,i_k$.

In a similar fashion to the directed acyclic case, we obtain the following result.

**Theorem 45.** Consider a mixed graph $G(V, D, H)$. Let $S_1, ..., S_k$ be subsets of $V$ with $\#S_1 = ... = \#S_k$. Then,

$$\det C_{S_1,...,S_k}^{(k)} = 0$$

if and only if every system of $k$-treks between $S_1, ..., S_k$ has a sided intersection.

Theorem 45 is proven in Appendix 7.2

**Example 46.** The following example shows that Theorem 45 enables us to determine whether random variables have a common cause. In the two graphs from Figure 3-10:

![Figure 3-10](image)

Figure 3-10: A model with a single hidden cause (Figure 3-10a) vs. a model with multiple hidden causes (Figure 3-10b).

let $A = \{1\}$, $B = \{2\}$, and $C = \{3\}$. In Figure 3-10a there is one tri-trek joining $A$, $B$, and $C$, hence $\det(C_{ABC}^{(3)}) = C_{123}^{(3)} \neq 0$. In Figure 3-10b there is no tri-trek joining $A, B$ and $C$ and, a fortiori, no tri-trek without sided intersection. Therefore $\det(C_{ABC}^{(3)}) = 0$.

### 3.5 Determinants of higher-order moments and multi-trek systems

At the start of this project, we focused on tensors of higher-order moments rather than cumulants. However, contrary to cumulant tensors (cf. Lemma 22), moment tensors of
order greater than 3 are not diagonal when the variables are independent. Extending Theorem 35 to higher-order moment tensors is therefore not straightforward. Nevertheless, we conjecture that the result still holds. Before stating this conjecture, we translate our results obtained for cumulant tensors to moment tensors.

Let \( \Phi^{(k)} \) be the tensor of \( k \)-th order moments of \( \epsilon \). Then, the \( k \)-th order moment tensor of \( X \) is given as follows.

**Proposition 47.** The tensor \( \mathcal{N}^{(k)} \) of \( k \)-th order moments of the random vector \( X \) with mean \((0, ..., 0)\) equals

\[
\mathcal{N}^{(k)} = \Phi^{(k)} \cdot (I - \Lambda)^{-k}.
\]  

Let \( G = (V, D, H) \) as in Section 3.4. In order to account for the non-zero off-diagonal entries in the tensor \( \Phi^{(k)} \) of higher-order moments of \( \epsilon \), we need to adapt our definition of a \( k \)-trek as follows.

**Definition 48.** A \( k \)-split-trek in \( G \) between \( k \) nodes \( v_1, \ldots, v_k \) is either:

(a) an ordered collection of \( k \) directed paths \((P_1, \ldots, P_k)\) where \( P_i \) has sink \( v_i \) and \( P_1, \ldots, P_k \) have the same source; or

(b) an ordered collection of \( k \) directed paths \((P_1, \ldots, P_k)\) where \( P_i \) has sink \( v_i \) and \( P_1, \ldots, P_k \) may have different sources, but each source must be shared by at least two paths.

![Figure 3-11: Types of \( k \)-split-treks for \( k = 4 \)](image)

**Definition 49.** The *top* of a \( k \)-split-trek \((P_1, \ldots, P_k)\) in \( G \) is either:
(a) a node that is the source of each of the $k$ paths $P_1, \ldots, P_k$.

(b) a set of $j$ nodes where each of the $j$ nodes is a source of $\ell_j$ of the paths $P_1, \ldots, P_k$ such that $\sum_j \ell_j = k$ and $\ell_j \geq 2$ for all $j$.

**Example 50.** In Figure 3-11, subfigure (a) illustrates the first type of a top, while subfigures (b-f) illustrate the second type.

Using equation (3.6), we rewrite the entries of $(I - \Lambda)^{-1}$ in equation (3.11) and thus express the entries of the tensor of $k^{th}$-order moments as follows.

**Proposition 51.** We have that

$$\mathcal{N}^{(k)}_{i_1, \ldots, i_k} = \sum_{(P_1, \ldots, P_k) \in \mathcal{S}(i_1, \ldots, i_k)} \Phi^{(k)}_{top(P_1, \ldots, P_k), \ldots, top(P_1, \ldots, P_k)} \lambda_{P_1} \cdots \lambda_{P_k}$$

(3.12)

where $\mathcal{S}(i_1, \ldots, i_k)$ is the set of all $k$-split-treks between $i_1, \ldots, i_k$.

Similarly to the case of higher-order cumulants, the determinant of the subtensor of $k^{th}$-order moments indexed by the sets $S_1, \ldots, S_k$ can be rewritten in terms of the split-trek systems between $S_1, \ldots, S_k$.

**Proposition 52.** Let $S_1, \ldots, S_k \subseteq V$ be $k$ sets of nodes such that $\#S_1 = \ldots = \#S_k = n$. Then,

$$\det \mathcal{N}^{(k)}_{S_1, \ldots, S_k} = \sum_{T \in \mathcal{S}(S_1, \ldots, S_k)} \text{sign}(T) m_T,$$

where $m_T$ is the split-trek-system monomial of the split-trek system

$$T = \{(P_1^{(1)}, \ldots, P_k^{(1)}), \ldots, (P_1^{(n)}, \ldots, P_k^{(n)})\},$$

defined as

$$m_T = \prod_{i=1}^n \Phi^{(k)}_{top((p_1^{(i)}, \ldots, p_k^{(i)}), \ldots, top((p_1^{(i)}, \ldots, p_k^{(i)})) \prod_{j=1}^k \lambda_{p_j^{(i)}}.}$$

Furthermore, the sum can be taken over the set $\tilde{S}(S_1, \ldots, S_k)$ of $k$-split-trek-systems without sided intersection.
We prove an analog of Theorem 35 for the tensors of 3rd-order moments ($k = 3$).

**Theorem 53.** Let $S_1, S_2, S_3$ be subsets of $V$ with $\#S_1 = \#S_2 = \#S_3$. Then,

$$\det \mathcal{N}_{S_1,S_2,S_3}^{(3)} = 0$$

if and only if every system of 3-split-treks between $S_1, S_2, S_3$ has a sided intersection.

The proofs of Proposition 51 and 52 and Theorem 53 can be found in Appendix 7.2.

For higher-order moments ($k > 3$), we conjecture the following theorem by analogy with Theorem 35.

**Conjecture 54.** Let $S_1, \ldots, S_k$ be subsets of $V$ with $\#S_1 = \ldots = \#S_k$. Then,

$$\det \mathcal{N}_{S_1,\ldots,S_k}^{(k)} = 0$$

if and only if every system of $k$-split-treks between $S_1, \ldots, S_k$ has a sided intersection.

Note that the if direction is straightforward since we can express the determinant as a sum of split-trek monomials of $k$-split-trek systems without sided intersection, as in Proposition 52. Provided Conjecture 54 is true, we can show the following relationship between moment tensors of different orders.

**Proposition 55.** Consider $k \geq 4$ sets of vertices $S_1, \ldots, S_k \subseteq V$ such that $\#S_1 = \ldots = \#S_k = n$. Let us suppose that the tensor of $k$-th order moments $\mathcal{N}_{S_1,\ldots,S_k}^{(k)}$ indexed by $S_1, \ldots, S_k$ has null determinant: $\det \mathcal{N}_{S_1,\ldots,S_k}^{(k)} = 0$. Then, for any $2 \leq h \leq k - 2$ and any partition $\{1, \ldots, k\} = \{i_1, \ldots, i_h\} \cup \{j_1, \ldots, j_{k-h}\}$, either the determinant of the $h$-th order moment tensor $\mathcal{N}_{S_{i_1},\ldots,S_{i_h}}^{(h)}$ is zero, or the determinant of the $(k-h)$-th order moment tensor $\mathcal{N}_{S_{j_1},\ldots,S_{j_{k-h}}}^{(k-h)}$ is zero.

**Proof.** Let us suppose that $\det \mathcal{N}_{S_1,\ldots,S_k}^{(k)} = 0$, but there exists $2 \leq h \leq k - 2$ and a partition $\{1, \ldots, k\} = \{i_1, \ldots, i_h\} \cup \{j_1, \ldots, j_{k-h}\}$ such that both $\det \mathcal{N}_{S_{i_1},\ldots,S_{i_h}}^{(h)} \neq 0$ and $\det \mathcal{N}_{S_{j_1},\ldots,S_{j_{k-h}}}^{(k-h)} \neq 0$. Then, by Conjecture 54 there exists an $h$-split-trek system $T_1$ with no sided intersection between $S_{i_1}, \ldots, S_{i_h}$ and a $(k-h)$-split-trek system $T_2$ with no sided intersection between $S_{j_1}, \ldots, S_{j_{k-h}}$. But then, combining together $T_1$ and $T_2$, we get a valid
$k$-split-trek system between $S_1, \ldots, S_k$ with no sided intersection, which contradicts the assumption $\det \mathcal{N}(k) = 0$. Therefore, one of $\det \mathcal{N}(h) = 0$ and $\det \mathcal{N}(k-h) = 0$ is true.

**Example 56.** We illustrate Proposition 55 for the case $k = 4$ and $h = 2$. In the graph from Figure 3-12(a), $\det \mathcal{N}(4)_{S_1, S_2, S_3, S_4} = 0$. In Figure 3-12(b), $\det \mathcal{N}(2)_{S_1, S_2} = 0$ and in Figure 3-12(c), $\det \mathcal{N}(2)_{S_3, S_4} \neq 0$.

![Figure 3-12: Illustration of Example 56](image)

Proposition 55 would give a nice relationship between the vanishing of determinants of high-order moment tensors and low-order moment tensors. One of our initial hopes was that in the case of Gaussian random variables, the vanishing of high-order moment determinants would be able to explain constraints in the model that are not subdeterminants of the covariance matrix (sometimes called Verma constraints) [Verma and Pearl, 1991]. However, Proposition 55 implies that if a high-order moment determinant vanishes, then so does a lower-order one. On the other hand there are lots of Gaussian graphical models, for which there are no covariance determinants vanishing [Drton et al., 2018], thus the vanishing of determinants would not suffice to describe the model.

### 3.6 Conclusion

In this chapter, we give implicit constraints on linear non-Gaussian structural equation models by providing a relationship between the vanishing of subdeterminants of the tensors of $k$-th order cumulants and a combinatorial criterion on the corresponding graph. Specifically, we show that the determinant of the subtensor of the $k$-th order cumulants for $k$ sets of vertices with equal cardinality vanishes if and only if there is no system of
$k$-treks between these sets without sided intersection. One of the main contributions of our work is the introduction of multi-directed edges in the hidden variable case, and our multi-trek criterion which allows us to, for example, detect the presence of a common cause of multiple vertices.
Chapter 4

Simplicial complexes as a modeling paradigm for networks

4.1 Introduction

In many concrete problems, interactions between nodes in graphs are not only pairwise but polyadic. This is the case for instance in communication networks (emails for instance), co-authorship network or protein-interaction networks. Therefore, tools and methods from graph theory (based on pairwise interactions) may not be sufficient to analyze such data.

In this chapter, we introduce the modeling of networks by simplicial complexes. We mainly follow the presentation on this topic by [Schaub et al., 2018, Lim, 2015]. This framework enables to account for high-order interactions in networks. We highlight to the reader that other choices of formalism for higher-order interactions were possible and have already been explored. We can cite for instance hypergraphs [Gaudelet et al., 2018].

4.2 Simplicial complexes

Let us consider a finite set of vertices $\mathcal{V}$. A $k$-simplex $\mathcal{S}^k$ is a subset of $\mathcal{S}$ with cardinality $k + 1$. Notice that there are no repeated elements in $\mathcal{S}^k$. A simplicial complex $\mathcal{X}$ is a set of simplices such that if $\mathcal{S} \in \mathcal{X}$, then every subsets of $\mathcal{S}$ is in $\mathcal{X}$. For instance, let’s consider
the graph $G$ with vertices $\{1, 2, 3, 4, 5, 6, 7\}$. If the 2-simplex $\{1, 3, 4\}$ is in $\mathcal{X}$, then $\mathcal{X}$ also contains the 1-simplices $\{1, 3\}, \{3, 4\}, \{4, 1\}$ and the nodes (0-simplices) $\{1\}, \{3\}, \{4\}$.

Figure 4-1: Networks with higher-order interactions represented by faces (shaded areas)

A face of $S^k$ is a subset of $S^k$ with cardinality $k$. A co-face of $S^k$ is a simplex $S^{k+1}$ of cardinality $k + 1$ such that $S^k \subset S^{k+1}$.

**Example 57.** In Figure 4-1, $\{3, 4\}, \{4, 1\}$, and $\{1, 3\}$ are faces of $\{1, 3, 4\}$. $\{5, 6, 7\}$ is a co-face of $\{5, 6\}, \{6, 7\}$ and $\{7, 5\}$.

We say that two $k$-simplices $S^k_i$ and $S^k_j$ are upper-adjacent if there exists a simplex $S^{k+1}$ with cardinality $k + 1$ such that $S^k_i \subset S^{k+1}$ and $S^k_j \subset S^{k+1}$. Two $k$-simplices $S^k_i$ and $S^k_j$ are said to be lower-adjacent is there exists a simplex $S^{k-1}$ of cardinality $k - 1$ such that $S^{k-1}_i \subset S^k_i$ and $S^{k-1}_j \subset S^k_j$.

**Example 58.** In Figure 4-1, $\{3, 4\}$ and $\{4, 1\}$ are upper-adjacent. $\{5, 2\}$ and $\{2, 3\}$ are not upper-adjacent. $\{3, 4\}$ and $\{4, 1\}$ are lower-adjacent. $\{5, 6, 7\}$ and $\{5, 8, 7\}$ are lower adjacent. $\{5, 3\}$ and $\{4, 6\}$ are not lower-adjacent.

In order to conduct operations on simplicial complexes, we need to define an orientation for each simplex. The choice of orientation is only for computation. As long as the orientation of faces is consistent along computations, the initial choice of orientation does not matter. This constitutes a major difference with the graph setting since in this latter case, nodes (i.e., 0-simplices) have no orientation.

We will say that two orderings of vertices are equivalent with respect to the orientation of a $k$-simplex (which constitutes a class equivalence for the orderings) as long as we can switch from one to the other orientation by applying an even number of permutations of the vertices.
Example 59. In Figure 4-1, the orderings [1, 3, 4] and [3, 4, 1] are equivalent. The orderings [5, 6, 7] and [6, 5, 7] are not equivalent.

The orientation of simplices is necessary to apply computations on simplicial complexes. These operations occur in the space of *chains* and *co-chains*.

### 4.3 Spaces of chains and co-chains

For a given $k$, let us define the finite dimensional vector space $C_k$ whose basis elements are the $k$-simplices $s^k_i$. An element $c_k$ in $C_k$ can be described as a linear combination of the element basis, i.e., $c_k = \sum_{i=1}^{K} \alpha_i s^k_i$, where $K = |\mathcal{X}^k|$ is the cardinality of $\mathcal{X}^k$ (i.e., number of $k$-simplices in the graph). $c_k$ can thus be represented by a vector $(\alpha_1, \ldots, \alpha_K)$. We call $c_k$ a chain and $C_k$ is the space of $k$-chains. Notice that any change in the orientation of the basis elements $s^k_i$ will induce a change in the sign of the corresponding coefficients $\alpha_j$ (i.e., multiplied by -1). We equip the vector space $C_k$ with the $\ell^2$ inner product $\langle c_1, c_2 \rangle = c_1^T c_2$.

We can then define the dual space $C^k$ of $C_k$. Elements in $C^k$ are called *alternating functions* or *co-chains*. Alternating functions take inputs in $C_k$ and outputs a value in $\mathbb{R}$.

Let us consider the following orientation of basis elements of $C_k$: $\{[i_0, \ldots, i_k] : i_0 < \ldots < i_k\}$. A $k$-cochain (or alternating function) is defined by the value that it takes on every chain that form the element basis, i.e., by the values it takes on each $[i, j], i < j$. We can then define the $\ell^2$ inner product in $C^k$ as

$$\langle f_1, f_2 \rangle = \sum_{i_0 < i_1 < \ldots < i_k} f_1([i_0, \ldots, i_k]) f_2([i_0, \ldots, i_k]) = f_1^T f_2$$

### 4.4 Boundary and co-boundary operators

We define the boundary operator $\partial_k$ as follows

$$\partial_k : C_k \rightarrow C_{k-1}$$

$$\partial_k([i_0, \ldots, i_k]) = \sum_{j=0}^{k} (-1)^j [i_0, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k]$$
The boundary operators map any element in the space of chains to its boundary constituents, i.e., the \(k-1\)-simplices that are lower-adjacent to the chain considered. Let us denote by \(\text{im}(\cdot)\) the image of an operator, then \(\text{im}(\partial_k)\) is the space of \((k-1)\)-boundaries. If \(c_k \in C_k\) is a cycle (i.e., the starting point and the ending point are the same), then \(\partial_k(c_k) = 0\). Therefore \(\ker(\partial_k)\) is called the space of \(k\)-cycles.

In the same way that the space \(C^k\) of cochains in the dual space of the space \(C_k\) of chains, we can define a dual operator \(\partial^T_k\) to the boundary operator. \(\partial^T_k\) is called the co-boundary operator and is defined as

\[
\partial^T_k : C_k \rightarrow C_{k+1} \tag{4.3}
\]

\[
(\partial^T_k f)([i_0, \ldots, i_{k+1}]) = \sum_{j=0}^{k+1} (-1)^j f([i_0, \ldots, i_{j-1}, i_{j+1}, \ldots, i_{k+1}]) \tag{4.4}
\]

Notice that boundary and co-boundary operators are linear maps between finite dimensional vector spaces. We can represent these mappings by matrices. The boundary operator \(\partial_k\) is given by the higher-order incidence matrix \(B_k\). The co-boundary operator \(\partial^T_k\) is given by \(B^T_k\).

**Example 60.** In Figure 4-1, the boundary maps \(B_1, B_2\) are given by the following matrices.

\[
\begin{bmatrix}
1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 1 & 0 \\
6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1
\end{bmatrix}
\]
The higher-order boundary maps enable to generalize the graph Laplacian operator for graphs with higher-order interactions.

### 4.5 Graph Laplacian and Hodge Laplacian

#### 4.5.1 Graph Laplacian

A keystone operator in graph theory is the graph Laplacian $L_0$. Let us first define the adjacency matrix $A$ whose entries are defined as follows

\[
A_{ij} = \begin{cases} 
1, & \text{if } i \text{ is connected to } j \\
0, & \text{otherwise}
\end{cases}
\quad (4.5)
\]

\[
B_2 = \begin{pmatrix}
[1, 2] & 1 & 0 & 0 \\
[2, 3] & 0 & 0 & 0 \\
[3, 4] & 1 & 0 & 0 \\
[4, 1] & 1 & 0 & 0 \\
[4, 6] & 0 & 0 & 0 \\
[5, 2] & 0 & 0 & 0 \\
[5, 3] & 0 & 0 & 0 \\
[5, 6] & 0 & 1 & 0 \\
[5, 8] & 0 & 0 & 1 \\
[6, 7] & 0 & 1 & 0 \\
[7, 5] & 0 & 1 & 1 \\
[8, 7] & 0 & 0 & 1 
\end{pmatrix}
\]
Furthermore, let $\mathbf{D}$ be the matrix of degrees. The degree $d(i)$ of a node $i$ is the number of neighbors of $i$. Hence we have

$$D_{ij} = \begin{cases} d(i), & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \quad (4.6)$$

The matrix $\mathbf{D}$ can be obtained by

$$\mathbf{D} = \text{diag}(\mathbf{A1}) \quad (4.7)$$

where $\text{diag}(\cdot)$ is the diagonal operator and $\mathbf{1}$ is the vector of ones.

Then, the graph Laplacian operator is defined as

$$L_0 = \mathbf{D} - \mathbf{A} \quad (4.8)$$

$$= \mathbf{B}_1 \mathbf{B}_1^T \quad (4.9)$$

where $\mathbf{B}_1$ is the incidence matrix.

The graph Laplacian gives an algebraic description of a graph. The spectral properties of this operator are commonly used in order to learn topological properties of a graph.

The Hodge Laplacian operators, also called Eckmann Laplacians ([Eckmann, 1944]), are a generalization of the graph Laplacian for graphs with higher-order interactions.

### 4.5.2 Hodge Laplacians

The Hodge Laplacians are defined by the boundary operators presented in [4.4]

$$L_k = \mathbf{B}_k^T \mathbf{B}_k + \mathbf{B}_{k+1} \mathbf{B}_{k+1}^T \quad (4.10)$$

Notice that the graph Laplacian is a special case of the Hodge Laplacian for $k = 0$ (where $\mathbf{B}_0 = 0$).

The Hodge Laplacian $L_k$ can be seen as a discretized version of the Laplace equation $\Delta x = 0$ ([Schaub et al., 2018] [Lim, 2015]). Therefore, the elements $\mathbf{h} \in \ker(L_1)$ are called
harmonic functions.

The harmonic functions also have a topological meaning in terms of "holes" in the graph. This interpretation from the definition of boundary maps (section 4.4). Let us recall that

\[ \partial_k \circ \partial_{k+1} = 0 \quad (4.11) \]
\[ \partial_{k+1}^T \circ \partial_k^T = 0 \quad (4.12) \]

Equations (4.11) and (4.12) are induced by the idea that the boundary of a boundary is empty. In matrix form, it follows that

\[ B_k B_{k+1} = 0 \quad (4.13) \]
\[ B_{k+1}^T B_k^T = 0 \quad (4.14) \]

From equation (4.11), we have \( \partial_k \circ \partial_{k+1} = 0 \), which implies that \( \text{im}(\partial_{k+1}) \subset \text{ker}(\partial_k) \). From this, we define the homology vector spaces \( H_k \) of \( X \) over \( \mathbb{R} \). An element \( x \) is in \( H_k \) if \( x \in \text{ker}(\partial_k) \) and \( x \notin \text{im}(\partial_{k+1}) \). In other words

\[ H_k := H_k(X, \mathbb{R}) = \ker(\partial_k)/\text{im}(\partial_{k+1}) \quad (4.15) \]

Elements in \( H_k \) are \( k \)-cycles that are not engendered by a \( k \)-boundary and can be interpreted as \( k \)-holes in the simplicial complex \( X \).

The dimension of \( H_k \) thus gives the number of \( k \)-holes in \( X \) and is called the \( k \)-th Betti number. If we recall that \( L_k = B_k^T B_k + B_{k+1}^T B_{k+1}^T = \partial_k^T \partial_k + \partial_{k+1} \partial_{k+1}^T \), we can prove that the \( k \)-th Betti number is also the dimension of the null space of the \( k \)-th Hodge Laplacian \( \ker(L_k) \) which is the set of harmonic functions [Lim, 2015].
4.6 Hodge decomposition

Since \( L_k = B_k^T B_k + B_{k+1} B_{k+1}^T \), any element \( h \in \ker(L_k) \) satisfies

\[
\begin{align*}
    h &\in \ker(B_k) & (4.16) \\
    h &\in \ker(B_{k+1}^T) & (4.17)
\end{align*}
\]

Furthermore, from standard linear algebra, we have

\[
\begin{align*}
    \text{im}(B_k^T) &\perp \ker(B_k) & (4.18) \\
    \text{im}(B_{k+1}) &\perp \ker(B_{k+1}^T) & (4.19)
\end{align*}
\]

Combining these results, the space of chains \( C_k \) can be decomposed as the union of the orthogonal subspaces \( \text{im}(B_{k+1}), \text{im}(B_k^T) \) and \( \ker(L_k) \) with respect to the standard inner product

\[
C_k = \text{im}(B_{k+1}) \oplus \text{im}(B_k^T) \oplus \ker(L_k) \quad (4.20)
\]

This decomposition is called the Hodge decomposition and provides additional insights to the centrality measure on simplicial complexes that we investigate in Chapter 5.
Chapter 5

Random walks on simplicial complexes and edge PageRank

5.1 Introduction

One main motivation of this thesis relates to higher-order interactions in networks. In Chapter 4, we presented the tools from algebraic topology that generalize the graph Laplacian operator in order to analyze observations modeled by simplicial complexes. We are now ready to investigate the initial question stated in the introduction of this thesis: Can we get additional insight about data using higher-order interactions?

One possible way to address this question is to follow Schaub et al. developments ([Schaub et al., 2018]) about diffusion on simplicial complexes. More precisely, how to define diffusion on simplicial complexes? What is a random walk on simplicial complexes? By analogy to the pairwise interaction case, can we define a centrality measure based on diffusion on simplicial complexes and if so, what are the properties of this centrality measure?

We start by reminding results on diffusion processes on graph.
5.2 Diffusion processes on graphs

A diffusion process on a graph can be modeled via a random walk, where the states of the random walker are the vertices and the transition proceed along edges. A standard way to define a random walk on a graph \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \) is through the following transition rule

\[
p_{t+1} = AD^{-1}p_t \tag{5.1}
\]

where \( A \) is the adjacency matrix of the graph, \( D \) is the matrix of degrees and \( p_t \) is a vector of size \( |\mathcal{V}| \) whose \( i \)-th entry \( p_t(i) \) records the probability that the random walker is at the node \( i \) at time \( t \). The initial distribution is given by \( p_0 \). Since the graph Laplacian \( L_0 \) is given by \( L_0 = D - A \), we can also write

\[
p_{t+1} = (D - L_0)D^{-1}p_t \tag{5.2}
\]

\[
= (I - L_0D^{-1})p_t \tag{5.3}
\]

As noted in [Schaub et al., 2018], one interesting property of this transition rule is that the probability distribution at time \( t \) depends on the graph Laplacian. In agreement with intuition, this shows that the transition depends on the topology of the graph since the harmonic functions of the graph Laplacian corresponds to holes in the graphs.

If we want to extend a random walk on simplicial complexes, one main issue arises: how to keep track of the orientation of the simplices? Notice that \( k \)-simplices correspond to the state of the random walk in this case. For a random walk on the node space, the answer is trivial since nodes do not have orientation. However, this is no longer the case for \( k \)-simplices \( (k > 0) \).

In order to address this issue, we adopt Schaub et al. definition of random walk on simplicial complexes ([Schaub et al., 2018]). As far as we know, this is the first work that provides a framework for diffusion processes on simplicial complexes. The next section closely follows section 3 in [Schaub et al., 2018].
5.3 Diffusion processes on simplicial complexes

In order to describe a diffusion process on simplicial complexes, Schaub et al. propose to consider a random walk on a high-dimensional lifted space. By analogy with equation (5.2), the Hodge Laplacian $L_k$ will act on a cochain vector $f \in C^k$ (instead of the vector $p_t$) that can be interpreted as a flow. Prior to this, $f$ must be lifted to a higher-dimensional space. We then apply $L_k$ on it. More precisely, we apply a linear operator that depends on $L_k$ and that must be normalized in order to respect the property of a diffusion process. We finally project back the output into the original space. These three operations - lift, apply, project - enables us to distinguish the magnitude and the orientation of the flow $f$. The magnitude of the flow can be interpreted as a volume. In this view, the magnitude is analog to a probability: a higher magnitude corresponds to a higher probability for the random walker to make a transition. On the other hand, the sign of the coefficients in $f$ corresponds to the direction of the walk and refers to the orientation of reference that we have chosen for the simplicial complexes (cf. section 4.2).

5.3.1 Lifting of an edge-flow in a higher-dimensional space

In this section, we describe the action of the Hodge Laplacian $L_1$ on a cochain $f$ that is lifted in a higher-dimensional space.

A lifting of a cochain $f \in C^1$ into a higher-dimensional space $D^1$ is equivalent to accepting all possible orientations for the corresponding chains. We give an example in Figure 5-1 where the two possible orientations of the edges are considered in the space $D^1$.

There are two possible orientations for each chain, hence $|D^1| = 2|C^1|$. Let us define the map $C^1 \to D^1$ that maps each cochain to its two possible orientations. In other words, the mapping $V$ represents the two possible directions of the corresponding edge. Let us define the matrix representation of $V$ as

$$V = \begin{pmatrix} +I_{n_1} \\ -I_{n_1} \end{pmatrix} \in \mathbb{R}^{2n_1 \times n_1} \quad (5.4)$$
where $n_1 = |\mathcal{C}^1|$ and $I_{n_1}$ is the identity matrix with dimension $n_1$.

So far, we have defined the lifting of a cochain (i.e., a flow). Let us now define the lifting of a linear operator.

### 5.3.2 Lifting of a linear matrix operator

**Definition 61** ([Schaub et al., 2018, Definition 3.1]). A matrix $N \in \mathbb{R}^{2n_1 \times 2n_1}$ is a lifting of a matrix $M \in \mathbb{R}^{n_1 \times n_1}$ if the following condition holds:

$$V^T N = MV^T$$

(5.5)

From the definition of $V$, we have

$$V^T V = 2I_{n_1},$$

(5.6)

which implies that the Moore-Penrose pseudoinverse of $V$ is

$$V^+ = \frac{1}{2} V^T$$

(5.7)
From this, if the matrix operator $M$ has a lifting $N$, then we get

$$V^T N = MV^T$$  \hspace{1cm} (5.8)  
$$V^T NV = MV^TV$$  \hspace{1cm} (5.9)  
$$V^T NV = 2M$$  \hspace{1cm} (5.10)

This can be rewritten as

$$M = \frac{1}{2} V^T NV$$  \hspace{1cm} (5.11)  
$$= V^\dagger NV$$  \hspace{1cm} (5.12)

Equation (5.12) shows that the matrix operator $M$ can be interpreted as a succession of a lifting to a higher-dimensional space (apply $V$), a linear transformation (apply $N$) and a projection back to the initial lower-dimensional space (apply $V^\dagger$).

We are now ready to consider the lifting of the Hodge Laplacians $L_1$.

### 5.3.3 Lifting of Hodge Laplacians

Following [Schaub et al., 2018] notation, let us define the following matrices

$$\hat{B}_1 := B_1 V^T = (B_1, -B_1) \text{ and } \hat{B}_2 := VB_2 = \begin{pmatrix} B_2 \\ -B_2 \end{pmatrix}$$  \hspace{1cm} (5.13)

Let us also denote by $\hat{B}^+_1, \hat{B}^-_1$ the positive and negative parts of $\hat{B}_1$ respectively. We can define the lifting of the Hodge-Laplacian $L_1$ as follows.

**Lemma 62** ([Schaub et al., 2018, Lemma 3.2]). The negative of the Hodge Laplacian $L_1 = B_1^T B_1 + B_2 B_2^T$ has a lifting $\hat{A}$:

$$-L_1 V^T = V^T \hat{A} \iff -L_1 = V^\dagger \hat{A} V,$$  \hspace{1cm} (5.14)
where \( \hat{A} = \hat{A}^l + \hat{A}^u \), with

\[
\hat{A}^l = (\hat{B}_1^+)^T B_1^+ + (\hat{B}_1^-)^T \hat{B}_1^- \quad \text{and} \quad \hat{A}^u = \hat{B}_2^+(\hat{B}_2^-)^T + \hat{B}_2^- (\hat{B}_2^+)^T
\]  

(5.15)

The matrix \( \hat{A} \) is nonnegative and symmetric. We can interpret \( \hat{A} \) as the weighted adjacency matrix of an undirected graph with \( 2n_1 \) nodes. \( \hat{A}^l \) provides the relationships between lower-adjacent edges and its entries are given by

\[
((\hat{B}_1^-)^T \hat{B}_1^+)[i_1,i_2][j_1,j_2] = \begin{cases} 
1, & \text{if } j_2 = i_1 \\
0, & \text{otherwise}
\end{cases}
\]

(5.16)

The first matrix is called a forward walk while the second matrix is called a backward walk.

In the graph of Figure 5-1, let us assume that the starting edge is [5, 4]. The edges that are connected to [5, 4] by a lower-adjacent backward walk are [2, 5], [3, 5], [4, 5][6, 5]. The edges that are connected to [5, 4] by a lower-adjacent forward walk are [4, 3] and [4, 5].

The matrix \( \hat{A}^u \) describes relationships between upper adjacent edges that share a common co-face \( \mathcal{S}^3 \).

\[
(\hat{A}^u)[i_1,i_2][j_1,j_2] = \begin{cases} 
1, & \text{if } [j_1, j_2] \sim [i_1, i_2] \\
0, & \text{otherwise}
\end{cases}
\]

(5.17)

where \( \sim \) means that \([j_1, j_2]\) and \([i_1, i_2]\) have a different orientation with respect to their joint co-face. Going back to Figure 5-1, an upper adjacent walk starting from the edge [5, 4] can make a transition to [4, 5], [5, 3], [3, 4].

Lemma 62 gives us the lifting of the Hodge Laplacian \( L_1 \) in a higher-dimensional space. However, in order to define a diffusion process, this operator must be appropriately normalized. While various normalizations are possible (by changing the weights), we adopt the same normalization as in [Schaub et al., 2018].

**Definition 63** ([Schaub et al., 2018, Definition 3.3]). Consider a simplicial complex \( \mathcal{X} \),
whose boundary operators can be represented by the matrices $B_1$ and $B_2$. The normalized Hodge Laplacian matrix is then defined as:

$$
\mathcal{L}_1 = D_2 B_1^T D_1^{-1} B_1 + B_2 D_3 B_2 D_2^T,
$$

(5.18)

where $D_2$ is the diagonal matrix of adjusted degrees of each edges via

$$(D_2)_{[i_1,i_2],[i_1,i_2]} = \max\{\deg([i_1,i_2]), 1\},$$

(5.19)

$D_1 = 2 \cdot \text{diag}(|B_1 D_2| 1)$ is the diagonal matrix of weighted degrees of the nodes, and $D_3$ is the diagonal matrix with $1/3$ on the diagonal.

The next result shows that the normalized Hodge Laplacian $\mathcal{L}_1$ enables to define a random walk in the lifted space.

**Theorem 64** ([Schaub et al., 2018, Theorem 3.4]). The matrix $-\frac{1}{2} \mathcal{L}_1$ has a stochastic lifting, i.e., there exists a stochastic matrix $\widehat{P}$ such that:

$$
-\frac{1}{2} \mathcal{L}_1 V^T = V^T \widehat{P}
$$

(5.20)

$$
\widehat{P} := \frac{1}{2} P_{\text{lower}} + \frac{1}{2} P_{\text{upper}}
$$

(5.21)

where $P_{\text{lower}}$ is the transition matrix of a random walk determined by the lower-adjacent connections and $P_{\text{upper}}$ is the transition matrix of a random walk determined by the upper adjacent connections. The transition matrix $P_{\text{lower}}$ is defined by a ‘forward walk’ and ‘backward walk’ component moving in the orientation of the edges or against it, respectively:

$$
P_{\text{lower}} := \frac{1}{2}(P_{\text{lower, forward}} + P_{\text{lower, backward}})
$$

(5.22)

$$
P_{\text{lower, forward}} = M_f \text{diag}(1^T M_f)^{-1}
$$

(5.23)

$$
P_{\text{lower, backward}} = M_b \text{diag}(1^T M_b)^{-1}
$$

(5.24)

where $M_f = \hat{D}_2 (\hat{B}_1^{-1})^T \hat{B}_1^+$ and $M_b = \hat{D}_2 (\hat{B}_1^+)^T \hat{B}_1$ are weighted lower adjacency matrices corresponding to forward and backward walks along the edges, and $\hat{D}_2 = \text{diag}(D_2, D_2)$. The transition
matrix $P_{\text{upper}}$ describes a random walk along upper adjacent faces as follows:

$$P_{\text{upper}} = \hat{A}^u \hat{D}_2^\# + \frac{1}{2} \left( \begin{array}{ccc} I & I \\ I & I \end{array} \right) \hat{D}_2^{(0)},$$  \hspace{1cm} (5.25)

where $\hat{A}^u = \hat{B}_2^+ (\hat{B}_2^-)^T + \hat{B}_2^- (\hat{B}_2^+)^T$ is the matrix of upper adjacent connections, $\hat{D}_2^\#$ denotes a diagonal (inverse matrix) defined as:

$$\hat{D}_2^\#_{[i_1,i_2],[i_1,i_2]} = \begin{cases} 
1, & \text{if } \deg([i_1,i_2]) = 0 \\
1/(3 \deg([i_1,i_2])), & \text{otherwise}
\end{cases} \hspace{1cm} (5.26)$$

and $\hat{D}_2^{(0)}$ is the diagonal matrix selecting all edges with no upper adjacent faces:

$$\hat{D}_2^{(0)}_{[i_1,i_2],[i_1,i_2]} = \begin{cases} 
1, & \text{if } \deg([i_1,i_2]) = 0 \\
0, & \text{otherwise}
\end{cases} \hspace{1cm} (5.27)$$

For the applications presented in Chapter 6, we need to define a normalized Hodge decomposition.

### 5.3.4 Normalized Hodge decomposition

Following [Schaub et al., 2018], we define the following normalized (weighted) Hodge decomposition from the normalized Hodge Laplacian:

$$\mathbb{R}^{n_1} = \text{im}(B_2) \oplus_{D_2^{-1}} \text{im}(D_2 B_1^T) \oplus_{D_2^{-1}} \ker(L_1)$$  \hspace{1cm} (5.28)

where $\oplus_{D_2^{-1}}$ is the union of orthogonal subspaces with respect to the inner product $\langle u, v \rangle_{D_2^{-1}} = u^T D_2^{-1} v$.

For the application that we investigate in the next chapter, we use a symmetrized version of the normalized Hodge Laplacian $L_1^s = D_2^{-1/2} L_1 D_2^{1/2}$, and the corresponding
Hodge decomposition is given by

\[ \mathbb{R}^{n_1} = \text{im}(D_2^{-1/2}B_2) \oplus \text{im}(D_2^{1/2}B_1^T) \oplus \ker(L_1^z) \] (5.29)

The Hodge decomposition is an orthogonal decomposition of a vector space. It shows that every vector \( x \in \mathbb{R}^{n_1} \) can be decomposed as follows

\[ x = g \oplus r \oplus h \] (5.30)

where \( g \) is the projection of \( x \) onto \( \text{im}(D_2^{-1/2}B_1^T) \), \( r \) is the projection of \( x \) into \( \text{im}(D_2^{-1/2}B_2) \), and \( h \) satisfies \( L_1^zh = 0 \).

This decomposition has a nice interpretation.

- \( \text{im}(D_2^{-1/2}B_1^T) \) is the weighted cut space of the edges ([Schaub et al., 2018]). This subspace can be seen as the linear combinations of simplices whose cyclic components are zero. \( g \) is a gradient flow between two nodes.

- \( \text{im}(D_2^{-1/2}B_2) \) is made of the weighted flows that can be described via local circulations along 2-simplices (i.e. made of 3 nodes). \( r \) is a circulation around a filled triangle in the graph and can be seen as a curl flow.

- \( h \in \ker(L_1^z) \) and therefore \( h \) can be seen as a global (by contrast to local in the curl flow case) circulation around a cycle whose sum of components is not zero. By analogy with the Laplace equation, \( h \) is the harmonic component of \( x \).

So far, we have extended the graph theoretical tools from graph theory to simplicial complexes. In particular, the Hodge Laplacian and the Hodge decomposition enable us to take into account higher-order interactions in networks, instead of only pairwise interactions.

We are now ready to address the following question: can we extend the concept of centrality measure to simplicial complexes?
5.4 Edge PageRank

In the graph setting, a centrality measure aims at ranking nodes by their importance. Centrality measures vary by the definition of importance that they adopt. Most commonly used centrality measures are degree centrality (number of neighbors), betweenness centrality (number of shortest paths between pair of nodes that pass by a given node), eigenvector centrality (account for the importance of one’s own neighbors). The PageRank centrality measure has reached a large audience since the 2000’s as it was developed by Google’s founders in their webpages ranking algorithm. Schaub et al. have recently proposed a version of the PageRank centrality on simplicial complexes, more precisely for edges (instead of nodes). The idea is to measure the influence of each edge on the other edges. We first review the standard node PageRank and then present the edge PageRank. Again, this section closely follows Schaub et al.’s developments (Section 6 in [Schaub et al., 2018]).

5.4.1 Node PageRank on graphs

The personalized PageRank centrality can be defined as follows

\textbf{Definition 65 ([Gleich, 2015, Definition 2.1])}. Let \( P \) be a column-stochastic matrix where all entries are nonnegative and the sum of entries in each column is 1. Let \( \bar{v} \) be a column-stochastic vector (\( e^T v = 1 \)), and let \( 0 < \alpha < 1 \) be the teleportation parameter. Then the PageRank problem is to find the solution of the linear system

\[
(I - \alpha P) \pi = (1 - \alpha)^{-1},
\]

where the solution \( \pi \) is called the PageRank vector.

The interpretation of equation (5.31) is as follows. The vector \( \pi \) is the stationary distribution of the random walk on the graph. The movements of the random walker occurs as follows:

- with probability \( \alpha \), the random walker makes a step according to the transition matrix \( P \)
• with probability $1 - \alpha$, the random walker jumps to a node in the graph according to the probability distribution $\mu$

There are two standard ways to assign the vector $\mu$.

• each entry of $\mu$ is equal to $1/n$, where $n$ is the length of $\mu$. In other words, the probability to jump to a given node is uniform

• $\mu$ is an indicator vector. In this case, when teleportation occurs, the random walker restarts from the node $i$ whose corresponding entry $\mu_i$ in $\mu$ is 1. In this case, we can interpret the solution $\pi$ of (5.31) as the influence of the node $i$ on every other nodes.

The next section focuses on the generalization of the PageRank centrality measure to edges, as proposed in [Schaub et al., 2018].

### 5.4.2 PageRank for $k$-simplices

Focusing on the edge case, we follow Schaub et al.’s definition of PageRank vector for edges in simplicial complexes.

**Definition 66 ([Schaub et al., 2018, Definition 6.1]).** Let $\mathcal{X}$ be a simplicial complex with normalized Hodge Laplacian $L_1$, $x$ be a vector of the form $x = V^T\bar{x}$ where $\bar{x}$ is a probability vector, and $\beta \in (2, \infty)$. The PageRank vector $\pi_1$ of the edges is then defined as the solution to the linear system

$$(\beta I + L_1)\pi_1 = (\beta - 2)x$$

(5.32)

Let us assume that $x$ is an indicator, whose $i$-th entry is non-zero. Then the entry $j$ of $\pi_1$ can be interpreted as the influence of the edge $i$ onto the edge $j$.

For the applications, using the normalized Hodge decomposition defined in Section 5.3.4 we will also consider the projections of the vector $\pi_1$ onto the gradient, curl and harmonic spaces.

Furthermore, in order to provide a score to each edge, we compute the two-norm of the stationary distribution vector $\pi_1$ (or alternatively, the two-norm of its projection vectors).
Chapter 6

Applications and empirical results

In this chapter, we investigate the edge PageRank centrality measure defined in Chapter 5. In particular, we aim at addressing the following questions:

- What features in the graph does edge PageRank identify?
- What is the meaning of these features?
- Does edge PageRank extract information from the data that is not retrieved by using node-based methods?

The outline of this chapter follows the chronological order of our investigation. We first start by analyzing synthetic data. We then apply the edge PageRank centrality on true datasets. Except when explicitly mentioned, we use a parameter $\beta = 2.5$ in equation (5.32).

6.1 Application of edge PageRank to synthetic data

We first map a graph of 41 nodes that contains 56 triangles. This map has the property that all triangles are filled except one of them. We choose the open triangle in the middle of the map so as to avoid boundary effects.

We then apply the edge PageRank and its projections (gradient, curl, harmonic) into the subspaces. The results are shown in Figures 6-2a, 6-2b, 6-3a and 6-3b. We observe...
that the harmonic component (Figure 6-3b) is mostly concentrated around the hole. We notice a concentric effect as edges are getting closer to the hole. By contrast, there is no gradient component around the hole (Figure 6-2b). This confirm the intuition that a harmonic component is a global circulation around a cycle, while a gradient is induced by
Figure 6-4: Map of 54 filled triangles and 2 holes

a flow. The figure showing the curl component (Figure 6-3a) does not provide more insights since the empty cycle is surrounded by filled cycles. As a consequence, the edges around the hole have both harmonic components (due to the empty cycle) and curl component (due to the filled cycle). The figure showing the two-norm of the edge PageRank vector before projection merges the effects described above. We can nevertheless observe the predominance of the harmonic component around the hole.

In order to confirm these observations, we repeat the experiments with 2 holes instead (Figure 6-4).

Applying the edge PageRank centrality, we obtain Figures 6-5a, 6-5b, 6-6a and 6-6b

![Figure 6-5: 2-norm of edge PageRank (Figure 6-5a) and 2-norm projected in the gradient space (Figure 6-5b)](image)

Again, the most insightful results are obtained for the gradient and harmonic components and confirm our expectations based on the 1-hole case.

We can further explore the gradient component by considering a synthetic graph with two communities. More precisely, we build a graph of two communities, each of 20 nodes.
Figure 6-6: 2-norm of edge PageRank projected in the curl space (Figure 6-6a) and 2-norm projected in the harmonic space (Figure 6-6b)

Each community is a clique and the two communities are connected by a single edge. Furthermore, we add a (filled) face for every triangle in the graph. Filled triangles are represented in shaded color. We plot below the network structure and the edge PageRank score for each component (Figure 6-7 and Figure 6-8).

Figure 6-7: Network structure and edge PageRank projected in the gradient space

Figure 6-8: Edge PageRank projected in the curl space and in the harmonic space
We can observe that the curl component is high around filled triangles (i.e., within communities) and the gradient component is high on the edge that link the two communities. This confirms the intuition that the gradient corresponds to a flow or a bottleneck.

We pursue this analysis of synthetic data with communities by considering a stochastic block model (Figure 6-9). We construct a graph of 20 nodes and two communities, each made of 10 nodes. The probability for two nodes in the same community to be connected is $p = 0.75$. The probability for two nodes in different communities is $q = 0.25$. We test the edge PageRank measure with different $\beta$ parameters without observing any modification in the relative measure. Only the absolute scale is changed. Figure 6-9 is represented for $\beta = 5$ (i.e. teleportation probability $\alpha = 0/4$). We observe that the curl components are exclusively around the filled triangles. The gradient components are found between components even though we can observe some edges within communities with a gradient component. The harmonic component is very low on edges around filled triangles.

Finally, we conclude this exploratory phase on synthetic data by considering graphs with global or local bridges. A global bridge is an edge between two nodes such that if it is removed, the two nodes are in two different components of the graph. This was the case of the edge connecting the two communities in Figure 6-7 and Figure 6-8. A local

Figure 6-9: SBM - 20 nodes - $p = 0.75$ - $q = 0.25$
bridge is an edge between two nodes such that if it is removed, the distance between the two nodes is strictly more than two. We test the edge PageRank measure (Figure 6-11) on graphs with such structural properties using examples from Easley and Kleinberg (Figure 6-10; [Easley and Kleinberg, 2010], pages 50-51).

Figure 6-10: Graph with one local bridge

We also plot the node PageRank centrality measure (Figure 6-12). Node PageRank identifies the same edge as edge PageRank centrality measure in this case (one local bridge).

We now consider a graph with two local bridges, for which we plot the structure
We observe that for both cases (1 and 2 local bridges), local bridges have a high harmonic and gradient component. Furthermore, we observe that the nodes with the highest node PageRank in Figure 6-15 are not all part of the local bridges. In other words, the nodes identified by the node PageRank and the edge PageRank are not necessarily identical.

We also tested the edge PageRank measure on graphs with local bridges that do not present symmetries as in Figures 6-10 and 6-13. We obtain similar conclusions.

Before analyzing true datasets, we summarize below the observations from our exper-
Figure 6-14: Visualization of edge PageRank on the graph with two local bridges

Figure 6-15: Node PageRank on the graph with two local bridges

- Gradient component of edge PageRank identifies global bridge and corresponds to a flow or a bottleneck
- Curl component of edge PageRank identifies edges around filled triangles
- Harmonic component of edge PageRank identifies edges around holes in the graph
- Local bridges have a high gradient and harmonic component of edge PageRank
Node PageRank and edge PageRank do not necessarily provide identical information about the graph.

We are now ready to analyze true datasets. The objective is to understand what we learn about the data by using the edge PageRank.

### 6.2 Application of edge PageRank to real datasets

The results of this section were obtained in collaboration with Arnab Sarker.

#### 6.2.1 Communication network

The first type of dataset that we analyze is a communication network. It consists of the communication by emails between the employees of the company Enron. Enron was an energy, commodities, and services company that is remembered for its sudden bankruptcy in December 2001 after many years of success. Management failures have often been suggested in order to explain this fall. In this dataset, a simplex includes the employee that sends the email and all the employees that receive it. There are 148 nodes in this dataset and we observe the communications for the period 1998-1999. The visualization of the network is too dense to be instructive. However, we are able to record the number of local bridges in the network over time. We then compare the evolution of the number of local bridges with the evolution of the stock price of the company.

![Figure 6-16: Evolution of the number of local bridges (left) and of the stock price (right).](image-url)
a simultaneity between the rise of the number of local bridges and the rise of the stock price. Similarly, there is a coincidence between the fall of the number of local bridges and the fall of the stock price.

In sociology, for a long time, a question has remained unanswered and still motivates research: during periods of shocks, do people rely on their weak ties or their strong ties? The notion of weak ties has been popularized by Granovetter in 1973 ([Granovetter, 1973a]) who shows that people often rely on distant relatives rather than close parents in order to find a job. Recently, Romero et al. have analyzed instant messages data from an hedge fund showing that the network tends to turtle up under stress, i.e., people tends to rely on their strong ties ([Romero et al., 2016]). Assuming that local bridges are the structural translation of weak ties, our result about the Enron data seems to corroborates this study. During prolific period (increase of stock price), people would rely on their weak ties (increase of the number of local bridges). During periods of stress (decrease of stock price), people would rely on their weak ties (decrease of the number of local bridges).

### 6.2.2 Co-authorship network

The second type of network with higher-order interactions that we analyze is a co-authorship network. More specifically, the dataset at hand is the DBLP dataset. This dataset records the academic articles related to the field of computer science. In this co-authorship network, a simplex corresponds to one article and includes all the authors. More specifically, we analyze the period 2010-2016 and the papers that were accepted at the six following conferences: VLDB, SIGMOD, ICML, NIPS, AAAI, IJCAI. The network that we analyze has about 17,000 nodes, 50,000 edges and 42,000 triangles.

The dataset is interesting because it provides metadata, namely the citations for each paper. Our initial hypothesis was that articles with high edge PageRank, in particular a high harmonic component, may have a large number of citations since they bridge a gap (i.e., a hole in the topological framework).

We first plot the number of citations as a function of each component of the edge PageRank (Figure 6-17). We then plot the number of publication for each pair of authors
Figure 6-17: Citations as a function of edge PageRank (i.e., for each edge) as a function of each component of the edge PageRank (Figure 6-18).

Figure 6-18: Number of publications as a function of edge PageRank as a function of edge PageRank

We also plot the span of each edge as a function of the components of the edge PageRank (Figure 6-19). If an edge is a local bridge joining two nodes, then the span of that edge...
is the length of the path between these two nodes if we remove the local bridge. A global bridge has an infinity span (since we disconnect the graph). For visualization purpose, we set the span to 20 in this case. An edge that is neither a global bridge nor a local bridge has a span equal to 0.

Figure 6-19: Span of the edges as a function of edge PageRank

We observe that there is no obvious relationships between citations and edge PageRank (Figure 6-17), thus giving some evidence that our hypothesis does not hold. Furthermore, from Figure 6-18 we notice that edge with high harmonic edge PageRank tends to have few publications. On the contrary, edges with high curl edge PageRank tends to have many publications. Therefore, we could make the hypothesis that high curl would be a good predictor of high collaboration. In other words, an edge participating in a lot of filled triangles would indicate that there are multiple copies of this edge. Finally, from Figure 6-19 we can see that local bridges tends to have high harmonic component, medium gradient component and zero curl. This observation is consistent with our results on synthetic data.

Even though local bridges seem to have a high harmonic edge PageRank, an edge with high harmonic component is not necessarily a local bridge. We could say that local bridges are a subset of the high harmonic edge PageRank.
In Figure 6-20 below, we distinguish the edges with high harmonic component that are local bridges from those that are not. We observe that edges with high harmonic component but that are not a local bridge have a very similar behavior to those that are local bridges, i.e., they also have a low curl component and a medium gradient. The interpretation of these edges is left for future research.

From our analysis on true data, we draw the following conclusions:

- **High harmonic edge PageRank** identifies local bridges in networks
- **Local bridges** have a high harmonic component, low curl and medium gradient
- **Local bridges** are a subset of the edges with high harmonic edge PageRank
- **High gradient edge PageRank** identifies global bridges in networks
- **Global bridges** have a high gradient component, low curl and low harmonic

We found evidence that the edge PageRank centrality measure is connected to the notion of bridges in networks. This notion refers to the structure of the graph that represents
the network. Bridges in networks actually connect to the well-defined sociological concept called *weak tie* that we present in the section 6.4. This connection makes the edge PageRank centrality measure a relevant tool to analyze network data and in particular to identify weak ties in networks.

### 6.3 Further experiments

Other experiments have been tempted without providing conclusive results. However, they deserve to be listed for the sake of completeness.

- We have considered the small-world graph model (Watts-Strogatz network) with four neighbors. We compute the edge PageRank of the non-edge, i.e., edges that are actually not in the graph. We then add these edges one by one (removing the edge previously added when we add a new edge) and measure the speed of the spread of an infection in the network. Our hypothesis was that edges with high edge PageRank would be good at spreading the infection in the network. This experiment does not enable us to conclude. The error on the average speed is too large to conclude whether edges with high edge PageRank perform better than edges picked at random.

- Searching for the positions of the employees of Enron in the company, we have tried to find whether edges with high edge PageRank connects nodes with specific positions. We haven’t identified any pattern.

- As for the co-authorship network, we have also carefully looked at the names of the nodes of the edges with high edge PageRank. Again, we haven’t identified any significant pattern.

### 6.4 The concept of *bridges* in the sociology literature

By definition, the notion of bridges in networks is a topological concept that describes the structure of the graph. In this section, we give some motivations about why bridges...
actually provide additional information on the data itself. [Easley and Kleinberg, 2010] have shown that under specific assumptions, bridges closely relate to the notion of *weak ties*, which is a well-known concept in sociology.

### 6.4.1 The notion of weak ties in sociology

The notion of weak ties dates back to the seminal paper by Mark Granovetter in 1973 [Granovetter, 1973a]. In this work, Granovetter defines the strength of a tie between two nodes according to four criteria: the amount of time, the emotional intensity, the intimacy and the reciprocal services. Granovetter’s main thesis is that weak ties provide individuals with additional opportunities that would not be available through strong ties. More specifically, Granovetter asks people that recently changed their job how they found their new position. In most cases, people found a new employer via their personal relationships. However, these personal relationships are usually described as acquaintances rather than close friends or relatives. Since this article, many quantitative studies have been conducted in order to validate or refute Granovetter’s conclusion. A nice overview of these studies is presented in [Aral, 2016]. These articles actually diverge in their conclusions. Some of them show that weak ties provide an efficient way to access scarce information and resources while others show that strong ties are more advantageous in this respect. In 2011, Aral and Van Alstyne propose a theory to reconcile these two contradicting views. According to the *diversity-bandwidth trade-off theory*, both strong and weak ties theory are valid. Which one prevails depends on the informational context. More precisely, Aral et al show that in a context of rapidly changing information, many topics and overlapping information, then strong ties deliver more information. By contrast, in an environment with slowly changing information, few topics and less information overlap, then weak ties provide more novelty.

Despite this debate about which ones of weak or strong ties are more advantageous in order to get new information, it is commonly accepted that weak ties enable an agent to be connected to a community or group of individuals that would be otherwise unreachable. In this respect, identifying weak ties could be of interest in the context of information
spreading.

Surprisingly enough, the notion of weak and strong ties is not defined from a mathematical point of view, i.e., with a quantitative definition. However, [Easley and Kleinberg, 2010] have shown the connection between weak ties and the structure of a network. Specifically, weak ties correspond to local bridges.

### 6.4.2 Connection between bridges and weak ties

Edges between nodes in a network can be seen as a snapshot of relationships at a specific time. The connection between bridges and weak ties actually requires to think about how a network evolves. A reasonable assumption about how networks form is the *triadic closure* property. This property assumes that *if two nodes have a common neighbor, then there is a high probability that they will be connected by an edge*. In the context of friendship, [Easley and Kleinberg, 2010] give three reasons in support of this assumption. First, if the nodes B and C have a common friend A, then it creates more opportunities for B and C to spend time together, thus increasing the probability to become friends. Furthermore, having a common friend provides B and C with a reason to trust each other. Finally, there is also an incentive for A to make B and C friends in order to avoid inconsistent relationships in her circle of personal contacts.

This property can be refined by considering the strength of the ties between nodes. Based on the intensity (closeness) and frequency of the relationship between two nodes, an edge is labelled as strong or weak tie. Then the *strong triadic closure property* works as follows: *a node A violates the strong triadic closure property if it has a strong tie with both nodes B and C but B and C are not connected by an edge (either a weak or strong tie)*. Even though the strong triadic closure property looks simple, it is of interest from a modelling perspective.

Using this property, [Easley and Kleinberg, 2010] prove that a weak tie - a local property - is equivalent to a local bridge - a global and structural concept. More formally, they state the following proposition.

**Proposition 67.** [Easley and Kleinberg, 2010] If a node A satisfies the strong triadic closure property and is involved in at least two strong ties, then any local bridge it is involved in must be
a weak tie.

The proof proceeds by contradiction.

Proof. Let’s consider a node $A$ with at least two strong ties and that satisfies the strong triadic closure. Assume that $A$ has a strong tie with $B$ and let’s suppose that the edge between $A$ and $B$ is a local bridge. $A$ has at least two strong ties. The link with $B$ is one of them. Let’s suppose that the node $C$ also has a strong tie with $A$. Since the edge $A – B$ is a local bridge, then there cannot be an edge between $B$ and $C$ (recall that if we remove a local bridge between two nodes, then the distance between these two nodes must be strictly greater than 2). However, since $A$ satisfies the strong triadic closure property, there must be an edge between between $B$ and $C$. From this contradiction, we deduce that the initial statement was wrong, i.e., a strong tie between $A$ and $B$ cannot be a local bridge.

Even though the strong triadic closure property may seem simplifying, quantitative studies have shown that it is a good approximation in practise. An overview of these research is available in [Easley and Kleinberg, 2010][Chapter 3.3].

6.4.3 Limitations

In our opinion, three limitations must be noticed about the definition of weak ties. First, this is not a quantitative definition, but instead it is grounded on the closeness of friends and the frequency of their interaction. Furthermore, this a binary definition that assumes that any edge can be classified as a weak or strong tie. Finally, it does not account for the evolution of friendship relationships over time. Despite these limitations, it remains an interesting concept in order to model interaction in networks.

6.5 Conclusion

The main lesson drawn from our empirical experiments on synthetic and real data is that local bridges in networks have a high harmonic edge PageRank centrality measure. By
definition, local bridges describe the structure - a global property - of a network. Using [Easley and Kleinberg, 2010] results, we can connect local bridges to the notion of weak ties - a local property - defined in sociology. This connection makes the harmonic edge PageRank centrality measure a good candidate to identify weak ties able to spread information in networks.
Chapter 7

Conclusion and directions for future research

7.1 Summary

The starting point of this thesis was the observation that in many data, interactions between entities (i.e. nodes) are not pairwise but polyadic. While network science methods have proved to be successful in analyzing pairwise interactions (e.g., centrality measures, community detection), these methods usually do not account for higher-order interactions.

In this thesis, we have proposed two perspectives in order to analyze higher-order interactions.

First, we assume that we do not know the interactions and the objective is to learn them based on observations. The uncertainty is on the nodes and we adopt the formalism of graphical models. Our main result is to propose a criterion that enables us to distinguish between a model where there is a single cause (Figure 7-1a) and a model where there are multiple causes (Figure 7-1b).

Notice that when the causes are observed random variables, then these two models can already be distinguished by looking at the conditional independence statements. Our main contribution is to prove that this criterion also holds when the causes are hidden
Notice also that when the random variables are Gaussian, our criterion does not provide more information than existing methods. This is because the criterion that we propose is based on higher-order moments while the covariance matrix is a sufficient statistics in the Gaussian case.

Furthermore, we analyzed higher-order interactions using tools of network science. In this setting, the uncertainty is on the edges. Specifically, we conduct many experiments in order to identify the contribution of the edge PageRank centrality measure introduced in [Schaub et al., 2018]. The main observation of this investigation is that there is an intimate connection between the edge PageRank centrality measure and the notion of local bridge in networks.
7.2 Future work

From the graphical models perspective, as shown in Example 36, Theorem 35 gives a criterion for checking if random variables have a common cause or not. It would be interesting to build a test statistic based on this criterion. Furthermore, our criterion can be coupled with existing algorithms to recover a mixed graph with multi-directed edges from observational data.

From the network science perspective, it would be interesting to determine whether the 2-norm of the edge PageRank vector is the most relevant measure in order to analyze higher-order interactions in networks. Indeed, by computing the 2-norm of the edge PageRank vector, we are loosing some information because the influence of one edge on the other edges is merged into a single number. For practical reasons it is easier to work with a scalar than a vector. We have tested other alternatives, such as taking the 1-norm, infinity norm, entropy or inverse participation ratio. As for now, we haven’t identified any measure that outperforms the others.

We may also ask whether edge PageRank is an efficient measure in order to identify edges in networks that would spread contagion (disease but also information or behavior). Alternatively, from a security network perspective, it would be interesting to investigate whether edge PageRank is an efficient measure to identify the most vulnerable edges in a network.

Finally, beyond the analysis of edge PageRank centrality measure, a lot remains to be done about higher-order interactions in networks. For instance, we could think about signal processing on simplicial complexes.
Appendix of Part I
Proof of Cauchy-Binet for tensors

We prove a tensor version of the Cauchy-Binet Theorem [Broida and Williamson, 1989] for the determinant of the product $AB$, where $A$ is a matrix and $B$ is a tensor of $k$-th order. We then apply this proposition to the Tucker decomposition of tensors $\mathcal{E}(k) \bullet (I - \Lambda)^{-k}$ in the proof of Proposition 33.

**Proposition 68.** Suppose that $A$ is a $p \times n \times p \times \ldots \times p$ $k$-tensor of order $k$ and $B$ is an $\times p$ matrix, $I$ and $J$ are subsets of $\{1, \ldots, p\}$. Then $\det(AB) = \sum_{I \subseteq [n], \#I = p} \det(A_I) \det(B_I)$, where the sum is over subsets $I$ of $\{1, \ldots, n\}$ such that $\#I = p$, and $A_I$ denotes the subtensor of $A$ given by $A_{[p],i,[p],i,\ldots,[p]}$, and $B_I$ denotes the submatrix of $B$ given by $B_{I,[p]}$.

**Proof.** Note that the entry $c_{i_1 \ldots i_k}$ of the product $C = AB$ is given by:

$$c_{i_1 \ldots i_k} = \sum_{I} A_{i_1 i_3 \ldots i_k} b_{i_2}.$$  \hfill(7.1)

Let’s adopt the following notation:

- $F$ is the set of functions with domain $[n] := \{1, \ldots, n\}$ and range $[n] := \{1, \ldots, n\}$
- $G$ is the set of injective functions that map $p$ ($p < n$) elements in $[n]$ to $p$ elements in $[n]$
- $H$ is the set of strictly increasing functions that map $p$ ($p < n$) elements in $[n]$ to $p$ elements in $[n]$
- $\mathcal{S}$ is the set of permutations of the elements $\{1, \ldots, p\}$
Using Definition 29 for the tensor determinant, we have:

\[
\det(AB) = \sum_{\sigma_2, \ldots, \sigma_{k-1}} \text{sign}(\sigma_2) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} (AB)_{i\sigma_2(1) \cdots \sigma_{k-1}(i)}
\]

\[
= \sum_{\sigma_2, \ldots, \sigma_{k-1}} \text{sign}(\sigma_2) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} \sum_{l} A_{il}\sigma_3(1) \cdots \sigma_{k-1}(i)B_{l\sigma_2(i)}
\]

\[
= \sum_{\sigma_2, \ldots, \sigma_{k-1}} \text{sign}(\sigma_2) \cdots \text{sign}(\sigma_{k-1}) \prod_{f \in F} \left( \sum_{\sigma_3, \ldots, \sigma_{k-1}} \text{sign}(\sigma_3) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} A_{if(1)\sigma_3(1) \cdots \sigma_{k-1}(i)} \right) \left( \sum_{\sigma_2} \text{sign}(\sigma_2) \prod_{i=1}^{p} B_{f(1)\sigma_2(i)} \right) \det(B_f)
\]

\[
= \sum_{f \in G} \left( \sum_{\sigma_3, \ldots, \sigma_{k-1}} \text{sign}(\sigma_3) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} A_{if(1)\sigma_3(1) \cdots \sigma_{k-1}(i)} \right) \det(B_f)
\]

because \(\det(B_f) = 0\) for \(f \notin G\)

\[
= \sum_{h \in H} \sum_{\gamma \in \mathcal{G}} \left( \sum_{\sigma_3, \ldots, \sigma_{k-1}} \text{sign}(\sigma_3) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} A_{ih(\gamma(1))\sigma_3(1) \cdots \sigma_{k-1}(i)} \right) \det(B_{h(\gamma)})
\]

Since \(A_{ih(\gamma(1))\sigma_3(1) \cdots \sigma_{k-1}(i)} = (A_{i\gamma(1)\sigma_3(1) \cdots \sigma_{k-1}(i)}h)_{h}\) where \((A_{i\gamma(1)\sigma_3(1) \cdots \sigma_{k-1}(i)}h)_{h}\) is the submatrix of \(A\) with columns selected by \(h\), and \(\det(B_{h(\gamma)}) = \text{sign}(\gamma) \det(B_h)\), we have:

\[
\det(AB) = \sum_{h \in H} \left( \sum_{\gamma \in \mathcal{G}} \sum_{\sigma_3, \ldots, \sigma_{k-1}} \text{sign}(\gamma) \text{sign}(\sigma_3) \cdots \text{sign}(\sigma_{k-1}) \prod_{i=1}^{p} (A_{i\gamma(1)\sigma_3(1) \cdots \sigma_{k-1}(i)}h)_{h} \right) \det(B_h)
\]

\[
= \sum_{h \in H} \det(A_h) \det(B_h).
\]

(7.2)

\[ \Box \]
Proofs of the theorems in Section 3.3 and Section 3.4

Proof of Theorem 35

The starting point of the proof of Theorem 35 is Lemma 21 that gives the relationship between the $k$-th order cumulant tensors of the random vectors $\varepsilon$ and $X$, $\mathcal{C}^{(k)} = \mathcal{E}^{(k)} \bullet (I - \Lambda)^{-k}$. We first apply the determinant operator to each side of this equation, using the tensor version of the Cauchy-Binet theorem proved above.

Proposition 69. Consider $k$ subsets of vertices $S_1, ..., S_k \subset V$ with $#S_1 = ... = #S_k$. Then, the determinant of the tensor of $k$th-order cumulants can be written as:

$$\det \mathcal{C}^{(k)}_{S_1, ..., S_k} = \sum_{R_1, ..., R_k} \det \mathcal{E}^{(k)}_{R_1, ..., R_k} \det (I - \Lambda)^{-1}_{R_1, S_1} ... \det (I - \Lambda)^{-1}_{R_k, S_k} \quad (7.3)$$

Proof. We apply the tensor version of the Cauchy-Binet Theorem stated in Proposition 68 $k$ times to the Tucker product in equation (3.5) and we obtain equation (7.3). \qed

We can now prove Proposition 33 which gives a combinatorial expression for the determinant of the $k$-th order cumulant tensor of $X$ in terms of the DAG $G$.

Proof of Proposition 33 By Proposition 69, we have:

$$\det \mathcal{C}^{(k)}_{S_1, ..., S_k} = \sum_{R_1, ..., R_k} \det \Phi_{R_1, ..., R_k} \det(I - \Lambda)^{-1}_{R_1, S_1} ... \det(I - \Lambda)^{-1}_{R_k, S_k} \quad (7.4)$$
Additionally, from Proposition 27, we can write
\[
C_{i_1, \ldots, i_k}^{(k)} = \sum_{T \in \mathcal{T}(i_1, \ldots, i_k)} m_T, \quad (7.5)
\]
where \( m_T \) is the \( k \)-trek monomial defined by \( m_T = \phi_{\text{top}}(p_1, \ldots, p_k) \lambda^1 \cdots \lambda^k \), where \( \lambda^p = \prod_{k \rightarrow i \in P^p} \lambda^k_{ij} \).

Assuming that \( \#R_1 = \cdots = \#R_k = \#S_1 = \cdots = \#S_k = n \) and using the Leibniz expansion formula for determinants, we then get:

\[
\det C_{S_1, \ldots, S_k}^{(k)} = \sum_{\sigma_i \in \mathcal{S}_{n_i}} \left( \prod_{i=1}^{k-1} \sum_{T_i \in \mathcal{T}(S_{\sigma_i(1)}, \ldots, S_{\sigma_i(k-1)})} \cdots \sum_{T_{n_i} \in \mathcal{T}(S_{\sigma_i(n)}, \ldots, S_{\sigma_i(k-1)})}} \prod_{i=1}^{k-1} \text{sign}(\sigma_i) \prod_{x=1}^{n} m_{T_x} \right) \]

\[
= \sum_{\sigma_i \in \mathcal{S}_{n_i}} \left( \prod_{i=1}^{k-1} \sum_{T_i \in \mathcal{T}(S_{\sigma_i(1)}, \ldots, S_{\sigma_i(k-1)})} \cdots \sum_{T_{n_i} \in \mathcal{T}(S_{\sigma_i(n)}, \ldots, S_{\sigma_i(k-1)})} \prod_{x=1}^{k-1} \text{sign}(\sigma_i) \prod_{x=1}^{n} m_{T_x} \right) \]

\[
= \sum_{T \in \mathcal{T}(S_1, \ldots, S_k)} \text{sign}(T) m_T, \quad (7.6)
\]

where \( \mathcal{S}_{n_i} \) is the set of permutations of the nodes in \( S_{i+1} \), \( T \) runs over all \( k \)-trek systems between \( S_1, \ldots, S_k \) and \( \text{sign}(T) = \text{sign}(\sigma_1) \cdots \text{sign}(\sigma_{k-1}) \). In this expression, we have \( m_T = \prod_{x=1}^{k} m_{T_x} \), where \( m_{T_x} = \Phi_{\text{top}}(P_{S_1}, \ldots, P_{S_{k-1}}) \lambda^{P_1} \cdots \lambda^{P_{k-1}} \) is the trek monomial corresponding to the trek \( T_x \).

In order to now prove the equation (3.9)
\[
\det C_{S_1, \ldots, S_k}^{(k)} = \sum_{T \in \mathcal{T}(S_1, \ldots, S_k)} \text{sign}(T) m_T,
\]
we need to show that \( m_T = 0 \) when \( T \) is a trek-system between \( S_1, \ldots, S_k \) with a sided intersection. A nice tool to proof this is given by the Gessel-Viennot-Lindstrom Lemma. This Lemma is actually a cornerstone for our paper since we use it again in the proof of our main Theorem 35 itself.

**Lemma 70** ([Gessel and Viennot, 1985][Lindström, 1973]). Suppose \( G \) is a directed acyclic
graph with vertex set \([p] = \{1, \ldots, p\}\). Let \(R\) and \(S\) be subsets of \([p]\) with \(#R = #S = n\). Then,

\[
\det (I - \Lambda)^{-1}_{R,S} = \sum_{P \in N(R,S)} (-1)^P \lambda^P,
\]

where \(N(R, S)\) is the set of all collections of nonintersecting systems of \(n\) directed paths in \(G\) from \(R\) to \(S\), and \((-1)^P\) is the sign of the induced permutation of elements from \(R\) to \(S\). In particular, \(\det (I - \Lambda)^{-1}_{R,S}\) is identically if and only if every system of \(n\) directed paths from \(R\) to \(S\) has two paths which share a vertex.

Therefore, only the terms that involve trek-systems without sided intersection remain in equation (7.6), which completes the proof of equation (3.9).

We need a last intermediary result before proving Theorem 35.

**Lemma 71.** Consider \(k\) subsets of vertices \(S_1, \ldots, S_k \subseteq V\) with \(#S_1 = \ldots = #S_k = n\). Then, \(\det C_{S_1,\ldots,S_k}^{(k)}\) is identically 0 if and only if for any set \(R \subseteq V\) such that \(#R = n\), there exists \(i \in \{1, \ldots, k\}\) with \(\det (I - \Lambda)^{-1}_{R,S_i} = 0\).

**Proof.** Let us suppose that \(\det C_{S_1,\ldots,S_k}^{(k)}\) is identically 0. Using equation (21) and the Cauchy-Binet theorem, we compute \(\det C_{S_1,\ldots,S_k}^{(k)}\). We obtain equation (7.3):

\[
\det C_{S_1,\ldots,S_k}^{(k)} = \sum_{R_1,\ldots,R_k} \det \mathcal{E}_{R_1,\ldots,R_k}^{(k)} \det (I - \Lambda)^{-1}_{R_1,S_1} \cdots \det (I - \Lambda)^{-1}_{R_k,S_k},
\]

where the sum runs over subsets \(R_1, \ldots, R_k\) of \(V\) of cardinality \(#R_1 = \ldots = R_k = #S_1 = \ldots = #S_k = n\). However, since \(\mathcal{E}^{(k)}\) is a diagonal tensor, \(\det \mathcal{E}_{R_1,\ldots,R_k}^{(k)} = 0\) unless \(R_1 = R_2 = \ldots = R_k = R\). In this case, denoting \(\det \mathcal{E}_{R_1,\ldots,R_k}^{(k)} = \det \mathcal{E}_R^{(k)}\) we get:

\[
\det C_{S_1,\ldots,S_k}^{(k)} = \sum_{R \subseteq V} \det \mathcal{E}_R^{(k)} \det (I - \Lambda)^{-1}_{R,S_1} \cdots \det (I - \Lambda)^{-1}_{R,S_k}, \quad (7.7)
\]

Each monomial \(\det \mathcal{E}_R^{(k)}\) appears only once, therefore for any set \(R\) satisfying \(#R = #S_1 = \ldots = #S_k = n\), there exists \(i \in \{1, \ldots, k\}\) such that \(\det (I - \Lambda)^{-1}_{R,S_i} = 0\), which proves the if-direction.

Let us now suppose that for any set \(R\) satisfying \(#R = #S_1 = \ldots = #S_k = n\), there exists
\( i \in \{1, \ldots, k\} \) such that \( \det(I - \Lambda)^{-1}_{R, S_i} = 0 \). Then from the expression (7.3), we conclude that \( \det C_{S_1, \ldots, S_k}^{(k)} = 0 \). \( \Box \)

Equipped with the Gessel-Viennot-Lindstrom Lemma (Lemma 70) and Lemma 71, we are now ready to prove our main theorem.

**Proof of Theorem 35** Let us first suppose that \( \det C_{S_1, \ldots, S_k}^{(k)} = 0 \) and let \( T \) be a \( k \)-trek system between \( S_1, \ldots, S_k \).

- If all elements of the multiset \( \text{top}(T) \) are distinct, then Lemma 71 implies that there exists an integer \( i \in [k] \) such that \( \det(I - \Lambda)^{-1}_{\text{top}(T), S_i} = 0 \). By Lemma 70, the path system from \( \text{top}(T) \) to \( S_i \) has a sided intersection, therefore \( T \) has a sided intersection.

- If \( \text{top}(T) \) has repeated elements, then at least two \( k \)-treks in \( T \) intersect, namely at their top, hence \( T \) has a sided intersection.

Conversely, let’s suppose that every \( k \)-trek system \( (T) \) between \( S_1, \ldots, S_k \) has a sided intersection. Let’s consider any set \( R \subseteq V \) that satisfies \( \#R = \#S_1 = \ldots = \#S_k \).

- If \( R \) forms the top of a \( k \)-trek system that ends at \( S_1, \ldots, S_k \), then there exists at least one integer \( i \in [k] \) such that there is a sided intersection in the path system between \( R \) and \( S_i \). By the Gessel-Viennot-Lindstrom Lemma 70, \( \det(I - \Lambda)^{-1}_{R, S_1} \cdots \det(I - \Lambda)^{-1}_{R, S_k} = 0 \) and therefore \( \det \mathcal{E}_R^{(k)} \det(I - \Lambda)^{-1}_{R, S_1} \cdots \det(I - \Lambda)^{-1}_{R, S_k} = 0 \).

- Alternatively, if \( R \) does not form the top of a \( k \)-trek system that ends at \( S_1, \ldots, S_k \), then there is no \( k \)-path system from \( R \) to \( S_1, \ldots, S_k \), i.e., there is no path system between \( R \) and at least one of \( S_1, \ldots, S_k \). This implies that at least one of \( \det(I - \Lambda)^{-1}_{R, S_i} = 0 \) and therefore \( \det \mathcal{E}_R^{(k)} \det(I - \Lambda)^{-1}_{R, S_1} \cdots \det(I - \Lambda)^{-1}_{R, S_k} = 0 \).

Thus, \( \det C_{S_1, \ldots, S_k}^{(k)} = \sum_{R \subseteq V} \det \mathcal{E}_R^{(k)} \det(I - \Lambda)^{-1}_{R, S_1} \cdots \det(I - \Lambda)^{-1}_{R, S_k} = 0 \). \( \Box \)

**Proof of Theorem 45**

Theorem 45 extends Theorem 35 to the case of mixed graphs. We use a common argument in the graphical models literature that enables us to convert the multidirected case to the
directed case. We replace every multidirected edge joining $i_1, ..., i_k$ with a vertex $v$ and a
directed edge between $v$ and any of the vertices $i_1, ..., i_k$. We call this graph $\tilde{G}$, also known
as the canonical DAG associated to $G$.  

![Figure 7-3: Mixed graph $G$ in Figure 7-3a and its corresponding canonical DAG $\tilde{G}$ in Figure 7-3b](image)

**Definition 72.** The linear structural equation model $X_j = \sum_{i \in \text{pa}(i)} \lambda_{ij} X_j + \epsilon_j$ given by a
mixed graph $G = (V, \mathcal{D}, \mathcal{H})$, where $V = [p]$, is the family of distributions on $\mathbb{R}^p$ with
tensor of $k^{th}$-order cumulants in the set:

$$\mathcal{M}(k)(G) = \{ \mathcal{E}(k) \cdot (I - \Lambda)^{-k} : \Lambda = (\lambda_{ij}) \in \mathbb{R}^{\mathcal{D}}, \mathcal{E}(k) \in (\mathbb{R}^{\otimes k})^H \} (7.8)$$

where $(\mathbb{R}^{\otimes k})$  

H denotes the set of $p \times \cdots \times p$ ($k$-times) cumulant tensors of $\epsilon$, which are zero at all en-
tries $(i_1, \ldots, i_k)$ unless $i_1 = \cdots = i_k$ or there exists a multi-directed edge $(j_1, \ldots, j_\ell) \in \mathcal{H}$
such that $\{i_1, \ldots, i_k\} \subseteq \{j_1, \ldots, j_\ell\}$

**Proposition 73.** Let $S_1, ..., S_k \subset V$ be $k$ sets of vertices such that $\#S_1 = \cdots = \#S_k$. Then the
determinant of $\mathcal{C}(k)_{S_1, \ldots, S_k}$ is zero for all cumulant tensors $\mathcal{C} \in \mathcal{M}(k)(G)$ if and only if the determinant
of $\mathcal{C}(k)_{\tilde{S}_1, \ldots, \tilde{S}_k}$ is zero for all cumulant tensors $\tilde{\mathcal{C}} \in \mathcal{M}(k)(\tilde{G})$. In other words, if there is no $k$-trek
system without sided intersection between $S_1, ..., S_k$ in $G$, then there is no $k$-trek system without
sided intersection between $S_1, ..., S_k$ in $\tilde{G}$, and vice-versa.

109
Proof. Let $\mathcal{M}^{(k)}(G)$ and $\mathcal{M}^{(k)}(\tilde{G})$ be the sets of $k^{th}$-order cumulant tensors of $G$ and $\tilde{G}$, respectively. We will prove that $\mathcal{M}^{(k)}(G)$ and $\mathcal{M}^{(k)}(\tilde{G})$ have the same Zariski closure, i.e., an equation vanishes on $\mathcal{M}^{(k)}(G)$ if and only if it vanishes on $\mathcal{M}^{(k)}(\tilde{G})$. To do this, let us show that the two parametrizations give the same family of tensors near the identity tensor. Note that there exist distributions of independent variables $X_1, ..., X_p$ for which the $k^{th}$-order cumulant is the identity tensor. As shown in Proposition 51, $C_{i_1, ..., i_k}^{(k)}$ can be expressed as a sum of the trek monomials of all $k$-treks in $\mathcal{T}(i_1, ..., i_k)$ between $i_1, ..., i_k$ in $G$ as $C_{i_1, ..., i_k}^{(k)} = \sum_{(p_1, ..., p_k) \in \mathcal{T}(i_1, ..., i_k)} \phi_{\text{top}(p_1, ..., p_k)} \lambda_1^{p_1} ... \lambda_k^{p_k}$. Similarly, $\tilde{C}_{i_1, ..., i_k}^{(k)}$ is the sum of the trek monomials of all $k$-treks in $\tilde{\mathcal{T}}(i_1, ..., i_k)$ between $i_1, ..., i_k$ in $\tilde{G}$, i.e., $\tilde{C}_{i_1, ..., i_k}^{(k)} = \sum_{(\tilde{p}_1, ..., \tilde{p}_k) \in \tilde{\mathcal{T}}(i_1, ..., i_k)} \tilde{\phi}_{\text{top}(\tilde{p}_1, ..., \tilde{p}_k)} \tilde{\lambda}_1^{\tilde{p}_1} ... \tilde{\lambda}_k^{\tilde{p}_k}$.

Now, let’s set

$$E_{i_1, ..., i_k}^{(k)} = \tilde{E}_{v, ..., v}^{(k)} \tilde{\lambda}_{v, i_1} ... \tilde{\lambda}_{v, i_k}$$ \hspace{1cm} (7.9)

if there is a multi-directed edge between $i_1, ..., i_k$ in $G$, i.e., there is a directed edge from the vertex $v$ to each of the nodes $i_l, l \in \{1, ..., k\}$ in $\tilde{G}$, and let

$$E_{i, ..., i}^{(k)} = \tilde{E}_{i, ..., i}^{(k)} + \sum_{w \in \mathcal{H}} \tilde{E}_{v, ..., v}^{(k)} \tilde{\lambda}_{v, i}^k$$ \hspace{1cm} (7.10)

for each $i \in \{p\}$. As we initially assumed we are near the identity tensor $I^{(k)}$, we can switch from one parametrization to the other as follows. First, note that $I^{(k)} \in \mathcal{M}^{(k)}(G)$ and $I^{(k)} \in \mathcal{M}^{(k)}(\tilde{G})$. Then, given $\tilde{E}^{(k)}$ and $\tilde{\lambda}$, we can find $E^{(k)}$ from equations (7.9) and (7.10), and $\tilde{\lambda}_{ij} = \tilde{\lambda}_{ij}$ for $i \to j$, i.e., given $\tilde{C}^{(k)} \in \mathcal{M}^{(k)}(\tilde{G})$, using equation (7.9) and equation (7.10), we get that the corresponding $C^{(k)} \in \mathcal{M}^{(k)}(G)$, therefore $\mathcal{M}^{(k)}(\tilde{G}) \subseteq \mathcal{M}^{(k)}(G)$. Conversely, given $E_{i_1, ..., i_k}^{(k)}$ and $\lambda_{ij}$ small enough, we can choose $\tilde{E}_{v, ..., v}^{(k)} = \varepsilon > 0$, $\tilde{\lambda}_{ij} = \lambda_{ij}$ for $i \to j$ in $\mathcal{D}$, and $\tilde{\lambda}_{v, i} = \sqrt{\frac{|E_{i_1, ..., i_k}^{(k)}|}{\varepsilon}}$ for $l \in \{1, ..., k\}$. Therefore, equation (7.9) is satisfied. Since $E_{i_1, ..., i_k}^{(k)}$ is small and $E_{i, ..., i}^{(k)}$ is near 1, we can find $\tilde{E}_{i, ..., i}^{(k)} > 0$ such that equation (7.10) is satisfied and such that $\tilde{E}^{(k)}$ is a diagonal tensor. This shows that if $C^{(k)} \in \mathcal{M}^{(k)}(G)$ is in a neighborhood of $I^{(k)}$, then we can find the corresponding $\tilde{C}^{(k)} \in \mathcal{M}^{(k)}(\tilde{G})$ as well.
Note that this correspondence between $\mathcal{C}^{(k)}$ and $\tilde{\mathcal{C}}^{(k)}$ is a bijection. Thus $\mathcal{M}^{(k)}(G)$ and $\mathcal{M}^{(k)}(\tilde{G})$ are equal in an open neighborhood and so they have the same Zariski closure, i.e., $\overline{\mathcal{M}^{(k)}(G)} = \overline{\mathcal{M}^{(k)}(\tilde{G})}$, where $\overline{\mathcal{M}^{(k)}(G)}$ denotes the Zariski closure of $\mathcal{M}^{(k)}(G)$. Therefore the determinant of $\tilde{\mathcal{C}}^{(k)}_{S_1, \ldots, S_k}$ vanishes on $\mathcal{M}^{(k)}(\tilde{G})$ if and only if the determinant of $\mathcal{C}^{(k)}_{S_1, \ldots, S_k}$ vanishes on $\mathcal{M}^{(k)}(G)$. This is equivalent to saying that there is no system of $k$-treks without sided intersection between $S_1, \ldots, S_k$ in $\tilde{G}$ if and only if there is no system of $k$-treks without sided intersection between $S_1, \ldots, S_k$ in $G$.

Going back to the proof of Theorem 45, Proposition 73 enables us to reduce the multi-directed setting to a directed acyclic graph for which we proved the Theorem 35.
Proofs of the results in Section 3.5

In this section, we prove our results in the case of higher-order moment tensors. Proofs are analogous to the proofs for higher-order cumulant tensors.

Proof of Proposition 51. We have that \( \mathcal{N}^{(k)} = \Phi^{(k)} \bullet (I - \Lambda)^{-k} \). The entries of \( \Phi^{(k)} \) are non-zero if and only if they are of the form \( \mathbb{E}[\varepsilon_1^{x_1} \ldots \varepsilon_k^{x_k}] \) where \( x_1, \ldots, x_k \) are integers either equal to 0, or strictly greater than 1. Such entries correspond to the cases when there is a \( k \)-split-trek between \( i_1, \ldots, i_k \) as defined in Definition 48. Furthermore, we have \( (I - \Lambda)^{-1}_{ij} = \sum_{p \in \mathcal{P}(ij)} \lambda^p \) by equation (3.6), and replacing this expression in equation (3.11), we obtain equation (3.12), which completes the proof.

Proof of Proposition 52. By applying the tensor version of the Cauchy-Binet Theorem \( k \) times to equation (3.11), we get:

\[
\det \mathcal{N}^{(k)}_{S_1, \ldots, S_k} = \sum_{R_1, \ldots, R_k} \det \Phi^{(k)}_{R_1, \ldots, R_k} \det(I - \Lambda)^{-1}_{R_1, S_1} \ldots \det(I - \Lambda)^{-1}_{R_k, S_k} \quad (7.11)
\]

Additionally, from equation (7.11), we can write

\[
\mathcal{N}^{(k)}_{i_1, \ldots, i_k} = \sum_{T \in \sigma(i_1, \ldots, i_k)} \text{sign}(T) m_T, \quad (7.12)
\]

where \( m_T \) is the \( k \)-split-trek monomial defined by \( m_T = \phi_{\text{top}(P_1, \ldots, P_k)} \lambda^{P_1} \ldots \lambda^{P_k} \).
Assuming that \( R_1 = ... = R_k = S_1 = ... = S_k = n \), we then get:

\[
\det \mathcal{N}_{S_1,\ldots,S_k}^{(k)} = \sum_{\sigma \in \mathfrak{S}_{n_1} \cdot \ldots \cdot \mathfrak{S}_{n_k-1}} \left( \sum_{T_j \in \mathcal{S}(s_j, s_{\sigma_1}(j), \ldots, s_{\sigma_k-1}(j))} \prod_{l=1}^{k-1} \text{sign}(\sigma_l) \prod_{s=1}^{n} m_{T_s} \right)
\]

\[
= \sum_{T \in \mathcal{S}(S_1,\ldots,S_k)} \text{sign}(T) \ m_T
\]

where \( \mathfrak{S}_{n_i} \) is the set of permutations of the nodes in \( S_{i+1} \), \( T \) runs over all \( k \)-split-trek systems between \( S_1,\ldots,S_k \) and \( \text{sign}(T) = \prod_{k-1} \text{sign}(\sigma_l) \). In this expression, we have \( m_T = \prod_{x=1}^{n} m_{T_x} \), where \( m_{T_x} = \Phi_{\top}(p_1,\ldots,p_k) \lambda^{P_1} \ldots \lambda^{P_k} \), i.e., \( m_T \) is the product of the monomials of the \( n \) \( k \)-split-treks that form the \( k \)-split-trek-system \( T \).

Similarly to the proof of Proposition 33, we can use the Gessel-Viennot-Lindstrom Lemma to show that the sum in the expression of \( \det \mathcal{N}_{S_1,\ldots,S_k}^{(k)} \) can be taken over the set \( \tilde{\mathcal{S}}(S_1,\ldots,S_k) \) of \( k \)-split-trek systems between \( S_1,\ldots,S_k \) without sided intersection.

In order to prove Theorem 53, notice that in equation (7.11), \( \Phi^{(k)} \) is diagonal for \( k = 3 \). Furthermore, a 3-trek and a 3-split-trek are the same (and similarly for a 3-trek system and a 3-split-trek-system). Then the proof of Theorem 53 follows the same reasoning as for Theorem 35.

Notice finally that for \( k > 3 \), \( \Phi^{(k)} \) is not diagonal, and for that reason, we cannot easily extend Theorem 53 to higher-order moments.
Part II

An Econometric Analysis of Yield Gaps in Semi-Arid Tropics of India
Chapter 8

Introduction

India was the birth place of the Green Revolution in the 1970’s. This technological rupture enabled the agricultural sector to experience a dramatic increase in farm yields, thus allowing the country to reach grain independence. Nowadays, however, 52% of the population in India still work in the agricultural sector and many rural households in Semi-Arid Tropics still experience negative profits when growing crops. Low productivity has been identified as one of the main reasons for the economic deadlock of the farming sector in rural areas. For a long time, research has focused on yield gaps considering crop yields as a critical aspect of subsistence and economic well-being for local communities in developing countries. In particular, the main question remains:

**What are the most relevant explanatory variables of yield gaps?**

In this research, we examine this question using a panel dataset to identify the reason why yield gap originates for the particular of Semi-Arid Tropics (SAT) in rural India.

Having stated this scientific objective, it is necessary to start with a clear definition. The literature is not unanimous about the concept of yield gap. Most papers (Lobell et al. 2008 among others) refer to yield gap as the difference between the potential yield and the actual yield of farmers. The potential yield is given by crop models. These latter are obtained via simulations based on deterministic equations and consider ideal physical conditions in terms of solar radiation, temperature and water for a given geographical
location. Since the physical conditions on the field differ from crop models’ framework, yield gap has been also defined as the difference between the attainable yield and farmers’ actual yield. The attainable yield takes into account the biophysical constraint on the field. In other words, the attainable yield is measured with a field experiment.

However, these two approaches present some limitations for two reasons. First, field conditions considered by the crop models may be far from the true conditions experienced by households on their everyday life. Second, these models may be unable to take into account the intrinsic unobservable households’ heterogeneity, in particular characteristics stemming from individual behavior.

For these reasons, we rather adopt an alternative definition from the literature that measures the output gap between each individual household and the maximum observed output. In particular, we propose to take the maximum observed yield at the village level. We call this reference the Maximum Local Yield (MLY). In that respect, the yield of reference is effectively attained by some economic agents. Furthermore, since the reference is taken at the village level, this method enables us to take into consideration the diversity of physical conditions, in particular rain and soil characteristics and to account for a village fixed effect.

Based on this definition, this analysis seeks at answering two questions:

1. What are the variables that could explain the yield gap between the most and lowest productive farmers for a given village?

2. What are the variables that are likely to increase individual farmers’ output over time?

This part is organized as follows. In Chapter 9, we review the lessons from the literature related to yield gaps and behavioral economics that are relevant for our study. In Chapter 10, we provide a description and an overview of the dataset that we study. In Chapter 11, we conduct an econometric analysis of the dataset using a fixed effects mod-
els and we focus on finding the explanatory variables of farming output rate. In Chapter 12, we pursue the quantitative analysis of data by using econometric methods and we focus on the dynamics of households’ productivity. We conclude with a summary of our contributions, we propose policy recommendations and suggest open research questions for future work. We notice to the reader that Chapter 11 and Chapter 12 can be read as stand-alone papers.
Chapter 9

Literature review

9.1 Literature review of yield gap

Many previous studies focusing on yield gaps have shown a large spread in yield gaps and a range from 20 to 80% of the potential yield is considered to be a good estimation that accounts for measurement errors [Lobell et al., 2009]. Although usually small, yield gaps measurements vary depending on the yield taken as the reference. Water availability and irrigated crop as opposed to rain fed agriculture have a critical impact on yield gaps. For that reason, literature usually also distinguishes between the simulated yield gap estimated for irrigated land and the water-limited rain fed potential yield for rain fed regions. The profusion of yields of reference and the distinction between yields for irrigated versus rain fed agriculture raise concerns about the comparability of yield gaps from different studies. This key point highlights the importance of the measurement method for a given location.

Different methods have been commonly adopted to measure yield gaps and identify their origins. Following [Lobell et al., 2009], a first method consists in controlled field experiments that enable researchers to compare the impact of different agricultural practices and control for land characteristics. In particular, the International Rice Economic Network’s experiment (IREN) conducted in 1970 found that characteristics specific to the field and biotic factors like soil characteristics were the most important drivers of yield
gaps. The “field-specific factors” [Lobell et al., 2009] refer to farmers’ agricultural practices. In our study and relying on concepts from econometrics, we may also talk about households’ heterogeneity. In addition to that, this study found that high quality seeds, fertilizers and pesticides help farmers to improve their yields. However, the additional yield obtained thanks to improved inputs is hardly justified in terms of farming economic profitability.

Alternatively, other research are based on empirical studies. Based on surveys and or remote sensing data, these studies estimate the statistical correlations between yields and variables measuring soil characteristics, climatic parameters and management. For instance, water constraints [Calvino and Sadras, 2002], the amount of fertilizer and timing of irrigation [Lobell et al., 2009] were found to be the main drivers of yield gaps for the given location of these respective studies. Other studies are based on model simulations and aim at estimating the gains from changing farming parameters such as the sowing and harvesting periods [Yang et al., 2004, Lobell et al., 2005]. Incidentally, the comparison of simulated yields obtained under different initial parameters can shed light on farming practices used by households on the field. Finally, econometricians have investigated the question of yield gap via the price of output provided by time-series data sets (for instance, [Keeney and TW.Hertel, 2014]). More specifically, these studies are based on the elasticity of the crop yield with respect to the price of the crop output. The analysis is grounded on the underlying assumption that a high price-elasticity means that management practices, and in particular the use of inputs, are the main explanatory variables of yield gaps. On the contrary, a low yield response to the output price would show that inputs are not the constraining factors.

As far as yield gaps in India are concerned, an important study led by the Indian Council of Agricultural Research (ICAR) reported an average rice yield gap equal to 52.3% for the whole country over the period 1990/1991 to 1997/1998 (FAO, ICAR, 2000). This yield gap is measured in comparison to the field experimental yield given by the Frontline Demonstration Program launched in 1990 to raise farmers’ awareness about
yield gaps. It ranges from 15.6% for Tamil Nadu up to 75.6% in Rajasthan.

Generally speaking, the main biophysical variables usually put forward to explain yield gaps are water constraints, the low quality of seeds, climatic events such as floods and droughts, pest attacks and diseases, and problematic soils, in particular erosion, salinity, acidity and nutrient deficiency. In addition to that, socio-economic factors have to be taken into account such as low implementation of irrigation, the weak diffusion of high yield variety seeds, the overuse or underuse of fertilizer, price market variability of inputs and outputs, low accessibility to financial support, labor shortage, suboptimal management practices and the lack of knowledge and/or information.

Our study builds upon the previous developments of the yield gaps literature. Furthermore, our analysis brings to the existing body of literature a methodological contribution. As explained in the next section, we combine the two following approaches: the identification of statistical relationships between farming output rate and agronomic variables on one hand, and econometrics methods, in particular panel data analysis, on the other hand. Such a combination allows us to handle both biophysics factors and socio-economic parameters while accounting for the heterogeneity of cultivated land and households.

9.2 Literature review of heuristics and judgements in behavioral economics

The foundations of the theory of heuristics and biases are grounded on Tversky and Kahneman research [Tversky and Kahneman, 1971, Tversky and Kahneman, 1974]. In contrast with the classic economic theory, this theoretical framework assumes that individuals are not fully rational and in that respect, do not necessarily comply with Bayesian rules. Nevertheless, it does not mean that people adopt random behaviors. Rather, they select information and rely on rules of thumb or heuristics. These heuristics are generally
fast, frugal and usually require few computations. In this regard, they may contradict optimization as defined by the classical economic theory. Some research have however found that they might perform as well as more complex statistical models in a context of uncertainty [Gigerenzer et al., 1999].

The representativeness heuristic arises when people apply an equivalent of “law of large numbers” even in the presence of a small size sample. We then could talk about a ‘law of small numbers’. More precisely, for any sample size, individuals associate the population mean to the sample mean [Camerer, 1987, C.F., 1990, Camerer and Thaler, 1995]. In the case of farmers, it could happen when they draw some general rules about rainfall based only on the most recent past years that could have been different from the general trend.

The ‘gambler’s fallacy’ is characterized by a belief in negative autocorrelation, also known as ‘alternative bias’ or ‘negative recency’ [Tversky and Kahneman, 1974]. The basic example usually used to illustrate this heuristic is the reluctance of people to bet on a number in a lottery that has just appeared. Such a behavior violates the assumption of rationality since the conditional mathematical probability of a previously winning number has not changed. In the farming context, the failure of a crop one year could be seen as a motivation to no longer grow this crop.

By contrast to the ‘gambler’s fallacy’, the ‘hot hands fallacy’ characterizes by a belief in positive autocorrelation. More particularly, an event is attributed a higher probability when it has been recently frequently observed. This heuristic usually occurs in sport when a team or a player scores extensively. In particular, this heuristic is usually observed when the stochastic process is inanimate. In that respect, stores that have sold a winning lotto ticket have higher sales in the following periods, up to 40 weeks after the sale of the winning ticket [Guryan and Kearney, 2008]. As agriculture is concerned, farmers could be tempted to grow the same crop when they got a good yield the previous year. By contrast, agronomists would tend to favor crop rotation.
As it might be guessed, heuristics are usually intertwined. Such is the case with the three rules of thumb mentioned above: the ‘gambler’s fallacy’ and the ‘hot hands fallacy’ are grounded on the representativeness heuristic. The conjunction fallacy occurs when people assign a higher probability to the intersection of two events, compared to the individual probability of each event. The seminal example is the Linda problem presented by [Tversky and Kahneman, 1983] in which people are asked to describe the profession of a fictive lady called Linda. After being presented with some information about the fictive character, people choose a combination of individual descriptions instead of separate alternatives. Such a choice contradicts the mathematical rule that the probability of the intersection cannot be higher than the probability of each individual event included in the intersection. Some research has nevertheless stated that the strength of this heuristic is alleviated when people are presented with natural frequencies instead of probabilities [Hertwig and Gigerenzer, 1999], or when people take decisions in group of two or three persons [Charness et al., 2010]. In the farming context, it could happen if farmers assume that the probability of flooding and pest attack is higher than each individual event.

The availability heuristic occurs when individuals assess the probability of an event based on the facility to represent and come up with this event in their mind. In that context, media and advertisement have a critical role in the assignment of a probability [Lichtenstein et al., 1978, Eisenman, 1993]. Some rare events might prove to be more probable to people because they are largely related in newspapers and TV shows. For instance, [Lichtenstein et al., 1978] have shown a strong relationship between people’s assumed frequencies of death by different means and the availability of this events in minds. Similarly, we could think that farmers are more upset by pest attacks or flooding when such events are heavily related on radio for instance, even so such events occurs in other geographical regions.

[Zajonc, 1980, Slovic et al., 2002, Loewenstein et al., 2001] have identified the affect heuristic according to which individuals classify and assess events based on emotions.
As Kahneman highlights it [Kahneman, 2011], this heuristic has the peculiarity to substitute the question ‘What do I think about it?’ with the more relaxed one ‘what do I feel about it?’. Such a heuristic is particularly relevant in the case of hazards, for instance with smoking or nuclear energy [Fischhoff et al., 1978, Finucane et al., 2000]. Similarly, this heuristic applies to farmers, for instance with floods or droughts.

Anchoring has been identified as one the most widespread heuristic. It arises when people base their decision or behavior on an a priori assumption called ‘anchor’. In addition to influence or even bias the decision, the anchor is not necessarily appropriate. Furthermore, an anchor may also prove to slow the adaptation of people’s decision to their target, in particular when it is self-generated [Epley and Gilovich, 2001]. Tversky and Kahneman have provided the fundamental developments in that domain by studying numerical anchors. In particular, they developed a famous experiment consisting in turning a wheel with numbers from 1 to 100 but constrained to display 10 or 65. After turning the wheel, people are asked to estimate whether the number of African countries at the UN is greater or lower that the anchor given by the wheel (experiment called ‘comparative judgment task’). A subsequent question consists in estimating this number (called ‘absolute judgment task’). Such anchor could occur in agriculture when some specific numbers are communicated to people (for instance output rate to be attained, quantity of fertilizer) and they finally stick to this anchor.

Individuals might turn to adopt inconsistent behavior in case of probabilistic information that results in base rate neglect. This effect is the propensity to ignore the importance of a specific characteristic in the global population. A famous example is given by the car problem [Kahneman and Tversky, 1972]. In this problem, cars in one city are either green (85%) or blue (15%). An accident involving a cab occurs and a witness that can be trusted at 80% level, testifies that it was blue. According to Bayes’ rule (conditional probability), the probability that the accident involved a blue cab is 41.4% but experimental answer is 80% because people miss to take into account the base rate. Similarly, pest attack by an insect A could be overestimated compared to the Bayes’ rule, if we do not take into
account the frequency of this insect.

Research has shown that this underweighting behavior is still true even when the base rate is high [Icek and Fishbein, 1977, Bar-Hillel, 1980], when individuals are educated [Casscells et al., 1978], or when people display a higher level of measured intelligence [Stanovich and West, 2000].

The reverse of the base rate neglect also exists and occurs when people underestimate the likelihood of a sample [Camerer and Thaler, 1995]. The heuristic called hindsight bias reveals when people states a posteriori that the outcomes fit with their initial prediction. Hindsight bias is a particular case in which people misperceive their own past predictions. [Kahneman, 2011] explains such a phenomenon by the weak ability of human mind to mobilize past beliefs when opinion has changed.

“A general limitation of the human mind is its imperfect ability to reconstruct past states of knowledge, or beliefs that have changed. Once you adopt a new view of the world (or any part of it), you immediately lose much of your ability to recall what you used to believe before your mind changed.” [Kahneman, 2011], p. 201, quoted in [Dhami, 2016], p 1387).

Two other reasons may explain this heuristic:

- A mean to give an explanation to unpredictable events and in that case, this heuristic can be seen as a coping mechanism [Dhami, 2016].
- A way to demonstrate one’s skills to others [Rachlinski, 1998].

An individual in accordance with the neoclassic economic theory should verify:

\[ E[X|I_t] = E[E[X|I_t]|I_{t+1}] \]  

(9.1)

Where \( E[X|I_t] \) is the individual’s expectation about \( X \) knowing the information \( I_t \). The left-hand side of the equation is the predictive judgment while the right-hand side is the
remembered prediction also called postdictive judgement. This equality says that a rational individual should display consistent judgements between the present and the past. The hindsight bias or creeping determinism \cite{Fischhoff, 1975} arises when this inequality is violated. In that case, the individual bends towards the current result, which \cite{Camerer et al., 1989} mathematically represent by the following equation:

\[
E[E[X|I_t]|I_{t+1}] = aX + (1 - a)E[X|I_t]
\]  

The individual is fully biased when \(a = 1\) and has no hindsight bias when \(a = 0\). This bias is also called the “I knew it all along” effect. In the farming context, we can think that such a bias is quite common when people try to explain a posteriori crop success or failure. In particular, such behavior could be increased by the willingness to get respectability from one’s peers in the village.

Confirmation bias is another critical rule of thumb that could be relevant for our project. This heuristic, also called biased assimilation in psychology, consists in selecting interpretation of evidence that allow to back one’s preliminary statements. This phenomenon is more frequently observed in the following cases \cite{Rabin and Schrag, 1999}:

- The information is ambiguous and can imply different interpretations
- The pooling and analysis of information requires a selection process which is also named by a hypothesis-based filtering
- The task assigned to survey people consists in assessing the correlation between events that occur at different periods in time
- The judgement requires individuals to aggregate information provided by different sources

\cite{Stanovich and West, 2007} have shown that confirmation bias is not reduced by higher cognitive skills or open-mindedness. In the farming context, confirmation biases could happen in order to rationalize a crop failure, thus subsequently conducting to inefficient
farm practices (quantity of fertilizer and/or pesticide for instance).

Two particular frameworks of behavioral economics that include some of the heuristics presented above, have recently received deeper developments: mental models and bounded rationality.

Mental models are structures of beliefs that people form about the world. In particular, they include causal relations and the classification of various information into limited categories. Mental models have a particular role in reducing cognitive costs. Individually integrated, they may be transmitted from generation to generation and in that respect, might prove to be inertial. Depending on their correctness, mental models can support optimal decisions or on the contrary, can undermine them. These mental models may stem from individuals themselves or be the result of history [Nunn, 2008, Alesina et al., 2013, Acemoglu et al., 2013]. A common definition of mental models, based on [Denzau and North, 1994], given in the World Development Report (2015) and quoted in [Dhami, 2016] (p. 1407) states that: “When people think, they generally do not draw on concepts that they have invented themselves. Instead, they use concepts, categories, identities, prototypes, stereotypes, causal narratives, and worldviews drawn from their communities.”

It is highly critical to understand the persistence of mental models because some of them might prove to be non-optimal or outdated. Many reasons have been put forward to explain the continuation of mental models:

- Mental models may reinforce anchors and for that reason, they may be highly mobilized by individuals
- The availability heuristic could be responsible to prioritize mental models in mind when choices have to be made
- Confirmation bias could foster people to select information that comply with their models, and on the contrary, neglect information that contradict them
Finally, social identity could have a critical impact because align ones’ decision with the mental models of one’s social group could be a way to signal one’s affiliation to other members.

Mental models could be responsible of reducing the number of alternatives, thus canceling the possibility of these alternatives to challenge the current mental models.

As highlighted by [Dhami, 2016] (p. 1409), mental models are different than social norms because they are not socially imposed, but rather, they are “broad ideas about how the world works and one’s place in it” [Bank, 2015]. The comparison with the performance of a group of reference may be an effective tool to incentivize people to change their mental views. For instance, people, when shown with the electricity consumption of a peer group, reduce their own consumption by the same proportion as when prices are increased by 20% [Allcott and Mullainathan, 2010, Allcott and Rogers, 2014]. A comparable phenomenon has been observed in the case of water consumption [Ferraro and Price, 2013].

Some research has delved into ways to bend mental models. While only telling people about their wrong models seems to be unsuccessful, [Datta and Mullainathan, 2014] advise to show people with practical proofs that undermine their main wrong beliefs. Accounting for mental models is critical for our research because many farming practices and behaviors are grounded on multigenerational transmission while not necessarily updated. Furthermore, it might be necessary to understand mental models in agriculture in order to find the most appropriate to communicate information to farmers.

Bounded rationality was initially presented by Herbert Simon in 1978 [Simon, 1978]. This approach assumes cognitive limitations and consequently, people should not be expected to strictly comply with optimization, contrary to what assumes the neoclassical economic theory. Contrary to this latter, [Simon, 1978] stated: “But there are no direct observations that individuals or firms do actually equate marginal costs and revenues”. By contrast, Simon preconizes to focus on procedural rationality, namely the quality of
the procedure to reach a choice. An example of procedural rationality is given by the satisficing behavior that consists in a gradual adjustment of people’s behavior through iterative steps in order to reach a specific aspiration level [Simon, 1955, Simon, 1978, Simon, 2000, Gigerenzer and Selten, 2001]. Bounded rationality is extremely relevant for agriculture because management practices and decisions are not necessarily optimal for the farming output rate. In particular, farmers could adjust their initial production objective in order to achieve a certain level of satisfaction.

Gigerenzer and colleagues have also studied the topic of bounded rationality by introducing the concept of fast and frugal heuristics [Gigerenzer et al., 1999, Gigerenzer and Selten, 2001]. They assume that heuristics are not mistakes but rather an adaptation to bounded rationality. Among fast and frugal heuristics, two have received particular attention: the recognition heuristic and the take-the-best heuristic. The former states that in face of two alternatives, people will choose the one that they recognize. Such a heuristic has been observed in investment strategies and proved to be more efficient than random selection, expert assessment or a set of statistical techniques able to come up with a selection process [B. et al., 1999, Czerlinski et al., 1999]. In the agricultural context, farmers would for instance choose the fertilizer whose brand they are used to.

However, such developments about recognition heuristic are contested. In particular, research has shown the relevancy of fast and frugal heuristics in domains where optimization does not apply. In that respect, the benchmark that this framework is supposed to outperform would be absent. As an example, fast and frugal heuristics were applied to stock markets choice, but this sector has proved to be unable to validate the efficient market hypothesis, thus crowding out the comparison with the optimal solution based on the neoclassic economic theory. Furthermore, as mentioned by [Dhami, 2016], it does not appear obvious that people really use fast and frugal heuristics.

Such critics lead us to spot light on the principal debate called “the great rationality debate” that arose in behavioral economics. Theoretical debates in behavioral economics
have opposed two main streams. The first one was introduced in 1974 by Tversky and Kahneman with their seminal paper ‘Judgment under uncertainty: heuristics and biases’ and assumes that individual are inconsistent with the rationality principle on which the traditional economic maximization is grounded. The second trend was led by Gigerenzer and colleagues, and arose as a critic of the former. It asserts that heuristics are more efficient than statistical counterparts for some specific topics. As [Dhami, 2016] fairly notices, both views may be right depending on the subjects and the context. Finally, and in particular possibly relevant for research about issues in developing countries, recent research developments have focused on limited attention. This concept deals with the restrained cognitive skills of individuals.

In contrast with classical economic theory that assumes people’s full attention, experience shows that this statement proves to be invalid. According to [Datta and Mullainathan, 2014], taking into account limited attention is particularly critical when dealing with the poorest whose focus would be distracted by more essential issues like food or clean water. As a consequence, this crowding out of the attention towards vital issues, also called the problem of limited bandwidth [Banerjee and Mullainathan, 2008] could be assimilated to a cognitive tax [Shah et al., 2012, Mullainathan and Shafir, 2013]. Limited attention could prove to be very relevant in the farming context, in particular in the Semi-Arid Tropics in rural India, because many of them have to deal with droughts, flood, pest attacks, uncertainty, financial constraint and the sustainability of their family. The following chapters aim at quantifying in what extent the heuristics presented in this chapter are relevant in order to explain yield gap.
Chapter 10

Data description

10.1 Introduction

In this chapter, we present an overview and a description of the data that we study in the two following chapters. The dataset is kindly provided to us by the International Crop Research Institute for Semi-Arid Tropics (ICRISAT). ICRISAT has collected data since 1972 in semi-arid regions via the Village Level Studies (VLS). These surveys have permitted to gather micro and meso data that include census information and data about agro-climatic features, irrigation, land use, crop and area cultivated, infrastructure, prices and institutions. The first generation of data (VLS1) extends from 1975 to 1984. The second generation (VLS2) covers the period 2001-2008. The last campaign up to date extents from 2009 to 2014. 400 households from 10 villages in four different states (Andhra Pradesh, Maharashtra, Gujarat, Madhya Pradesh) are surveyed in the VLS1. As for the VLS2, 6 of the initial villages are studied within two states (Andhra Pradesh, Maharashtra) and includes 265 households. The most recent campaign launched in 2009 fits into the Village Dynamics in South Asia (VDSA) project and covers 42 villages in India and Bangladesh. In the next subsections, we show some lessons that can be drawn for the descriptive analysis of ICRISAT’s dataset. We provide some basic and intuitive ways to read the correlations between variables, even if it does not imply causation at all. More analysis is necessary to draw more robust statements. Notice that the results are provided for the year 2014.


10.2 Demographics

In this subsection, we focus the data analysis on the demographic variables.

10.2.1 Occupation

The following barplots display the main activity of the first head in the household. More precisely, for each occupational category, we show the percentage of households whose first head claims to work in the professional field.

Main occupation

Figure 10-1: Main occupation of the Head 1 in the household
Figure 10-2: Main occupation of the Head 2 in the household

Subsequent occupation

Figure 10-3: Subsequent occupation of the Head 1 in the household
For each professional category (from 1 to 12), the y-axis show the percentage of households in this category. For instance, from Figure 10-1, around 52% percent of the head 1 (main person of the household) declares to work in farming as their first activity. Overall, we can observe that most households have the primary head in farming (first activity for 52% of them). Farming (19%), livestock (19%), farm labor (17%) and domestic work (16%) are the principal second activities of the head in the household. By contrast, the second head of the household are mostly committed in farming (29%) and domestic work (29%) for their first activity. The second activity of the second head of the household is usually domestic work (for 32% of them) and farming (16%).
10.2.2 Education

All population

The x-axis represents the years of education and the y-axis displays the percentage of head 1 (respectively head) that have the corresponding level of education. As an example, around 45% of head 2 didn’t receive education.
We observe that on average, the proportion of head 2 (usually the woman) that didn’t receive education is higher than the proportion of head 1 that didn’t receive education.

**By landholding category**

![Figure 10-7: Level of education by landholding size](image)

We use the following encoding:

- LB: landless
- SM: small
- MD: Medium
- LA: Large

As an example, the medium farmers have on average 1.72 years of education. We use the same convention for the next graphs. We observe that the medium and large farmers seem to be slightly more educated. Nevertheless, it does not imply that this difference is statistically significant.
10.3 Farming production

10.3.1 Land area

The x-axis corresponds to intervals of land area in acre. The y-axis shows the percentage of households for each interval. Each bar corresponds to an interval of 0.5 acres. As an example, there are less than 1% of the farmers in our panel data that have between 0 and 0.5 acres. Overall, about 53% of farmers have less than 5 acres in total.
10.3.2 Number of plots

On this graph, the x-axis represents the number of plots cultivated. The y-axis displays the percentage of farmers for each category. For instance, around 29% of farmers do farming on one plot.
10.3.3 Percentage irrigation

Figure 10-10: Share of irrigated land by landholding size

As before, LB, SM, MD and LA stand for landless, small, medium and large size respectively. The y-axis represents the percentage of households that irrigate for each subcategories of farmers. For instance, around 0.41% of large farmers irrigate. Overall, we notice that large farmers use more widely irrigation. Nevertheless, the difference is very tiny and is probably not statistically significant.
10.3.4 Output rate for all population – Kharif Cotton

On this plot, we show the percentage of households that get an output in a specific interval. The y-axis represents the percentage of households. The intervals for the output rate (in kg/acre) is indicated by colors. We distinguish between irrigated and non-irrigated land. Results are shown for cotton during the kharif season. As an example, for 18% of the households that irrigate (left bar) have an output rate of cotton during Kharif season between 30 and 35 kg/acre (navy blue section of the bars). Overall, farmers that do not irrigate are more likely to get an output rate between 35 and 40 kg/acre (around 48% of them are in this case). There is a higher proportion of farmers that irrigate with an output rate between 40 and 45 kg/acs (around 22% of them) compared to farmers that do not irrigate (around 10%).

Figure 10-11: Output rate of cotton - Irrigated vs. non-irrigated
10.3.5 Output rate by landholding category

We use the same convention to indicate the landholding categories. The y-axis indicates the output rate of cotton obtained during the kharif season (in kg/acre). For instance, landless farmers get 42 kg/acre on average. We notice that on average, landless farmers and large farmers (respectively 42.5 and 43.5 kg/acre) do better than small and medium farmers (respectively 40 and 40.5 kg/acre).

Figure 10-12: Output rate of cotton during kharif season by landholding size
10.4 Revenue

10.4.1 Farm profit

We compute the economic profit (revenue minus cost) generated by households thanks to their farming activities. The red vertical line indicates the zero profit. The area under the curve and on the left of the red vertical line shows the proportion of farmers with a negative profit (basically, farmers that lose money from farming). We notice that most farmers lose money from their farming activities. Actually, in 2014, out of 871 farmers, 12 had a positive profit from farming (area under the curve and on the right of the red vertical line).
10.4.2 Total revenue

![Distribution of income for farm and non-farm households](image)

Figure 10-14: Distribution of income for farm and non-farm households

Instead of looking at only the income from farming, we analyze the total income, thus including off-farm activities such as livestock, trading or construction. Notice that contrary to the previous plot, we do not look at the profit (revenue minus cost) but the total revenue. In particular, we compare the total revenue for households that have only farming activities and the total revenue for households that also do off-farm activities. We notice a shift to the right of the revenue distribution for households that have off-farm activities in addition to farm activities. In other words, farmers with off-farm activities would tend to have higher total revenue on average.

10.5 Finance

10.5.1 Debt for all population

We report here the histogram of money borrowed by households. The x-axis shows the amount of money borrowed. Each bar represents an interval of 10,000 rupees. The y-axis displays the percentage of households for each interval of borrowed money. For instance, 25% of households have between 0 and 10,000 rupees of loans. Overall, 47% of the households have less than 30,000 rupees of loans.
10.5.2 Debt by category of landholding

On this graph, we show the amount of money borrowed for each category of farmers (as before, landless, small, medium and large farmers). It seems that large farmers have on average higher loans (220,000) than the other categories. Nevertheless, we notice that this plot must cautiously interpreted since we display the average by category. In each category, there are outliers (see previous plot) with huge debt. Furthermore, this plot
shows the cumulative debt while the previous one shows the loan during the previous year. 6. Market distance

![Market distance](image)

**Figure 10-17: Distance to market by landholding size**

We analyze the distance between farmers’ home residence and the closest market where they can sell their production and buy inputs. The difference between landholding categories is extremely small and very unlikely significant. On average, farmers live 13km far from the closest market.

In the next two chapters, we use quantitative methods in order to investigate these preliminary results.
Chapter 11

Econometric analysis of yield gap in Semi-Arid Tropics India

11.1 Introduction

In this chapter, we focus on identifying the determinants of yield gaps in rural areas of Semi-Arid Tropics in India. Rather than focusing on the gap between the ideal yield predicted by crop models, we study the gap between the yield of each household and the maximum yield for the same crop obtained in their village. For practical reason, we consider the difference between the maximum and the observed yield so as to work with positive values. This village-based approach enables us to overcome the geophysical aspects that characterize local situations. In that respect, we can take into account soil characteristics, that may not be necessarily measured. In our view, it would be irrelevant to compare the yield of farmers that live in two different locations (e.g. villages) where the underlying agronomic factors, and management practices differ sufficiently so as to violate the ceteris paribus principle. In this respect, we see the village level as the most appropriate reference for yield comparison rather block, district, state or national levels.
11.2 Methodology

We analyze five crops (cotton in kharif season, paddy in kharif season, pigeonpea in kharif season, sorghum in rabi season and wheat in rabi season). The Kharif season extends from July to October while the Rabi season covers October to March. Two reasons motivate this choice. First, these crops constitute staple food for many rural Indian households. Second, they are massively grown by farmers and in that respect, contrary to other crops, we have substantial amount of data. Households are spread into 30 different villages within 8 states (Andhra Pradesh, Maharashtra, Karnataka, Gujarat, Madhya Pradesh, Jharkhand, Odisha, Bihar). Observations cover six years from 2009 to 2014.

We first estimate a static model for which the output rate is the dependent variable. This model is built according to the following equation:

\[ Q_{i,t} = \alpha_{10} + \alpha_{11} \text{Variables}_{-\text{households}}_{i,t} + \alpha_{12} \text{Variables}_{-\text{farm}}_{i,t} + \alpha_{13} \text{Variables}_{-\text{soil}}_{i,t} + \alpha_{14} \text{Variables}_{-\text{inputs}}_{i,t} + \alpha_{15} \text{Variables}_{-\text{finance}}_{i,t} + \alpha_{16} \text{Rain}_{i,t} + \alpha_{17} \text{dummy}_\text{year} + \alpha_{18} \text{village}_\text{*year} \]

where \( Q \) is the output rate.

The output rate dependent variable is measured as the weight of the total production in kilograms (kg) divided by the area over which it has been cultivated in acres.

The independent variables included in the regression are summarized in the table below:

The income include wages. Wages are the part of revenue coming from other sources of revenue than farming production such as livestock, trade and employment (farm and non-farm). The monetary value are expressed under their logarithm form in order to deal with large variation and outliers. It also enables us to overcome the issue of inflation.

We decided to account for the variables measuring the income, the amount of bor-
Table 11.1: Description of variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Variables included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables_households</td>
<td>Age head of household</td>
</tr>
<tr>
<td></td>
<td>land area under crop$_i$, percentage of irrigated land under crop$_i$, dummy variable for ownership, number of crops, number of plots</td>
</tr>
<tr>
<td>Variables_farm</td>
<td>area clay, area clay loam, area loam, area sandy, area sandy-loam, area deep black, area medium black, area shallow-black murrum, area red soil, area problematic soils, area without degradability, area of soil with salinity/acidity, area of soil with nutrient depletion, eroded soil, area of soil with water logging</td>
</tr>
<tr>
<td>Variables_soil</td>
<td>quantity of fertilizer used for crop$_i$, quantity of pesticide used for crop$_i$, area under crop$_i$ with hybrid seed, area under crop$_i$ with High Yield Variety (HYV) seed, area under crop$_i$ with BT seed</td>
</tr>
<tr>
<td>Variables_inputs</td>
<td>total income earned at the previous period, amount of borrowings contracted at the previous period, government benefits received at the previous period</td>
</tr>
<tr>
<td>Variables_finance</td>
<td>total income earned at the previous period, amount of borrowings contracted at the previous period, government benefits received at the previous period</td>
</tr>
<tr>
<td>Rain</td>
<td>Total annual amount of rain in mm</td>
</tr>
</tbody>
</table>

rowings and government benefits lagged by one period. Effectively, the current value of these variables are unlikely to dramatically affect household’s productivity because the decisions for the current agricultural season are taken simultaneously or even before households hold these amounts of money. On the contrary, income, borrowed money and government benefits earned at the previous period are likely to impact household’s decisions, farm investments, financial flexibility and hence productivity.

The dummy variable entitled *binary_owner* is equal to 1 when the household owns at least one plot of land and 0 otherwise.

All areas are measured for each crop that are investigated. In other words and as an example, the area of red soil for cotton for the household is the land area cultivated with
cotton and for which the soil is of type red soil.

We must mention here that data about soil fertility and slope were at our disposal. However, we decided not to include it in the model because the accuracy of such data are highly questionable. Effectively, these soil characteristics are not measured but directly reported by the farmer. The classification is categorical, namely "very poor", "poor", "good" and "very good" for the fertility, and in percentage for the slope. This information have more to do with households’ perception than objective data. We consequently do not include it in the model.

We regress the dependent variable on the covariates first using a panel data estimation with fixed effects. The fixed effects estimator is the most relevant estimator in our case because the unobserved heterogeneity is likely correlated with the explanatory variables. Formally, this estimation is grounded on the following single-explanatory variable model:

$$y_{i,t} = \beta_1 x_{i,t} + a_i + u_{i,t} \text{ for } t = 1, 2, \ldots T$$

(11.1)

where $y_{i,t}$ is the dependent variable for the unit $i$ observed at time $t$; $x_{i,t}$ is the unique covariate of the model; $a_i$ is the unobserved fixed effect or unobserved heterogeneity; $u_{i,t}$ is the idiosyncratic error or time-varying error;

The fixed effects estimation method assume that $\text{Cov}(x_{i,t}, a_i) \neq 0$. Alternatively, using a random effects estimator would imply that unobserved heterogeneity is uncorrelated with the covariates. As a consequence, this estimation method would consider that the unobserved management ability of farmer is uncorrelated with the income. Such assumption is likely inaccurate in practice. As a robustness check, we perform the regression with a random effects estimator and confirm the relevance of the fixed effects estimator with a Haussman test (see part VI). We do not include a cross-pooling estimation. In our view,
it would be inappropriate to consider that the observations are independent in our study since they are grouped by household. We use the coefficients obtained with the fixed effects method in the discussion part.

Using a fixed effects estimation method enables us to identify the key determinants of the output rate. We use a dummy variable for the year in order to account for a "year effect", in other words the peculiarities that may have happened during each year, in particular climatic events such as droughts or floods. We also introduce an interaction term between the village and the year in order to account for a “village effect”. As mentioned below in the hypothesis section, this interaction term enables us to test whether the village membership is correlated with farm productivity.

Yet, fixed effects estimation does not allow us to estimate the coefficient of time-invariant variables. Similarly, the effects of variables with small variations will not be estimated. These two aspects constitutes two highly challenging research gaps in the analysis of panel data that may deserve further investigation (see for instance Arellano et al. 2016). However, the effect of such covariates is controlled in the fixed effects estimation method. It is of high importance in particular as far as households’ heterogeneity and ability are concerned.

We then use a quantile regression estimation of the same model. Rather than focusing on the characteristics at the origin of the output rate for the average farmer, quantile regression estimation enables us to estimate the effects of the covariates for each quantile of the output distribution. To the best of our knowledge, no previous study provides this insight for agricultural productivity. This aspect is interesting for two reasons. First, the explanatory variables at the origin of yield gap might differ depending on the position in the distribution of productivity. Second, such an analysis may enable us to identify latent class of farmers based on their productivity. These latent classes may differ from the traditional categorization into land areas usually adopted in agronomy and agricultural economics. The identification of such latent classes may prove to be critical for the
design of appropriate policies aiming at improving agricultural yields and rural households’ welfare.

The second model tackles the absolute change in production. Particularly, we focus the analysis on the dynamics of the output. The objective is twofold. First, we want to identify the determinants of growth in output. Second, we want to test for stasis in farm productivity level. We estimate the following model:

\[
\Delta Q_{i,t} = \alpha_0 + \alpha_1 Q_{i,t-1} + \alpha_{12} \text{Variables}_{\text{households},i,t} + \alpha_{13} \text{Variables}_{\text{farm},i,t} \\
+ \alpha_{14} \text{Variables}_{\text{soil},i,t} + \alpha_{15} \text{Variables}_{\text{inputs},i,t} + \alpha_{16} \text{Variables}_{\text{finance},i,t} \\
+ \alpha_{17} \text{Rain}_{i,t} + \alpha_{18} \text{dummy}_\text{year} + \alpha_{19} \text{village} \times \text{year}
\]

where

\[
\Delta Q_{i,t} = \frac{Q_{i,t} - Q_{i,t-1}}{Q_{i,t-1}}
\]

is the change in output rate.

Strictly speaking, this model is not dynamic. However, we are able to introduce a dynamical component by using the growth in productivity as our dependent variable and the lagged productivity at the previous period as one of our independent variables. A positive (respectively negative) value for the dependent variable of this model means that the output has increased (respectively decreased) over two consecutive periods.

As with the static model, we use a panel data estimation with fixed effects. We estimate first the coefficients of the covariates for the average farmer and then for each quantile of the output rate distribution. The same motivations as for the static model justify the use of both estimation methods. The model enables us to investigate the possibility of different patterns of dynamics of productivity depending on the position in the yield distribution. For that reason, a quantile regression estimation could be relevant to compare
different groups of farmers based on their productivity.

We finally focus on the change in yield gap compared to the maximum productivity observed in the village of each household. The objective is to identify the determinants of the change in yield gaps when the gap is taken with the maximum of the output rate distribution. With that purpose, we estimate the following model:

\[ \Delta YG_{i,t} = \alpha_{10} + \alpha_{11} YG_{i,t} + \alpha_{12} \text{Variables}_{\text{households}_{i,t}} + \alpha_{13} \text{Variables}_{\text{farm}_{i,t}} \]
\[ + \alpha_{14} \text{Variables}_{\text{soil}_{i,t}} + \alpha_{15} \text{Variables}_{\text{inputs}_{i,t}} + \alpha_{16} \text{Variables}_{\text{finance}_{i,t}} \]
\[ + \alpha_{17} \text{Rain}_{i,t} + \alpha_{18} \text{dummy}_{\text{year}} + \alpha_{19} \text{village} \times \text{year} \] (11.4)

where

\[ \Delta YG_{i,t} = \frac{YG_{i,t} - YG_{i,t-1}}{YG_{i,t-1}} \] (11.5)

is the change in yield gap.

Like the model (ii), this model includes a dynamical component by considering the change in yield gap as the dependent variable and the yield gap obtained at the previous period as one of the explanatory variable. As for the previous models, we first use a multi-linear regression and then a quantile regression to compare the quantiles of the yield gap distribution. This model (iii) investigates the mobility of farmers within the yield distribution. Since the yield gap is measured as the difference between the maximum output rate in the village of household(i) and the output rate of this household, a positive change (respectively negative) in yield gap, \( \Delta YG > 0 \) (respectively \( \Delta YG < 0 \)), means that the household has a less favorable (respectively more favorable) relative position with respect to the maximum yield obtained in his village compare to the previous period. A negative delta YG does not necessarily mean that the household’s situation has improved in terms of productivity. Only the relative position with respect to the maxi-
mum yield has improved. In other words, the model (iii) enables us to analyze the relative change in productivity within the output distribution.

As for the two precedent models, in addition to the fixed effects estimation method, we also use a random effects estimation in order to compare the coefficients under both methods and supposedly confirm the relevance of the fixed effect method.

Furthermore and as suggested by Wooldridge [Wooldridge, 2010, Wooldridge, 2019], we also use a correlated random effects (CRE) estimation method. This type of estimator enables us to keep going with fixed effects while introducing time-invariant variables.

Formally, we assume that the unobserved heterogeneity can be expressed by the following linear relation:

\[ a_i = \alpha + \gamma \bar{x}_i + r_i \]  \hspace{1cm} (11.6)

where \( \bar{x}_i = \frac{\sum_{t=1}^{T} x_{i,t}}{T} \) is the time average of the covariate \( x_{i,t} \) and \( r_i \) is the error term that is assumed to be uncorrelated with \( x_{i,t} \). We thus have:

\[ \text{Cov}(x_i, r_i) = 0 \] and consequently \( \text{Cov}(x_{i,t}, r_i) = 0 \) since \( \bar{x}_i \) is a linear function of the \( x_{i,t} \).

Combining the two equations 11.1 and 11.6 we end up with:

\[ y_{i,t} = \alpha + \beta x_{i,t} + \gamma \bar{x}_i + r_i + u_{i,t} \]  \hspace{1cm} (11.7)

The composite error \( r_i + u_{i,t} \) is uncorrelated with the explanatory variable \( x_{i,t} \). The average variable \( \bar{x}_i \) enables us to control for the correlation between the covariate and the unobserved heterogeneity. The equation 11.7 is finally the model underlying the random effects estimation method.

The advantage of this estimation method is to allow us to introduce the time-invariant variables and estimate their effects. In particular, we add the education level, the size of the family, the main and subsequent occupation of the household, the gender of the head
of the household, the distance to market, religion and caste. The analysis of the CRE regression will thus focus on the estimated coefficients of time-invariant variables in order to determine whether these variables enable us to further explain output rates and yield gaps.

We must also highlight the constraint involved by the unbalanced structure of our panel on the estimation. As explained by Wooldridge [Wooldridge, 2010, Wooldridge, 2019], the time-average of time-varying variables must only include the observations for which all variables of the regression are available. For that reason, the period T in the expression of the average differs for each unit i.

11.3 Hypotheses

Models (i) and (ii) enable us to test three hypotheses about the characteristics of farmers that may determine output rate and change in output rate:

- The output rate and change in output rates are mainly determined by physical and observable characteristics
- The output rate and change in output rates are mainly determined by unobserved variables thus illustrating the critical impact of unobserved heterogeneity between households that were overlooked by previous research
- Explanatory variables of output rate and change in output differ along the distribution of output distribution thus indicating the diversity of farmers beyond the landholding characteristics

Both the second and third models enable us to test the following hypothesis:

- Change in output rates and yield gaps follow a stable path, namely an increase in the change of the output rate and the change of the yield gap is likely to be followed by a decrease, thus indicating a capped production in the long term
Using data for five crops, we also aim at testing whether conclusions hold for all crops or if they differ depending on the crop considered. Finally, introducing an interaction term including the village affiliation, we might test whether the village membership has a critical impact on the level and dynamics of output rate, thus showing evidence for a “village effect” on productivity. If this hypothesis proves to be confirmed by data analysis, we would suggest three explanations for which the village variable might play a role on productivity level:

- Geophysical characteristics: in particular time-invariant soil characteristics for which we cannot estimate the effect in a fixed effects estimation since they are time-invariant. In addition, it is not necessarily possible to measure some of these characteristics.

- Geographical location: in particular, the proximity of urban areas and the endowments in infrastructure such as roads (cf. study Costinot - economic impact of roads)

- Village synergy: solidarity among people, networks building, exchange of information and good practice and economic activity spillovers. These aspects, usually built over long time, are peculiar to each village and might prove to be critical in building a productive environment and positive livelihood context. However, these aspects are difficult to be represented with measurable variables.

### 11.4 Data description

ICRISAT has collected data since 1972 in semi-arid regions via the Village Level Studies (VLS). These surveys have permitted to gather micro and meso data that include census information and data about agro-climatic features, irrigation, land use, crop and area cultivated, infrastructure, prices and institutions. The first generation of data (VLS1) extends from 1975 to 1984. The second generation (VLS2) covers the period 2001-2008. The last campaign up to date extents from 2009 to 2014. 400 households from 10 villages in four different states (Andhra Pradesh, Maharashtra, Gujarat, Madhya Pradesh) are surveyed in the VLS1. As for the VLS2, 6 of the initial villages are studied within two states (Andhra Pradesh, Maharashtra) and includes 265 households. The most recent campaign
launched in 2009 fits into the Village Dynamics in South Asia (VDSA) project and covers 42 villages in India and Bangladesh. The datasets and descriptions are available via http://vdsakb.icrisat.ac.in/Login.aspx?ReturnUrl=%2fdefault.aspx

11.5 Results

We now discuss the results of these models that are presented at the end of this chapter in tables 1 to 6. **Kharif Cotton**

- **Output rate**

As for the analysis of cotton during kharif season, we find a negative correlation between the output rate obtained at the previous period and the change in output rate over two consecutive periods with a p-value lower than 0.001. The amount of rain received in a year is also positively correlated (0.03240, p < 0.05) with the change in output rate. (see table 2).

- **Delta output rate**

The change in yield gap with respect to the maximum shows a positive coefficient for the percentage of land grown with cotton that are irrigated. The coefficient (0.0661) is significant at level 0.05 (see table 3).

**Kharif Paddy**

- **Output rate**

We find a positive correlation between the output rate of paddy and the total number of crops grown by households (0.046, p < 0.05). An interesting result comes from the quantile regression, in particular as far as the 90th-quantile of the output distribution is concerned. The total land area and land area with a soil of type medium black are both found to be positively correlated with the output rate of paddy, with coefficient respectively equal to 1.558(p < 0.05) and 0.952(0.001). By contrast, the land area presenting no degradation and
the quantity of fertilizer used for paddy crop are negatively correlated. Coefficients are respectively -1.504 (p<0.05) and -0.00345 (p<0.001) (see table 2).

- Delta output rate

The dependent variable is significantly and negatively correlated with the output rate, expressed in logarithm form, observed at the previous period. The coefficient equals -1.448 and is significant at level 0.05 (see table 3)

- Delta YG max

Regression of the change in the difference between farmer(i) output and the maximum yield measured in his own village is negatively correlated with the difference observed at the previous period (-2.484, p<0.01) (see table 4).

**Kharif Pigeonpea**

- Output rate

The output rate of the average farmer is negatively correlated (-7.340) with the land area characterized by a soil of type loam. This coefficient is significant at level (0.01). Similarly, land areas grown with pigeon pea and presenting problematic soils are negatively correlated with yields (-13.18, p<0.05). On the contrary, rain is positively correlated, the coefficient being equal to 0.993 and significant at level 0.05 (see table 2).

- Delta output rate

The output rate of pigeon pea measured at the previous period is negatively correlated with the change in output rate over two consecutive periods (-1.306, p<0.001) (see table 3).
• Delta YG max

The change in yield gap to the maximum is negatively correlated with the gap measured at the previous period. The coefficient equals -0.267 and is significant at 0.1%. Similarly, the dummy variable measuring the shift from landless to landowner status is negatively correlated with the change in yield gap with respect to the maximum (-908.5, p<0.001). On the contrary, the land area grown with pigeon pea and presenting a nutrient depletion is positively correlated with the change in yield gap (9.473, p<0.001). Similar positive correlations are found for land area with a soil of type loam, land area with a soil of type sandy-loam and annual amount of rain. Coefficients are respectively 4.762(p<0.05), 15.59(0.05) and 1.139(p<0.001). Interestingly, the quantile regression confirms these results for the 10th top quantile of the distribution of the change in yield gap to the maximum. The land area with soil of type loam and of type sandy are both positively correlated. Coefficients are respectively 13.30(p<0.001) and 5.612(p<0.001) (see table 4).

Rabi Sorghum

• Output rate

We find a significant coefficient of 2.162 with level of significance 5% when the output rate of sorghum during rabi season is regressed on the income obtained at the previous period (see table 2).

• Delta output rate

The output rate obtained at the previous period is negatively correlated with the change in output rate (-1.104, p<0.001). The quantile regression displays a similar result for the median with a coefficient (-0.970) significant at 0.1%. A negative correlation is also found for the land area characterized by problematic soils (-0.495, p<0.01). The total number of plots shows a positive coefficient equal to 0.0979, at significance level 1% (see table 3).
The gap with respect to the maximum yield observed in the village is found to be negatively correlated with the change in yield gap over two periods (-0.414, p<0.001). We also find negative correlation with the same dependent variables for the land area without degradability (-5.027), the land area presenting erosion (-5.194) and the quantity of fertilizer (-0.0106), with respective significance level of 5%, 1% and 5%. On the contrary, we find positive correlations for the land area of soil type loam (5.245, p<0.05), land area of type sandy loam (15.42, p<0.01), land area of soil type sandy (4.539, p<0.05) and annual amount of rain (0.711, p<0.001) (see table 4).

**Rabi wheat**

- **Output rate**

Considering the output rate of wheat during rabi season, we find that the annual amount of rain is negatively correlated with a coefficient equal to -0.373 and a significance level of 1%. The quantile regression finds a negative correlation between the amount of fertilizer used for cotton crop and the output rate of the 10th lower quantile of the output rate distribution (see table 2).

- **Delta output rate**

The output rate obtained at the previous period is negatively correlated with the change in output rate over two periods (-1.497, p<0.001) (see table 3).

- **Delta YG max**

We find a negative correlation between the dependent variable and the yield gap of the average farmer with respect to the maximum output rate in the same village observed at the previous period. The coefficient is equal to -0.690 and the p-value is lower than 0.001.
Coefficients are also negative for the following covariates: the quantity of fertilizer used for that crop (-0.00186, p < 0.01), the land area of wheat grown with hybrid seed - by contrast with local or High Yield Variety seed - (-4.985, p < 0.01) and the annual amount of rain (-0.0296, p < 0.001). The total land area dedicated to wheat is positively correlated, with a coefficient of 0.370 and a significance level of 1% (see table 4).

- The correlation random effects estimator for unbalanced panel data confirms the results of the fixed effects estimator, namely that soil characteristics are critical variables in explaining the heterogeneity in output rates. Coefficients for time-invariant variables are mostly insignificant, except in four cases. Distance to market is negatively correlated with the average farmer’s output rate of wheat (-0.195, p < 0.001). Similarly, the market distance is negatively correlated with the change in sorghum productivity (-0.0139, p < 0.01). We also find a positive correlation between religion and the change in pigeon pea output rate (0.0158, p < 0.05) (see table 6).

The figure below shows the distribution in output rate for cotton during the Kharif season in 2012. We observe a similar pattern for each year: a high pick at a certain level of productivity by village and a second smaller pick at zero productivity. This pattern shows that farmers are quite homogeneous in terms of productivity within each village. Furthermore, when farmers do not achieve a good yield, it usually stems from a complete failure of the crop. Only villages were sufficient data are collected are drawn. The villages that do not appear usually host no household growing the considered crop. We then show the summary statistics of the regressions both at the average and for quantile regressions. As a practical purpose for the reader, we summarize in figures 2 and 3 the significant variables of the regression at the average, indicating by a color code the sign of their correlation with the dependent variable.

Following the results of fixed effects estimation and quantile regression, we present the results of the correlated effects estimation for unbalanced panel data. Results are then followed by a discussion.
11.6 Discussion

When considering the dynamic models, we almost systematically find a significant and negative correlation between the output rate (respectively yield gap to the max) obtained at the previous period and the change in output rate (respectively change in yield gap to the max). This result simply shows that for all the crops under study, the production rate will not grow indefinitely and illustrates natural physical constraints. In that respect, such results validate the hypothesis of capped production in the long term. It also show the smoothing effect over time, meaning that a bad year has more chance to be compensated over time by a good year and inversely.

The positive correlation between the change in output rate of cotton and the annual amount of rain might suggest that rainfall is a critical parameter constraining the increase in cotton yield gap. These validations help us have confidence in our models.

The positive correlation between the change in yield gap with the maximum output rate and the share of land grown with cotton that are irrigated might imply that increasing
the irrigation capacity of the average farmer might increase its relative difference in terms of output with the top producer. In that respect and everything else controlled, irrigation might not enable the average farmer to bridge his gap with the most productive farmer. At this point, two explanations could be advanced: either irrigation is not determinant is cotton output rate, or there exists a hidden ability or skill to be mastered that prevents the average farmer to level up in the output rate distribution when land irrigation is increased.

Since the output rate of paddy and the number of crops appear to be two independent variables, the positive correlation found between these two covariates could let us think that an intrinsic and unobserved household’s competence might be the missing link between these two events. Effectively, households endowed with higher ability and management capability might be prone to grow more diverse crops and also would be more successful in their paddy productivity.

Results about pigeon pea give us good insights about the effect of soil type. In particular, it seems that growing pigeon pea in a soil of type loam is a counterproductive and inefficient cropping choice. Not surprisingly, we find that soil qualified as problematic soil by farmers themselves have a negative relationship with yields. As for cotton, rainfall amount has a positive effect on paddy yield.

As shown by the significant and negative relationship between the change in yield gap and the dummy variable expressing the landholding status, the access to ownership seems to have a strong effect on the reduction of the gap between the average farmer and the most productive farmer in the same village. This effect might be explained by the positive effect of learning-by-doing and gain of experience. Stability in the land plots cultivated and in particular in soil type would enable farmers to draw the lessons from one season to the other. Yet, this interpretation is not confirmed by the covariate of age. Furthermore, the regression of change in yield gap to the maximum show the critical impact of the type of soil. In that respect, nutrient depletion, loam, sandy-loam does not enable
the average farmer to positively move in the output rate distribution relative to the most productive farmers. The quantile regression enables us to refine this interpretation, confirming the positive correlation loam and sandy soil for the 90th quantile, meaning those that have the greatest gap with the maximum yield.

As far as sorghum is concerned, the income earned at the previous period seems to have a critical role in the output rate of the present year. Everything else controlled, the interpretation of such a positive relationship might be quite hard to explain. A higher income at the previous period might play as a risk-alleviating mechanism thus fostering initiative and risk-taking. This result is not found for the other crops and the interpretation remains hypothetical.

Logically, the area of problematic soils shows a negative correlation in the change of sorghum output rate. This results might suggest that there effectively exists a land-fertility path dependence [C. et al., 2015] that would trap some farmers in a low productivity stasis. In addition to that, the positive correlation between the change in output rate and the number of plots could illustrate the positive effect of diversifying the risk of crop failure. By having separated plots, farmers could reduce the risk to fail all their production as it would be the case if the crop is grown on a single plot and hit by an exogenous event during the season such as a pest attack for instance. This result could also illustrate that sorghum is a riskier crop. Such an interpretation would be consistent with the risk-taking interpretation of the positive relationship between income and output rate mentioned above.

Results about the change in yield gap confirm one more time the importance of the type of soil. In particular, land without problem of degradability tends to reduce the gap between the average farmer and the maximum yield over time. Increasing the quantity of fertilizer, at least for sorghum cultivation appears to have the same effect. The positive coefficients found for loam soil, sandy-loam soil and sandy soil would suggest that these type of soils are not the most appropriate for growing sorghum since they do not allow
the lowest productive farmers to bridge their gap of productivity with the most productive ones.

We also found a negative correlation of the change in yield gap with land erosion. Despite being significant, this relationship is highly questionable since it would imply a decrease between the average farmer and the maximum yield. We have to acknowledge the oddity of this result that we are unable to soundly interpret, except if we consider an homogenization of output rate triggered by soil erosion.

The negative correlation for rainfall amount would suggest that water is not the main constraint for wheat cultivation. Interestingly, this result is confirmed for the most productive farmers via the quantile regression that shows the same negative correlation for the 10th highest quantile of the output rate distribution.

The negative correlation found between change in yield gap and the amount of fertilizer would suggest that the average farmer might gain from using more fertilizer in terms of output rate compare to the maximum yield. We must insist that this result is crop specific and hold only for wheat in that particular case. Similarly, hybrid seeds could have the same reducing effect on the variance of the distribution of wheat output rate. While rain seems to have a negative impact on wheat output rate, it seems to impact the output rate distribution by reducing the difference between the average farmer and the most productive farmer.

The negative effect of increasing land area cultivated with wheat on the change in yield gap could let us think that everything else controlled, smaller landholders might be more efficient than big landowner, thus confirming results of other research about the inverse land size-efficiency relationship [Gautam and Ahmed, 2019].

The correlated random effects estimation enables us to introduce time-invariant covariates in the model. Among the time-invariant covariates that we include in our model,
level of education and size of the household are expected to have a positive impact on output rate according to the literature while distance to market is expected to have a negative impact. Only the relationship involving the distance to market appears to be significant in our regression and solely for the case of wheat. Otherwise, the results of the correlated random effects confirm the results of the fixed effects estimation, namely that soil characteristics are the critical variable explaining households' farm productivity. Generally speaking, coefficients for time-invariant variables do not appear to be significant. Whether or not they impact farm productivity, at least their effect seems to be displaced by the effect of soil type.

Chapter conclusion

Our study does not enable to draw general pattern of variables responsible for low or high productivity that would be valid for all types of crops. On the contrary, variables that determine productivity seem to be crop specific. We show in next page a summary table of the critical variables for each crop. A green box indicates a positive correlation between the variable and the output rate, a red box indicates a negative correlation between the variable and the output rate.

Figure 11-2: Summary table of the critical variables that are correlated to output rate
In addition, two aspects emerge from this econometric analysis of agricultural productivity. First, soil characteristics have a critical impact on output rate, change in output rate and the variance in output within a village. More precisely, problematic soils and inappropriate type of soil such as loam and sandy loam have critical negative effects on output rates and could partly explain why some farmers are locked in with low productivity.

Second, a “hidden ability” that we were able to account for thanks to the fixed effects estimation of our models, seems to greatly impact the success of failure of rural households. In other words, with the same geophysical and financial conditions, households with similar characteristics perform differently. Our models and chosen estimation methods enable us to confirm the striking effect of households’ heterogeneity on agricultural productivity. Yet, they do not allow us to explain this heterogeneity. Further research might be necessary to delve into that topic.

We show below some of the regression tables that result from the econometric analysis. We provide below the keys to read the table:

- age.h1: age of the head of the household
- area: total area cultivated under the specific crop
- irri_percent: percentage of the plot cultivated with a specific crop that is irrigated
- biry_owner: indicates whether the household owns the plot of not
• are_no_degra: land area cultivated with the specific crop that presents no soil degradability

• area erosion: land area cultivated with the specific crop that presents evidence of erosion

• area_nutrient_dep: area erosion: land area cultivated with the specific crop that presents nutrients depletion

• Area_salinity_acidity: land area cultivated with the specific crop has soil characterized by acidity/salinity

• area_water_logging: land area cultivated with the specific crop has soil characterized by water logging

• area_loam: land area cultivated with the specific crop with soil of type loam

• area_sandy_loam: land area cultivated with the specific crop with soil of type sandy loam

• area_prob_soil: land area cultivated with the specific crop with problematic soil

• area_deep_black: land area cultivated with the specific crop with soil of type deep black

• area_medium_black: land area cultivated with the specific crop with soil of type medium black

• area_shallow_black_murrum: land area cultivated with the specific crop with soil of type shallow black murrum

• area_sandy: land area cultivated with the specific crop with soil of type sandy

• area_red_soil: land area cultivated with the specific crop with soil of type red soil

• nb_plots: number of plots cultivated by the farmer with the specific crop

• nbr_crops: number of crops cultivated by the farmer
• qty_fertilizer: quantity of fertilizer used by the farmer for the specific crop

• qty_pesticide: quantity of pesticide used by the farmer for the specific crop

• area_seed_2: land area cultivated with improved/High Yield Variety seed for the specific crop

• area_seed_3: land area cultivated with hybrid seed for the specific crop

• area_seed_4: land area cultivated with BT (genetically modified) seed for the specific crop

• L.log.income: Income at the previous period (logarithm)

• L.log.borrowing_tot: Total amount of money borrowed by the household

• L.log.gov.benefit: Total amount of money received via government support

• rain: Amount of rain for the entire year

The header of the columns must be read as follows:

• kc.fe: fixed effects model for kharif cotton

• kc.re: random effects model for kharif cotton

• kp.fe: fixed effects model for kharif paddy

• kp.re: random effects model for kharif paddy

• kpp.fe: fixed effects model for kharif pigeonpea

• kpp.re: random effects model for kharif pigeonpea

• rs.fe: fixed effects model for rabi sorghum

• rs.re: random effects model for rabi sorghum

• ws.fe: fixed effects model for wheat sorghum

• rs.re: random effects model for wheat sorghum
We first show the results for the regression with fixed and random effects. We then show the results obtained via quantile regression. For this later case, we only display the case of the output rate of kharif cotton. We finally show the results of the correlated random effects models. Due to space limitation, we only show the case where the dependent variable is the output rate. We display the results for each crop under investigation.
<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_h1</td>
<td>4.265</td>
<td>0.0186</td>
<td>0.196</td>
<td>-0.0230</td>
<td>0.176</td>
<td>0.0450</td>
<td>0.0122</td>
<td>-0.00572</td>
<td>-0.00756</td>
<td>-0.0101*</td>
</tr>
<tr>
<td></td>
<td>(1.31)</td>
<td>(0.62)</td>
<td>(1.64)</td>
<td>(-1.85)</td>
<td>(0.68)</td>
<td>(0.84)</td>
<td>(0.06)</td>
<td>(-0.20)</td>
<td>(-0.27)</td>
<td>(-2.12)</td>
</tr>
<tr>
<td>area</td>
<td>-0.396</td>
<td>0.182</td>
<td>0.322</td>
<td>-0.0342</td>
<td>0.177</td>
<td>-0.0905</td>
<td>-0.469</td>
<td>-0.0305</td>
<td>-0.0842</td>
<td>0.0205</td>
</tr>
<tr>
<td></td>
<td>(-0.61)</td>
<td>(0.37)</td>
<td>(0.72)</td>
<td>(-1.41)</td>
<td>(0.39)</td>
<td>(-0.28)</td>
<td>(-0.95)</td>
<td>(-0.10)</td>
<td>(-1.31)</td>
<td>(0.52)</td>
</tr>
<tr>
<td>irri_percent</td>
<td>-0.0151</td>
<td>0.00722</td>
<td>0.00467</td>
<td>0.00220</td>
<td>0.00906</td>
<td>0.0173</td>
<td>0.0184</td>
<td>-0.00898</td>
<td>0.00387</td>
<td>0.00474</td>
</tr>
<tr>
<td></td>
<td>(-1.17)</td>
<td>(-0.78)</td>
<td>(0.72)</td>
<td>(-1.41)</td>
<td>(0.32)</td>
<td>(0.84)</td>
<td>(-1.17)</td>
<td>(-0.90)</td>
<td>(0.68)</td>
<td>(1.83)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>2.342</td>
<td>4.123**</td>
<td>-3.453</td>
<td>-1.057</td>
<td>-8.058</td>
<td>-0.871</td>
<td>0</td>
<td>0.537</td>
<td>-0.751</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>(0.96)</td>
<td>(2.69)</td>
<td>(-1.53)</td>
<td>(-0.92)</td>
<td>(-0.68)</td>
<td>(-0.30)</td>
<td>(0.29)</td>
<td>(-0.49)</td>
<td>(0.38)</td>
<td></td>
</tr>
<tr>
<td>area_no_degra</td>
<td>0.0500</td>
<td>0.166</td>
<td>0.0763</td>
<td>0.250</td>
<td>1.168</td>
<td>0.497</td>
<td>0.283</td>
<td>0.748</td>
<td>0.933</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.27)</td>
<td>(0.06)</td>
<td>(0.35)</td>
<td>(0.62)</td>
<td>(0.32)</td>
<td>(0.39)</td>
<td>(1.73)</td>
<td>(1.84)</td>
<td>(1.18)</td>
</tr>
<tr>
<td>area_erosion</td>
<td>1.133</td>
<td>0.458</td>
<td>-0.488</td>
<td>0.375</td>
<td>-0.698</td>
<td>-0.223</td>
<td>0.597</td>
<td>0.701</td>
<td>0.944</td>
<td>0.209</td>
</tr>
<tr>
<td></td>
<td>(1.01)</td>
<td>(0.71)</td>
<td>(-0.28)</td>
<td>(0.45)</td>
<td>(-0.37)</td>
<td>(-0.14)</td>
<td>(0.83)</td>
<td>(1.67)</td>
<td>(1.86)</td>
<td>(1.27)</td>
</tr>
<tr>
<td>area_nutrient_dep</td>
<td>0</td>
<td>0.555</td>
<td>0</td>
<td>-8.514*</td>
<td>-2.826</td>
<td>0</td>
<td>0</td>
<td>0.280</td>
<td>-0.596</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.)</td>
<td>(0.12)</td>
<td>(.)</td>
<td>(2.35)</td>
<td>(1.24)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.23)</td>
<td>(0.65)</td>
<td></td>
</tr>
<tr>
<td>Area_salinity_acidity</td>
<td>0.135</td>
<td>0.0696</td>
<td>0.253</td>
<td>0.686</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.671</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.12)</td>
<td>(0.14)</td>
<td>(0.74)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.57)</td>
<td>(0.92)</td>
</tr>
<tr>
<td>area_water_logging</td>
<td>1.237</td>
<td>0.260</td>
<td>-24.30</td>
<td>-1.974</td>
<td>7.092</td>
<td>2.843</td>
<td>0</td>
<td>806.6</td>
<td>1.387</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td>(0.74)</td>
<td>(0.24)</td>
<td>(-0.24)</td>
<td>(-0.28)</td>
<td>(1.86)</td>
<td>(1.23)</td>
<td>(.)</td>
<td>(1.83)</td>
<td>(0.99)</td>
<td>(0.65)</td>
</tr>
<tr>
<td>area_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.394</td>
<td>-3.730**</td>
<td>-4.200*</td>
<td>-0.996</td>
<td>-1.021*</td>
<td>-1.225</td>
<td>0.705</td>
</tr>
<tr>
<td></td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(-0.16)</td>
<td>(-2.87)</td>
<td>(-2.05)</td>
<td>(-1.52)</td>
<td>(-2.31)</td>
<td>(-0.55)</td>
<td>(0.67)</td>
</tr>
<tr>
<td>area_sandy_loam</td>
<td>1.286</td>
<td>-11.04**</td>
<td>0</td>
<td>-1.960</td>
<td>-4.360</td>
<td>-2.540</td>
<td>-3.402</td>
<td>-1.074</td>
<td>-0.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(-3.26)</td>
<td>(.)</td>
<td>(-1.59)</td>
<td>(-0.44)</td>
<td>(-0.44)</td>
<td>(-0.84)</td>
<td>(-1.05)</td>
<td>(.)</td>
<td>(-0.67)</td>
</tr>
<tr>
<td>area_prob_soil</td>
<td>-0.575</td>
<td>-0.454</td>
<td>0.128</td>
<td>-0.840</td>
<td>-17.56*</td>
<td>-13.18**</td>
<td>-0.909</td>
<td>-5.188***</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.32)</td>
<td>(-0.46)</td>
<td>(0.07)</td>
<td>(-1.02)</td>
<td>(-2.51)</td>
<td>(-3.01)</td>
<td>(-0.42)</td>
<td>(-3.60)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>area_deep_black</td>
<td>-2.164</td>
<td>-1.378</td>
<td>-0.0296</td>
<td>-0.177</td>
<td>-1.008</td>
<td>0.206</td>
<td>-0.377</td>
<td>-0.515</td>
<td>-0.912</td>
<td>-0.239</td>
</tr>
<tr>
<td></td>
<td>(-1.78)</td>
<td>(-1.96)</td>
<td>(-0.02)</td>
<td>(-0.20)</td>
<td>(-0.52)</td>
<td>(0.13)</td>
<td>(-0.46)</td>
<td>(-1.17)</td>
<td>(-1.73)</td>
<td>(-1.28)</td>
</tr>
<tr>
<td>area_medium_black</td>
<td>-0.108</td>
<td>-0.265</td>
<td>-0.225</td>
<td>-0.525</td>
<td>-0.438</td>
<td>0.177</td>
<td>-0.581</td>
<td>-1.789***</td>
<td>-0.866</td>
<td>-0.226</td>
</tr>
<tr>
<td></td>
<td>(-0.10)</td>
<td>(-0.44)</td>
<td>(-0.16)</td>
<td>(-0.73)</td>
<td>(-0.23)</td>
<td>(0.11)</td>
<td>(-0.75)</td>
<td>(-3.83)</td>
<td>(-1.69)</td>
<td>(-1.40)</td>
</tr>
<tr>
<td>area_shallow_black_murrum</td>
<td>0.319</td>
<td>0.0334</td>
<td>3.027</td>
<td>3.689*</td>
<td>-1.450</td>
<td>-1.080</td>
<td>-0.656</td>
<td>-0.877</td>
<td>-1.022</td>
<td>-0.271</td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.05)</td>
<td>(0.98)</td>
<td>(2.33)</td>
<td>(-0.76)</td>
<td>(-0.68)</td>
<td>(-0.75)</td>
<td>(-1.49)</td>
<td>(-1.88)</td>
<td>(-1.46)</td>
</tr>
<tr>
<td>area_sandy</td>
<td>-0.387</td>
<td>-0.569</td>
<td>-0.838</td>
<td>-0.440</td>
<td>-6.141</td>
<td>-1.897</td>
<td>-0.961</td>
<td>-0.685</td>
<td>-0.838</td>
<td>-0.340</td>
</tr>
<tr>
<td></td>
<td>(-0.31)</td>
<td>(-0.85)</td>
<td>(-0.54)</td>
<td>(-0.59)</td>
<td>(-1.82)</td>
<td>(-1.08)</td>
<td>(-0.12)</td>
<td>(-1.36)</td>
<td>(-1.40)</td>
<td>(-1.51)</td>
</tr>
<tr>
<td>area_red_soil</td>
<td>0.323</td>
<td>0.0474</td>
<td>-0.429</td>
<td>-0.343</td>
<td>0.613</td>
<td>0.243</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.30)</td>
<td>(0.08)</td>
<td>(-0.32)</td>
<td>(0.47)</td>
<td>(0.32)</td>
<td>(0.15)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
</tbody>
</table>

N = 417  417  276  276  584  584  264  264  426  426

*t statistics in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb_plots</td>
<td>0.0276</td>
<td>0.264</td>
<td>-0.280</td>
<td>0.192</td>
<td>-1.571</td>
<td>-0.645</td>
<td>1.532</td>
<td>-0.0339</td>
<td>-0.0955</td>
<td>-0.0690</td>
</tr>
<tr>
<td></td>
<td>(0.03 )</td>
<td>(0.51 )</td>
<td>(-0.57 )</td>
<td>(0.65 )</td>
<td>(-1.19 )</td>
<td>(-0.68 )</td>
<td>(1.33 )</td>
<td>(-0.08 )</td>
<td>(-0.67 )</td>
<td>(-0.70 )</td>
</tr>
<tr>
<td>nbr_crops</td>
<td>0.324</td>
<td>0.252</td>
<td>0.464*</td>
<td>0.0526</td>
<td>-0.791</td>
<td>-0.224</td>
<td>0.493</td>
<td>0.572**</td>
<td>0.0548</td>
<td>-0.0160</td>
</tr>
<tr>
<td></td>
<td>(0.96 )</td>
<td>(1.19 )</td>
<td>(2.37 )</td>
<td>(0.42 )</td>
<td>(-1.16 )</td>
<td>(-0.60 )</td>
<td>(1.22 )</td>
<td>(2.82 )</td>
<td>(0.77 )</td>
<td>(-0.37 )</td>
</tr>
<tr>
<td>qty_fertilizer</td>
<td>0.0000999</td>
<td>0.000295</td>
<td>0.00140</td>
<td>0.00201</td>
<td>-0.00837</td>
<td>-0.00154</td>
<td>0.0136</td>
<td>0.00930</td>
<td>0.000419</td>
<td>-0.000698</td>
</tr>
<tr>
<td></td>
<td>(0.08 )</td>
<td>(0.28 )</td>
<td>(1.29 )</td>
<td>(2.45 )</td>
<td>(-1.16 )</td>
<td>(-0.29 )</td>
<td>(1.64 )</td>
<td>(1.55 )</td>
<td>(0.84 )</td>
<td>(-0.20 )</td>
</tr>
<tr>
<td>qty_pesticide</td>
<td>0.00371</td>
<td>0.0197</td>
<td>0.000251</td>
<td>0.00877</td>
<td>0.373</td>
<td>0.895**</td>
<td>0.153</td>
<td>0.0699</td>
<td>0.139</td>
<td>0.180*</td>
</tr>
<tr>
<td></td>
<td>(0.04 )</td>
<td>(0.26 )</td>
<td>(0.01 )</td>
<td>(0.58 )</td>
<td>(1.00 )</td>
<td>(3.16 )</td>
<td>(0.64 )</td>
<td>(0.37 )</td>
<td>(1.39 )</td>
<td>(2.39 )</td>
</tr>
<tr>
<td>area_seed_2</td>
<td>-0.432</td>
<td>-0.509</td>
<td>0.143</td>
<td>0.240</td>
<td>-0.232</td>
<td>-0.378</td>
<td>0.160</td>
<td>0.0600</td>
<td>0.0166</td>
<td>-0.0201</td>
</tr>
<tr>
<td></td>
<td>(-0.66 )</td>
<td>(-0.95 )</td>
<td>(0.88 )</td>
<td>(1.85 )</td>
<td>(-0.68 )</td>
<td>(-1.28 )</td>
<td>(0.50 )</td>
<td>(0.25 )</td>
<td>(0.47 )</td>
<td>(-0.78 )</td>
</tr>
<tr>
<td>area_seed_3</td>
<td>0.138</td>
<td>-0.153</td>
<td>-0.636</td>
<td>0.316</td>
<td>-0.793</td>
<td>-0.635</td>
<td>-0.108</td>
<td>-1.283</td>
<td>0.132</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>(0.12 )</td>
<td>(-0.20 )</td>
<td>(-0.39 )</td>
<td>(0.24 )</td>
<td>(-0.77 )</td>
<td>(-0.92 )</td>
<td>(-0.07 )</td>
<td>(-0.92 )</td>
<td>(0.13 )</td>
<td>(0.28 )</td>
</tr>
<tr>
<td>area_seed_4</td>
<td>0.168</td>
<td>-0.294</td>
<td>0</td>
<td>0</td>
<td>-1.002</td>
<td>0.0891</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.32 )</td>
<td>(-0.63 )</td>
<td>()</td>
<td>()</td>
<td>(-0.23 )</td>
<td>(0.02 )</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>L.log_income</td>
<td>-0.0670</td>
<td>-0.139</td>
<td>0.148</td>
<td>0.0756</td>
<td>0.205</td>
<td>-0.0781</td>
<td>2.162*</td>
<td>0.262</td>
<td>0.0314</td>
<td>0.171*</td>
</tr>
<tr>
<td></td>
<td>(-0.09 )</td>
<td>(-0.30 )</td>
<td>(0.43 )</td>
<td>(0.39 )</td>
<td>(0.11 )</td>
<td>(-0.08 )</td>
<td>(2.22 )</td>
<td>(0.65 )</td>
<td>(0.15 )</td>
<td>(2.07 )</td>
</tr>
<tr>
<td>L.log_borrowings_tot</td>
<td>0.436</td>
<td>0.292</td>
<td>-0.0323</td>
<td>-0.0795</td>
<td>1.727</td>
<td>1.162*</td>
<td>-0.366</td>
<td>-0.411</td>
<td>0.0250</td>
<td>0.0677</td>
</tr>
<tr>
<td></td>
<td>(0.86 )</td>
<td>(0.94 )</td>
<td>(-0.14 )</td>
<td>(-0.60 )</td>
<td>(1.62 )</td>
<td>(2.22 )</td>
<td>(-0.90 )</td>
<td>(-1.92 )</td>
<td>(0.24 )</td>
<td>(1.38 )</td>
</tr>
<tr>
<td>L.log_gov_benefit</td>
<td>-0.371</td>
<td>-0.00802</td>
<td>-0.0573</td>
<td>-0.101</td>
<td>-0.613</td>
<td>-0.877</td>
<td>-0.731</td>
<td>-0.209</td>
<td>0.00964</td>
<td>0.0402</td>
</tr>
<tr>
<td></td>
<td>(-1.02 )</td>
<td>(-0.03 )</td>
<td>(-0.25 )</td>
<td>(-0.60 )</td>
<td>(-0.84 )</td>
<td>(-1.66 )</td>
<td>(-1.33 )</td>
<td>(-0.63 )</td>
<td>(0.12 )</td>
<td>(0.81 )</td>
</tr>
<tr>
<td>rain</td>
<td>0.0673</td>
<td>0.397</td>
<td>0.00652</td>
<td>0.0398</td>
<td>0.993*</td>
<td>0.557*</td>
<td>-0.282</td>
<td>-0.276</td>
<td>-0.373***</td>
<td>-0.374***</td>
</tr>
<tr>
<td></td>
<td>(0.10 )</td>
<td>(0.69 )</td>
<td>(0.08 )</td>
<td>(0.57 )</td>
<td>(2.29 )</td>
<td>(1.97 )</td>
<td>(-1.57 )</td>
<td>(-1.91 )</td>
<td>(-3.73 )</td>
<td>(-5.97 )</td>
</tr>
<tr>
<td>2012.year</td>
<td>-21.82</td>
<td>-88.62</td>
<td>1.338</td>
<td>-5.576</td>
<td>-204.6*</td>
<td>-104.7</td>
<td>-44.86</td>
<td>-39.86</td>
<td>64.07***</td>
<td>64.15***</td>
</tr>
<tr>
<td></td>
<td>(-1.16 )</td>
<td>(-0.71 )</td>
<td>(0.08 )</td>
<td>(-0.37 )</td>
<td>(-2.18 )</td>
<td>(-1.71 )</td>
<td>(-1.75 )</td>
<td>(-1.94 )</td>
<td>(3.96 )</td>
<td>(6.44 )</td>
</tr>
<tr>
<td>2013.year</td>
<td>-53.49</td>
<td>-271.0</td>
<td>1.105</td>
<td>-21.91</td>
<td>-647.6*</td>
<td>-345.6</td>
<td>36.62</td>
<td>37.83</td>
<td>202.6***</td>
<td>203.8***</td>
</tr>
<tr>
<td></td>
<td>(-0.12 )</td>
<td>(-0.69 )</td>
<td>(0.02 )</td>
<td>(-0.46 )</td>
<td>(-2.18 )</td>
<td>(-1.79 )</td>
<td>(1.37 )</td>
<td>(1.77 )</td>
<td>(3.77 )</td>
<td>(6.07 )</td>
</tr>
<tr>
<td>2014.year</td>
<td>-15.64</td>
<td>-11.61</td>
<td>4.052</td>
<td>2.371</td>
<td>-37.64</td>
<td>-16.74</td>
<td>20.00</td>
<td>20.94</td>
<td>29.21***</td>
<td>29.37***</td>
</tr>
<tr>
<td></td>
<td>(-0.76 )</td>
<td>(-0.75 )</td>
<td>(1.52 )</td>
<td>(1.05 )</td>
<td>(-1.47 )</td>
<td>(-0.95 )</td>
<td>(1.44 )</td>
<td>(1.93 )</td>
<td>(3.95 )</td>
<td>(6.55 )</td>
</tr>
<tr>
<td>N</td>
<td>417</td>
<td>417</td>
<td>276</td>
<td>276</td>
<td>584</td>
<td>584</td>
<td>264</td>
<td>264</td>
<td>426</td>
<td>426</td>
</tr>
</tbody>
</table>

1. statistics in parentheses
2. * p < 0.05, ** p < 0.01, *** p < 0.001
Table 11.4: Regression panel data fixed effects
Change in output rate (Page 1/2)

<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.log_output_rate</td>
<td>-1.413***</td>
<td>-0.982***</td>
<td>-1.448***</td>
<td>-1.111***</td>
<td>-1.506***</td>
<td>-1.069***</td>
<td>-1.104***</td>
<td>-1.006***</td>
<td>-1.497***</td>
<td>-0.908***</td>
</tr>
<tr>
<td></td>
<td>(-22.21)</td>
<td>(-21.92)</td>
<td>(-20.59)</td>
<td>(-12.78)</td>
<td>(-17.15)</td>
<td>(-18.80)</td>
<td>(-10.38)</td>
<td>(-15.95)</td>
<td>(-14.23)</td>
<td>(-18.99)</td>
</tr>
<tr>
<td>age_h1</td>
<td>-0.00812</td>
<td>0.000229</td>
<td>0.0137</td>
<td>-0.000349</td>
<td>0.00133</td>
<td>-0.00244</td>
<td>-0.00103</td>
<td>-0.000186</td>
<td>-0.000779</td>
<td>-0.000537</td>
</tr>
<tr>
<td></td>
<td>(-0.18)</td>
<td>(0.71)</td>
<td>(1.74)</td>
<td>(-0.56)</td>
<td>(0.19)</td>
<td>(-0.32)</td>
<td>(-0.22)</td>
<td>(-0.21)</td>
<td>(-0.34)</td>
<td>(-1.52)</td>
</tr>
<tr>
<td>area</td>
<td>-0.00133</td>
<td>0.000250</td>
<td>0.0148</td>
<td>0.00172</td>
<td>-0.00469</td>
<td>-0.00452</td>
<td>-0.00122</td>
<td>0.00116</td>
<td>-0.00659</td>
<td>-0.00110</td>
</tr>
<tr>
<td></td>
<td>(-0.16)</td>
<td>(0.04)</td>
<td>(0.53)</td>
<td>(0.10)</td>
<td>(-0.44)</td>
<td>(-0.82)</td>
<td>(-0.86)</td>
<td>(0.14)</td>
<td>(-1.46)</td>
<td>(-0.42)</td>
</tr>
<tr>
<td>irri_percent</td>
<td>0.000124</td>
<td>0.0000725</td>
<td>0.000167</td>
<td>0.000296</td>
<td>-0.000443</td>
<td>0.000270</td>
<td>0.000244</td>
<td>-0.00137</td>
<td>0.000303</td>
<td>-0.000312</td>
</tr>
<tr>
<td></td>
<td>(0.72)</td>
<td>(0.65)</td>
<td>(0.19)</td>
<td>(0.10)</td>
<td>(-0.74)</td>
<td>(0.77)</td>
<td>(0.53)</td>
<td>(-0.51)</td>
<td>(0.59)</td>
<td>(-0.15)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>-0.00320</td>
<td>-0.000532</td>
<td>-0.238</td>
<td>-0.111</td>
<td>0</td>
<td>0.00960</td>
<td>0</td>
<td>0.315***</td>
<td>-0.0642</td>
<td>0.00982</td>
</tr>
<tr>
<td></td>
<td>(-0.10)</td>
<td>(-0.03)</td>
<td>(-1.94)</td>
<td>(-1.92)</td>
<td>(.)</td>
<td>(0.20)</td>
<td>(.)</td>
<td>(4.97)</td>
<td>(-0.64)</td>
<td>(0.31)</td>
</tr>
<tr>
<td>area_no_degra</td>
<td>-0.0326</td>
<td>0.00203</td>
<td>0.0171</td>
<td>0.0286</td>
<td>0.0195</td>
<td>0.0277</td>
<td>0.0428</td>
<td>-0.00412</td>
<td>0.0656</td>
<td>-0.00190</td>
</tr>
<tr>
<td></td>
<td>(-0.69)</td>
<td>(0.27)</td>
<td>(-0.24)</td>
<td>(0.83)</td>
<td>(0.64)</td>
<td>(1.31)</td>
<td>(0.66)</td>
<td>(-0.21)</td>
<td>(1.72)</td>
<td>(-0.15)</td>
</tr>
<tr>
<td>area_erosion</td>
<td>-0.0309</td>
<td>0.00189</td>
<td>-0.0132</td>
<td>0.0509</td>
<td>0.00664</td>
<td>0.0151</td>
<td>0.0267</td>
<td>0.00294</td>
<td>0.0660</td>
<td>-0.000313</td>
</tr>
<tr>
<td></td>
<td>(-0.65)</td>
<td>(0.24)</td>
<td>(-0.13)</td>
<td>(1.17)</td>
<td>(0.21)</td>
<td>(0.71)</td>
<td>(0.42)</td>
<td>(0.15)</td>
<td>(1.72)</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>area_nutrient_dep</td>
<td>0</td>
<td>-0.000514</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0236</td>
<td>0</td>
<td>0</td>
<td>0.0292</td>
<td>-0.0519</td>
</tr>
<tr>
<td></td>
<td>(.)</td>
<td>(-0.01)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.68)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.35)</td>
<td>(-0.91)</td>
</tr>
<tr>
<td>Area_salinity_acidity</td>
<td>0.00188</td>
<td>-0.00255</td>
<td>0.0383</td>
<td>0.0109</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0764</td>
<td>-0.0158</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(-0.39)</td>
<td>(0.31)</td>
<td>(0.24)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.35)</td>
<td>(-0.36)</td>
</tr>
<tr>
<td>area_water_logging</td>
<td>-0.0718</td>
<td>0.00154</td>
<td>0</td>
<td>0</td>
<td>-0.319</td>
<td>0.00179</td>
<td>0</td>
<td>0</td>
<td>0.0926</td>
<td>-0.00476</td>
</tr>
<tr>
<td></td>
<td>(-1.04)</td>
<td>(0.12)</td>
<td>(.)</td>
<td>(.)</td>
<td>(-0.21)</td>
<td>(0.05)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.98)</td>
<td>(-0.30)</td>
</tr>
<tr>
<td>area_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.0626</td>
<td>-0.141</td>
<td>-0.113***</td>
<td>-0.0518</td>
<td>-0.00533</td>
<td>-0.139</td>
<td>0.0438</td>
</tr>
<tr>
<td></td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(-0.56)</td>
<td>(-0.82)</td>
<td>(-0.8)</td>
<td>(-0.26)</td>
<td>(-0.26)</td>
<td>(-0.92)</td>
<td>(0.66)</td>
</tr>
<tr>
<td>area_sandy_loam</td>
<td>-0.0303</td>
<td>-0.0113</td>
<td>0</td>
<td>-0.266***</td>
<td>-0.00507</td>
<td>0.0115</td>
<td>0.0370</td>
<td>0.0669</td>
<td>0</td>
<td>0.00243</td>
</tr>
<tr>
<td></td>
<td>(-0.33)</td>
<td>(-0.27)</td>
<td>(.)</td>
<td>(-4.56)</td>
<td>(-0.03)</td>
<td>(0.14)</td>
<td>(0.24)</td>
<td>(0.58)</td>
<td>(.)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>area_prob_soil</td>
<td>0.0487</td>
<td>-0.00397</td>
<td>0.0159</td>
<td>-0.0377</td>
<td>0</td>
<td>0</td>
<td>-0.495**</td>
<td>-0.346***</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.81)</td>
<td>(-0.36)</td>
<td>(0.14)</td>
<td>(-0.89)</td>
<td>(.)</td>
<td>(.)</td>
<td>(-2.98)</td>
<td>(-3.43)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>area_deep_black</td>
<td>0.0281</td>
<td>-0.00267</td>
<td>0.0496</td>
<td>0.00678</td>
<td>-0.0194</td>
<td>-0.0263</td>
<td>-0.0421</td>
<td>0.00294</td>
<td>-0.0610</td>
<td>0.00248</td>
</tr>
<tr>
<td></td>
<td>(0.58)</td>
<td>(-0.30)</td>
<td>(0.58)</td>
<td>(0.16)</td>
<td>(-0.61)</td>
<td>(-1.20)</td>
<td>(-0.65)</td>
<td>(0.12)</td>
<td>(-1.57)</td>
<td>(0.17)</td>
</tr>
<tr>
<td>area_medium_black</td>
<td>0.0344</td>
<td>-0.00221</td>
<td>-0.0289</td>
<td>-0.0571</td>
<td>-0.0124</td>
<td>-0.0147</td>
<td>-0.0931</td>
<td>-0.0208</td>
<td>-0.0605</td>
<td>0.000807</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
<td>(-0.30)</td>
<td>(-0.40)</td>
<td>(-1.64)</td>
<td>(-0.40)</td>
<td>(-0.69)</td>
<td>(-1.01)</td>
<td>(-0.95)</td>
<td>(-1.57)</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>

N = 293 293 203 203 356 356 185 185 307 307

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>area_shallow_black_murrum</td>
<td>0.0369</td>
<td>0.00332</td>
<td>0.0380</td>
<td>0.222*</td>
<td>-0.0147</td>
<td>-0.0189</td>
<td>-0.0563</td>
<td>0.00918</td>
<td>-0.0748</td>
<td>0.000640</td>
</tr>
<tr>
<td></td>
<td>(0.78)</td>
<td>(0.43)</td>
<td>(0.22)</td>
<td>(2.55)</td>
<td>(-0.43)</td>
<td>(-0.83)</td>
<td>(-0.77)</td>
<td>(0.31)</td>
<td>(-1.81)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>area_sandy</td>
<td>0.0255</td>
<td>-0.00801</td>
<td>-0.0750</td>
<td>-0.0209</td>
<td>0.316</td>
<td>-0.0135</td>
<td>-0.0491</td>
<td>-0.0122</td>
<td>-0.0585</td>
<td>-0.00168</td>
</tr>
<tr>
<td></td>
<td>(0.53)</td>
<td>(-1.01)</td>
<td>(-0.89)</td>
<td>(-0.58)</td>
<td>(0.21)</td>
<td>(-0.42)</td>
<td>(-0.76)</td>
<td>(-0.59)</td>
<td>(-3.36)</td>
<td>(-0.10)</td>
</tr>
<tr>
<td>area_red_soil</td>
<td>0.0411</td>
<td>0.00169</td>
<td>-0.00497</td>
<td>-0.0317</td>
<td>0.00519</td>
<td>-0.0121</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.86)</td>
<td>(0.22)</td>
<td>(-0.07)</td>
<td>(-0.91)</td>
<td>(0.13)</td>
<td>(-0.53)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>nb_plots</td>
<td>0.0104</td>
<td>0.00437</td>
<td>0.000459</td>
<td>0.00577</td>
<td>0.0151</td>
<td>-0.0139</td>
<td>0.0979**</td>
<td>0.0229</td>
<td>-0.00138</td>
<td>-0.00653</td>
</tr>
<tr>
<td></td>
<td>(0.84)</td>
<td>(0.82)</td>
<td>(0.35)</td>
<td>(0.59)</td>
<td>(-0.89)</td>
<td>(2.97)</td>
<td>(1.25)</td>
<td>(-1.3)</td>
<td>(-0.96)</td>
<td>(-0.96)</td>
</tr>
<tr>
<td>nbr_crops</td>
<td>-0.00213</td>
<td>0.00253</td>
<td>0.00747</td>
<td>0.0267</td>
<td>-0.0118</td>
<td>-0.00727</td>
<td>0.00439</td>
<td>-0.00786</td>
<td>0.0402</td>
<td>0.0445</td>
</tr>
<tr>
<td></td>
<td>(-0.50)</td>
<td>(1.04)</td>
<td>(0.56)</td>
<td>(0.37)</td>
<td>(-1.09)</td>
<td>(-1.3)</td>
<td>(0.37)</td>
<td>(-1.39)</td>
<td>(0.77)</td>
<td>(1.45)</td>
</tr>
<tr>
<td>qty_fertilizer</td>
<td>-0.00000200</td>
<td>0.0000104</td>
<td>0.0000600</td>
<td>0.000496</td>
<td>-0.000923</td>
<td>-0.0000619</td>
<td>0.000332</td>
<td>0.000223</td>
<td>0.0000919</td>
<td>0.0000160</td>
</tr>
<tr>
<td></td>
<td>(-0.14)</td>
<td>(0.89)</td>
<td>(0.84)</td>
<td>(0.93)</td>
<td>(-0.57)</td>
<td>(-0.66)</td>
<td>(1.49)</td>
<td>(1.38)</td>
<td>(0.25)</td>
<td>(0.67)</td>
</tr>
<tr>
<td>qty_pesticide</td>
<td>-0.000923</td>
<td>-0.000526</td>
<td>-0.000639</td>
<td>0.000732</td>
<td>0.00321</td>
<td>-0.0204</td>
<td>-0.00109</td>
<td>0.0987</td>
<td>0.0166**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-0.75)</td>
<td>(-0.57)</td>
<td>(-0.44)</td>
<td>(0.93)</td>
<td>(0.40)</td>
<td>(-1.6)</td>
<td>(-0.24)</td>
<td>(-0.23)</td>
<td>(1.32)</td>
<td>(3.24)</td>
</tr>
<tr>
<td>area_seed_2</td>
<td>-0.00174</td>
<td>-0.000595</td>
<td>0.00377</td>
<td>0.00370</td>
<td>0.00493</td>
<td>0.00491</td>
<td>0.0152</td>
<td>0.000572</td>
<td>0.00110</td>
<td>-0.00126</td>
</tr>
<tr>
<td></td>
<td>(-0.22)</td>
<td>(-0.09)</td>
<td>(0.38)</td>
<td>(0.47)</td>
<td>(0.07)</td>
<td>(0.99)</td>
<td>(1.51)</td>
<td>(0.09)</td>
<td>(0.46)</td>
<td>(-0.75)</td>
</tr>
<tr>
<td>area_seed_3</td>
<td>-0.00385</td>
<td>-0.00450</td>
<td>0.0369</td>
<td>0.0318</td>
<td>0.00176</td>
<td>-0.00594</td>
<td>0.0268</td>
<td>-0.00336</td>
<td>-0.0559</td>
<td>0.00185</td>
</tr>
<tr>
<td></td>
<td>(-0.32)</td>
<td>(-0.46)</td>
<td>(0.39)</td>
<td>(0.50)</td>
<td>(-0.58)</td>
<td>(0.14)</td>
<td>(-0.03)</td>
<td>(-0.47)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>area_seed_4</td>
<td>0.000636</td>
<td>-0.00221</td>
<td>0</td>
<td>0</td>
<td>0.0298</td>
<td>0.0125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(-0.40)</td>
<td>(.)</td>
<td>(.)</td>
<td>(0.48)</td>
<td>(0.24)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>L.log_income</td>
<td>0.0104</td>
<td>-0.0102</td>
<td>-0.0177</td>
<td>-0.0114</td>
<td>-0.00365</td>
<td>-0.0129</td>
<td>0.0455</td>
<td>0.0145</td>
<td>0.00752</td>
<td>0.01000</td>
</tr>
<tr>
<td></td>
<td>(0.91)</td>
<td>(-1.74)</td>
<td>(-0.76)</td>
<td>(-1.06)</td>
<td>(-0.86)</td>
<td>(-0.86)</td>
<td>(1.58)</td>
<td>(1.19)</td>
<td>(0.42)</td>
<td>(1.53)</td>
</tr>
<tr>
<td>L.log_borrowings_tot</td>
<td>0.00444</td>
<td>0.00699</td>
<td>-0.0113</td>
<td>0.00577</td>
<td>0.0152</td>
<td>0.00624</td>
<td>0.00800</td>
<td>-0.0661</td>
<td>0.00375</td>
<td>0.00351</td>
</tr>
<tr>
<td></td>
<td>(0.62)</td>
<td>(1.85)</td>
<td>(-0.69)</td>
<td>(0.82)</td>
<td>(0.77)</td>
<td>(0.68)</td>
<td>(0.61)</td>
<td>(-0.99)</td>
<td>(0.42)</td>
<td>(0.94)</td>
</tr>
<tr>
<td>L.log_gov_benefit</td>
<td>-0.00488</td>
<td>-0.00260</td>
<td>-0.00748</td>
<td>-0.00760</td>
<td>-0.0130</td>
<td>-0.00303</td>
<td>-0.0287</td>
<td>-0.00851</td>
<td>0.00725</td>
<td>-0.00203</td>
</tr>
<tr>
<td></td>
<td>(-0.99)</td>
<td>(-0.83)</td>
<td>(-0.49)</td>
<td>(-0.80)</td>
<td>(-0.98)</td>
<td>(-0.36)</td>
<td>(-1.66)</td>
<td>(-0.83)</td>
<td>(1.24)</td>
<td>(-0.55)</td>
</tr>
<tr>
<td>rain</td>
<td>0.0324*</td>
<td>0.0193</td>
<td>-0.00604</td>
<td>-0.00113**</td>
<td>-0.00222</td>
<td>0.00344</td>
<td>-0.00283</td>
<td>-0.00314</td>
<td>-0.0175</td>
<td>-0.0164*</td>
</tr>
<tr>
<td></td>
<td>(2.26)</td>
<td>(1.75)</td>
<td>(-1.00)</td>
<td>(-2.63)</td>
<td>(-0.04)</td>
<td>(0.97)</td>
<td>(-0.58)</td>
<td>(-0.83)</td>
<td>(-1.45)</td>
<td>(-2.38)</td>
</tr>
<tr>
<td>2012.year</td>
<td>-7.053*</td>
<td>-4.254</td>
<td>0.352**</td>
<td>0.441***</td>
<td>-0.228</td>
<td>1.685</td>
<td>-0.427</td>
<td>-0.460</td>
<td>3.105</td>
<td>2.932**</td>
</tr>
<tr>
<td></td>
<td>(-2.26)</td>
<td>(-1.78)</td>
<td>(2.78)</td>
<td>(4.73)</td>
<td>(-0.08)</td>
<td>(0.83)</td>
<td>(-1.27)</td>
<td>(-1.77)</td>
<td>(1.63)</td>
<td>(2.74)</td>
</tr>
<tr>
<td>2013.year</td>
<td>-22.19*</td>
<td>-13.20</td>
<td>0.989*</td>
<td>1.201***</td>
<td>0.215</td>
<td>0.584</td>
<td>0.0652</td>
<td>0.128</td>
<td>9.675</td>
<td>8.959*</td>
</tr>
<tr>
<td></td>
<td>(-2.25)</td>
<td>(-1.74)</td>
<td>(2.59)</td>
<td>(4.21)</td>
<td>(0.42)</td>
<td>(1.58)</td>
<td>(0.09)</td>
<td>(0.23)</td>
<td>(1.50)</td>
<td>(2.45)</td>
</tr>
<tr>
<td>2014.year</td>
<td>-0.851</td>
<td>-0.529</td>
<td>0.415***</td>
<td>0.236***</td>
<td>0.351</td>
<td>0.0942</td>
<td>0.0541</td>
<td>0.0444</td>
<td>1.435</td>
<td>1.331**</td>
</tr>
<tr>
<td></td>
<td>(-1.90)</td>
<td>(-1.76)</td>
<td>(3.95)</td>
<td>(3.55)</td>
<td>(0.88)</td>
<td>(0.37)</td>
<td>(0.13)</td>
<td>(0.15)</td>
<td>(1.70)</td>
<td>(2.78)</td>
</tr>
</tbody>
</table>

N = 293

T statistics in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
Table 11.6: Regression panel data fixed effects
Change in yield gap (Page 1/2)

<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>L yg max village</td>
<td>-0.550</td>
<td>-0.232</td>
<td>-2.484</td>
<td>**</td>
<td>-0.893</td>
<td>-0.267</td>
<td>***</td>
<td>-0.253</td>
<td>***</td>
<td>-0.414</td>
</tr>
<tr>
<td></td>
<td>(-1.73)</td>
<td>(-1.92)</td>
<td>(-3.14)</td>
<td>(-4.44)</td>
<td>(-5.24)</td>
<td>(-4.31)</td>
<td>(-4.17)</td>
<td>(-5.95)</td>
<td>(-4.17)</td>
<td></td>
</tr>
<tr>
<td>age_h1</td>
<td>0.651</td>
<td>0.0231</td>
<td>1.013</td>
<td>0.00462</td>
<td>-0.188</td>
<td>0.0478</td>
<td>-0.0251</td>
<td>-0.00840</td>
<td>0.0303</td>
<td>0.0152</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.38)</td>
<td>(0.13)</td>
<td>(0.97)</td>
<td>(0.86)</td>
<td>(0.27)</td>
<td>(1.31)</td>
<td>(2.27)</td>
<td>(1.27)</td>
<td></td>
</tr>
<tr>
<td>area</td>
<td>0.770</td>
<td>0.0170</td>
<td>-0.254</td>
<td>-0.0887</td>
<td>0.540</td>
<td>0.405</td>
<td>0.497</td>
<td>0.0815</td>
<td>0.370</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.02)</td>
<td>(-0.07)</td>
<td>(-1.45)</td>
<td>(1.48)</td>
<td>(1.70)</td>
<td>(0.68)</td>
<td>(3.37)</td>
<td>(2.27)</td>
<td></td>
</tr>
<tr>
<td>irri_percent</td>
<td>0.0661</td>
<td>0.0392</td>
<td>0.0268</td>
<td>-0.00709</td>
<td>0.0434</td>
<td>0.0289</td>
<td>0.00405</td>
<td>-0.00385</td>
<td>-0.0671</td>
<td>-0.00463</td>
</tr>
<tr>
<td></td>
<td>(2.29)</td>
<td>(2.91)</td>
<td>(0.26)</td>
<td>(-0.45)</td>
<td>(1.99)</td>
<td>(1.67)</td>
<td>(0.38)</td>
<td>(-0.85)</td>
<td>(-0.91)</td>
<td>(-0.76)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>0.537</td>
<td>-1.178</td>
<td>1.514</td>
<td>-2.895</td>
<td>-908.5</td>
<td>**</td>
<td>-8.663</td>
<td>**</td>
<td>-8.663</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(-0.28)</td>
<td>(0.18)</td>
<td>(-1.11)</td>
<td>(-3.60)</td>
<td>(-2.91)</td>
<td>(-0.41)</td>
<td>(-0.44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area_no_degra</td>
<td>1.852</td>
<td>0.426</td>
<td>-0.597</td>
<td>-1.457</td>
<td>-0.569</td>
<td>-5.027</td>
<td>0.0254</td>
<td>-0.112</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.44)</td>
<td>(-0.11)</td>
<td>(-0.73)</td>
<td>(-0.56)</td>
<td>(-2.20)</td>
<td>(0.06)</td>
<td>(-0.32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area_erosion</td>
<td>0.683</td>
<td>0.517</td>
<td>15.23</td>
<td>-5.672</td>
<td>0.0213</td>
<td>-0.122</td>
<td>-0.298</td>
<td>-0.211</td>
<td>-0.297</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.51)</td>
<td>(1.16)</td>
<td>(-2.29)</td>
<td>(0.02)</td>
<td>(-0.12)</td>
<td>(-0.52)</td>
<td>(-0.47)</td>
<td>(-0.83)</td>
<td></td>
</tr>
<tr>
<td>area_nutrient_dep</td>
<td>0</td>
<td>-0.306</td>
<td>0</td>
<td>0</td>
<td>9.473</td>
<td>**</td>
<td>5.256</td>
<td>**</td>
<td>5.256</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>(-.17)</td>
<td>(-.17)</td>
<td>(-.17)</td>
<td>(-.17)</td>
<td>(4.07)</td>
<td>(3.06)</td>
<td>(-.17)</td>
<td>(-.17)</td>
<td>(0.60)</td>
<td>(0.31)</td>
</tr>
<tr>
<td>area_salinity_acidity</td>
<td>1.050</td>
<td>0.575</td>
<td>10.54</td>
<td>-2.446</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
<td>(0.70)</td>
<td>(0.82)</td>
<td>(-0.97)</td>
<td>(-0.97)</td>
<td>(-0.97)</td>
<td>(-0.97)</td>
<td>(-0.97)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area_water_logging</td>
<td>-0.217</td>
<td>1.961</td>
<td>0</td>
<td>0</td>
<td>-5.810</td>
<td>-5.078</td>
<td>0</td>
<td>0</td>
<td>-0.254</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-0.01)</td>
<td>(0.86)</td>
<td>(-.17)</td>
<td>(-1.84)</td>
<td>(-2.46)</td>
<td>(-2.46)</td>
<td>(-.17)</td>
<td>(-1.84)</td>
<td>(-0.54)</td>
<td></td>
</tr>
<tr>
<td>area_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.762</td>
<td>*</td>
<td>3.230</td>
<td>*</td>
<td>3.230</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(2.23)</td>
<td>(1.93)</td>
<td>(2.23)</td>
<td>(1.93)</td>
<td>(2.93)</td>
<td>(2.93)</td>
</tr>
<tr>
<td>area_sandy_clay_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
</tr>
<tr>
<td>area_clay</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(-.01)</td>
</tr>
<tr>
<td>area_sandy_loam</td>
<td>-15.93</td>
<td>-0.682</td>
<td>0</td>
<td>0.772</td>
<td>15.59</td>
<td>*</td>
<td>11.36</td>
<td>*</td>
<td>15.42</td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td>(-0.72)</td>
<td>(-0.13)</td>
<td>(-.01)</td>
<td>(0.26)</td>
<td>(2.48)</td>
<td>(2.28)</td>
<td>(3.14)</td>
<td>(0.79)</td>
<td>(-0.03)</td>
<td></td>
</tr>
<tr>
<td>area_prob_soil</td>
<td>-0.0131</td>
<td>-0.957</td>
<td>-17.03</td>
<td>4.893</td>
<td>0</td>
<td>0</td>
<td>5.466</td>
<td>0.291</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.00)</td>
<td>(-.65)</td>
<td>(-1.29)</td>
<td>(2.08)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(1.91)</td>
<td>(0.37)</td>
<td>(-.01)</td>
<td></td>
</tr>
<tr>
<td>area_deep_black</td>
<td>0.802</td>
<td>-0.111</td>
<td>0</td>
<td>0</td>
<td>0.974</td>
<td>0.486</td>
<td>4.453</td>
<td>0.201</td>
<td>-0.0320</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(-.10)</td>
<td>(-.01)</td>
<td>(-.01)</td>
<td>(0.86)</td>
<td>(0.67)</td>
<td>(2.00)</td>
<td>(0.67)</td>
<td>(-0.07)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>area_medium_black</td>
<td>-0.602</td>
<td>-0.0331</td>
<td>0.453</td>
<td>2.365</td>
<td>-0.260</td>
<td>0.173</td>
<td>4.303</td>
<td>0.232</td>
<td>-0.0342</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>(-0.03)</td>
<td>(-0.03)</td>
<td>(0.09)</td>
<td>(1.18)</td>
<td>(-0.21)</td>
<td>(0.16)</td>
<td>(1.96)</td>
<td>(0.43)</td>
<td>(-0.08)</td>
<td>(0.29)</td>
</tr>
</tbody>
</table>

N 267 267 132 132 343 343 167 167 209 209

*t statistics in parentheses
*p < 0.05, **p < 0.01, ***p < 0.001
Table 11.7: Regression panel data fixed effects
Change in yield gap (Page 2/2)

<table>
<thead>
<tr>
<th></th>
<th>kc_fe</th>
<th>kc_re</th>
<th>kp_fe</th>
<th>kp_re</th>
<th>kpp_fe</th>
<th>kpp_re</th>
<th>rs_fe</th>
<th>rs_re</th>
<th>rw_fe</th>
<th>rw_re</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.yg_max_village</td>
<td>-0.550</td>
<td>-0.232</td>
<td>-2.484 *</td>
<td>-0.893 *</td>
<td>-0.257 ***</td>
<td>-0.253 ***</td>
<td>-0.414 ***</td>
<td>-0.0862 **</td>
<td>-0.690 ***</td>
<td>-0.374 ***</td>
</tr>
<tr>
<td></td>
<td>(-1.73)</td>
<td>(-1.92)</td>
<td>(-3.14)</td>
<td>(-2.20)</td>
<td>(-4.44)</td>
<td>(-5.24)</td>
<td>(-4.31)</td>
<td>(-2.67)</td>
<td>(-5.59)</td>
<td>(-4.17)</td>
</tr>
<tr>
<td>age_h1</td>
<td>0.651</td>
<td>0.0231</td>
<td>1.013</td>
<td>0.00462</td>
<td>-0.188</td>
<td>0.0478</td>
<td>-0.0251</td>
<td>-0.00840</td>
<td>0.0303</td>
<td>0.0152</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.38)</td>
<td>(0.13)</td>
<td>(0.97)</td>
<td>(0.86)</td>
<td>(0.27)</td>
<td>(1.31)</td>
<td>(1.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area</td>
<td>0.770</td>
<td>0.0170</td>
<td>-0.254</td>
<td>-0.0887</td>
<td>0.540</td>
<td>0.405</td>
<td>0.497</td>
<td>0.0815</td>
<td>0.370 **</td>
<td>0.160 *</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.02)</td>
<td>(-0.07)</td>
<td>(1.45)</td>
<td>(1.48)</td>
<td>(1.70)</td>
<td>(0.68)</td>
<td>(3.37)</td>
<td>(2.27)</td>
<td></td>
</tr>
<tr>
<td>irri_percent</td>
<td>0.0661 *</td>
<td>0.0392 **</td>
<td>0.0268</td>
<td>-0.00709</td>
<td>0.0434 *</td>
<td>0.0289</td>
<td>0.00405</td>
<td>-0.00385</td>
<td>-0.00671</td>
<td>-0.00463</td>
</tr>
<tr>
<td></td>
<td>(2.29)</td>
<td>(2.91)</td>
<td>(0.26)</td>
<td>(-0.45)</td>
<td>(1.99)</td>
<td>(1.67)</td>
<td>(0.38)</td>
<td>(-0.85)</td>
<td>(-0.91)</td>
<td>(-0.76)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>0.537</td>
<td>-1.178</td>
<td>1.514</td>
<td>-2.895</td>
<td>-908.5 ***</td>
<td>-8.663 ***</td>
<td>0</td>
<td>-0.570</td>
<td>0</td>
<td>-0.486</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(-0.58)</td>
<td>(0.18)</td>
<td>(-1.11)</td>
<td>(-3.60)</td>
<td>(-2.91)</td>
<td>(0.41)</td>
<td>(0.44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area_no_degra</td>
<td>1.852</td>
<td>0.426</td>
<td>-0.597</td>
<td>-1.457</td>
<td>-0.570</td>
<td>-0.609</td>
<td>-5.027 *</td>
<td>-0.395</td>
<td>0.0254</td>
<td>-0.112</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.44)</td>
<td>(-0.11)</td>
<td>(-0.73)</td>
<td>(-0.47)</td>
<td>(-0.56)</td>
<td>(-2.20)</td>
<td>(-0.69)</td>
<td>(0.06)</td>
<td>(-0.32)</td>
</tr>
<tr>
<td>area_erosion</td>
<td>0.683</td>
<td>0.517</td>
<td>15.23</td>
<td>-5.672 *</td>
<td>0.0213</td>
<td>-0.122</td>
<td>-5.194 *</td>
<td>-0.298</td>
<td>-0.211</td>
<td>-0.297</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.51)</td>
<td>(1.16)</td>
<td>(-2.29)</td>
<td>(0.02)</td>
<td>(-0.11)</td>
<td>(-2.35)</td>
<td>(-0.52)</td>
<td>(-0.47)</td>
<td>(-0.83)</td>
</tr>
<tr>
<td>area_nutrient_dep</td>
<td>0</td>
<td>-0.863</td>
<td>0</td>
<td>0</td>
<td>9.473 ***</td>
<td>5.256 **</td>
<td>0</td>
<td>0</td>
<td>0.531</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>(...)</td>
<td>(-0.17)</td>
<td>(...)</td>
<td>(...)</td>
<td>(4.07)</td>
<td>(3.06)</td>
<td>(...)</td>
<td>(...)</td>
<td>(0.60)</td>
<td>(0.31)</td>
</tr>
<tr>
<td>area_salinity_acidity</td>
<td>1.050</td>
<td>0.575</td>
<td>10.54</td>
<td>-2.446</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
<td>(0.70)</td>
<td>(0.82)</td>
<td>(-0.97)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>area_water_logging</td>
<td>-0.217</td>
<td>1.961</td>
<td>0</td>
<td>0</td>
<td>-5.810</td>
<td>-5.078 *</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.254</td>
</tr>
<tr>
<td></td>
<td>(-0.01)</td>
<td>(0.86)</td>
<td>(...)</td>
<td>(...)</td>
<td>(-1.84)</td>
<td>(-2.46)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(-0.54)</td>
</tr>
<tr>
<td>area_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.762 *</td>
<td>3.330</td>
<td>5.245 *</td>
<td>0.246</td>
<td>3.376</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(2.23)</td>
<td>(1.93)</td>
<td>(2.35)</td>
<td>(0.43)</td>
<td>(1.56)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>area_sandy_clay_loam</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
</tr>
<tr>
<td>area_clay</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
<td>(...)</td>
</tr>
<tr>
<td>area_sandy_loam</td>
<td>-15.93</td>
<td>-0.682</td>
<td>0</td>
<td>0</td>
<td>0.772</td>
<td>15.59 *</td>
<td>11.36 *</td>
<td>15.42 **</td>
<td>1.579</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.72)</td>
<td>(-0.13)</td>
<td>(...)</td>
<td>(...)</td>
<td>(0.26)</td>
<td>(2.48)</td>
<td>(2.28)</td>
<td>(3.14)</td>
<td>(0.79)</td>
<td>(...)</td>
</tr>
<tr>
<td>area_prob_soil</td>
<td>-0.0131</td>
<td>-0.957</td>
<td>-17.03</td>
<td>4.893 *</td>
<td>0</td>
<td>0</td>
<td>5.466</td>
<td>0.291</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.00)</td>
<td>(-0.65)</td>
<td>(-1.29)</td>
<td>(2.08)</td>
<td>(...)</td>
<td>(...)</td>
<td>(1.91)</td>
<td>(0.37)</td>
<td>(...)</td>
<td>(...)</td>
</tr>
<tr>
<td>area_deep_black</td>
<td>0.802</td>
<td>-0.111</td>
<td>0</td>
<td>0</td>
<td>0.974</td>
<td>0.816</td>
<td>4.553</td>
<td>0.201</td>
<td>-0.0320</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(-0.10)</td>
<td>(...)</td>
<td>(...)</td>
<td>(0.76)</td>
<td>(0.72)</td>
<td>(2.00)</td>
<td>(0.35)</td>
<td>(0.07)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>area_medium_black</td>
<td>-0.602</td>
<td>-0.00331</td>
<td>0.453</td>
<td>2.365</td>
<td>-0.260</td>
<td>0.172</td>
<td>4.303</td>
<td>0.232</td>
<td>-0.0342</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>(-0.03)</td>
<td>(-0.00)</td>
<td>(0.09)</td>
<td>(1.18)</td>
<td>(-0.21)</td>
<td>(0.16)</td>
<td>(1.96)</td>
<td>(0.43)</td>
<td>(-0.08)</td>
<td>(0.29)</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>267</td>
<td>267</td>
<td>132</td>
<td>132</td>
<td>343</td>
<td>343</td>
<td>167</td>
<td>167</td>
<td>209</td>
<td>209</td>
</tr>
</tbody>
</table>

*t statistics in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
Table 11.8: Quantile regression - Kharif Cotton

<table>
<thead>
<tr>
<th>Output rate</th>
<th>0.10</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_h1</td>
<td>1.103</td>
<td>5.009</td>
<td>5.019</td>
<td>5.010</td>
<td>-0.0298</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
<td></td>
<td></td>
<td></td>
<td>(-0.20)</td>
</tr>
<tr>
<td>area</td>
<td>0.539</td>
<td>0.382</td>
<td>0.398</td>
<td>0.449</td>
<td>-0.520</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-0.71)</td>
</tr>
<tr>
<td>irri_percent_kharif_cotton</td>
<td>-0.00460</td>
<td>-0.0119</td>
<td>-0.000523</td>
<td>0.0142</td>
<td>0.0185</td>
</tr>
<tr>
<td></td>
<td>(-0.02)</td>
<td>(-0.16)</td>
<td>(-0.01)</td>
<td>(0.25)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>6.635</td>
<td>1.764</td>
<td>0.636</td>
<td>0.847</td>
<td>2.143</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.45)</td>
</tr>
<tr>
<td>area_no_degradability</td>
<td>-0.142</td>
<td>-0.347</td>
<td>-0.318</td>
<td>-0.762</td>
<td>-0.810</td>
</tr>
<tr>
<td></td>
<td>(-0.23)</td>
<td></td>
<td></td>
<td></td>
<td>(-0.39)</td>
</tr>
<tr>
<td>area_soil_erosion</td>
<td>0.0456</td>
<td>-0.284</td>
<td>-0.296</td>
<td>-0.760</td>
<td>-1.232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-0.62)</td>
</tr>
<tr>
<td>irri_percent_kharif_cotton</td>
<td>-0.00460</td>
<td>-0.0119</td>
<td>-0.000523</td>
<td>0.0142</td>
<td>0.0185</td>
</tr>
<tr>
<td></td>
<td>(-0.02)</td>
<td>(-0.16)</td>
<td>(-0.01)</td>
<td>(0.25)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>biry_owner</td>
<td>6.635</td>
<td>1.764</td>
<td>0.636</td>
<td>0.847</td>
<td>2.143</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.45)</td>
</tr>
<tr>
<td>area_no_degradability</td>
<td>-0.142</td>
<td>-0.347</td>
<td>-0.318</td>
<td>-0.762</td>
<td>-0.810</td>
</tr>
<tr>
<td></td>
<td>(-0.23)</td>
<td></td>
<td></td>
<td></td>
<td>(-0.39)</td>
</tr>
<tr>
<td>area_soil_erosion</td>
<td>0.0456</td>
<td>-0.284</td>
<td>-0.296</td>
<td>-0.760</td>
<td>-1.232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-0.62)</td>
</tr>
<tr>
<td>area_water_logging</td>
<td>0.147</td>
<td>0.0845</td>
<td>-0.159</td>
<td>-0.596</td>
<td>-1.465</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-0.13)</td>
</tr>
<tr>
<td>area_sandy_loam</td>
<td>-30.39</td>
<td>-35.02</td>
<td>0.109</td>
<td>0.325</td>
<td>5.121</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>area_deep_black</td>
<td>0.0739</td>
<td>0.368</td>
<td>0.342</td>
<td>0.106</td>
<td>1.290</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.11)</td>
</tr>
<tr>
<td>area_medium_black</td>
<td>0.0303</td>
<td>0.377</td>
<td>0.303</td>
<td>0.891</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.40)</td>
</tr>
<tr>
<td>area_shallow_black_murrum</td>
<td>0.000645</td>
<td>0.596</td>
<td>0.307</td>
<td>0.563</td>
<td>0.450</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.18)</td>
</tr>
<tr>
<td>area_sandy</td>
<td>-0.163</td>
<td>0.233</td>
<td>0.203</td>
<td>1.017</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.38)</td>
</tr>
<tr>
<td>area_red_soil</td>
<td>0.0444</td>
<td>0.335</td>
<td>0.284</td>
<td>0.831</td>
<td>1.069</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.84)</td>
</tr>
<tr>
<td>nb_plots</td>
<td>-0.192</td>
<td>-0.504</td>
<td>-0.236</td>
<td>0.700</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nbr_crops</td>
<td>0.264</td>
<td>0.442</td>
<td>-0.0384</td>
<td>-0.100</td>
<td>-0.196</td>
</tr>
<tr>
<td></td>
<td>(0.44)</td>
<td></td>
<td></td>
<td></td>
<td>(-0.15)</td>
</tr>
<tr>
<td>qty_fertilizer</td>
<td>0.00365</td>
<td>-0.000685</td>
<td>-0.0000862</td>
<td>-0.000275</td>
<td>-0.000182</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(-0.06)</td>
<td>(-0.05)</td>
<td>(-0.11)</td>
<td>(-0.04)</td>
</tr>
<tr>
<td>qty_pesticide</td>
<td>-0.00617</td>
<td>0.00238</td>
<td>-0.00580</td>
<td>0.0526</td>
<td>0.00651</td>
</tr>
<tr>
<td></td>
<td>(-0.02)</td>
<td>(0.00)</td>
<td></td>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td>area_seed_2</td>
<td>-0.150</td>
<td>2.920</td>
<td>0.00634</td>
<td>0.547</td>
<td>1.173</td>
</tr>
<tr>
<td></td>
<td>(1.27)</td>
<td></td>
<td></td>
<td></td>
<td>(1.46)</td>
</tr>
<tr>
<td>area_seed_3</td>
<td>0.185</td>
<td>-0.0164</td>
<td>0.0402</td>
<td>0.0581</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.20)</td>
</tr>
<tr>
<td>area_seed_4</td>
<td>-0.379</td>
<td>-0.159</td>
<td>-0.0222</td>
<td>0.118</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td>(-0.08)</td>
<td>(-0.05)</td>
<td>(-0.05)</td>
<td>(0.17)</td>
<td>(0.43)</td>
</tr>
<tr>
<td>llog_income</td>
<td>-0.958</td>
<td>1.736</td>
<td>0.0217</td>
<td>2.847</td>
<td>0.894</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.29)</td>
</tr>
<tr>
<td>llog_borrowings_tot</td>
<td>-0.0365</td>
<td>0.0820</td>
<td>-0.110</td>
<td>-0.155</td>
<td>-0.0244</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-0.01)</td>
</tr>
<tr>
<td>llog_gov_benefit</td>
<td>-0.141</td>
<td>0.074</td>
<td>0.224</td>
<td>0.177</td>
<td>0.499</td>
</tr>
<tr>
<td></td>
<td>(-0.07)</td>
<td>(0.01)</td>
<td>(0.24)</td>
<td></td>
<td>(1.10)</td>
</tr>
<tr>
<td>rain</td>
<td>0.00475</td>
<td>0.00177</td>
<td>0.000693</td>
<td>0.00157</td>
<td>0.00589</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.32)</td>
</tr>
</tbody>
</table>

N 417 417 417 417 417

t statistics in parentheses  **p < 0.01, ***p < 0.001
<table>
<thead>
<tr>
<th></th>
<th>kharif_cotton</th>
<th>kharif_paddy</th>
<th>kharif_pigeonpea</th>
<th>rabi_sorghum</th>
<th>rabi_wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender_h1</td>
<td>0.0334</td>
<td>0.150</td>
<td>-3.674</td>
<td>0.989</td>
<td>-0.170</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.16)</td>
<td>(-1.04)</td>
<td>(0.62)</td>
<td>(-0.77)</td>
</tr>
<tr>
<td>years_of_education_h1</td>
<td>-0.000198</td>
<td>0.0104</td>
<td>-0.0540</td>
<td>0.0674</td>
<td>0.0169</td>
</tr>
<tr>
<td></td>
<td>(-0.00)</td>
<td>(0.29)</td>
<td>(-0.29)</td>
<td>(0.50)</td>
<td>(0.88)</td>
</tr>
<tr>
<td>main_occp_h1</td>
<td>-0.0825</td>
<td>-0.0285</td>
<td>-0.515</td>
<td>0.0387</td>
<td>0.00339</td>
</tr>
<tr>
<td></td>
<td>(-0.53)</td>
<td>(-0.35)</td>
<td>(-1.24)</td>
<td>(0.25)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>subs_occp_h1</td>
<td>0.107</td>
<td>-0.124</td>
<td>0.0946</td>
<td>0.00664</td>
<td>-0.0439</td>
</tr>
<tr>
<td></td>
<td>(0.66)</td>
<td>(-1.30)</td>
<td>(0.32)</td>
<td>(0.02)</td>
<td>(-1.00)</td>
</tr>
<tr>
<td>family_size</td>
<td>0.0683</td>
<td>0.0577</td>
<td>-0.247</td>
<td>0.0961</td>
<td>-0.0342</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.81)</td>
<td>(-0.69)</td>
<td>(0.54)</td>
<td>(1.05)</td>
</tr>
<tr>
<td>market_dist</td>
<td>-0.0918</td>
<td>-0.0269</td>
<td>-0.0183</td>
<td>0.0826</td>
<td>-0.195***</td>
</tr>
<tr>
<td></td>
<td>(-0.97)</td>
<td>(-0.31)</td>
<td>(-0.25)</td>
<td>(0.49)</td>
<td>(-3.50)</td>
</tr>
<tr>
<td>land_group</td>
<td>-0.458</td>
<td>0.0915</td>
<td>-0.705</td>
<td>-1.042</td>
<td>-0.0646</td>
</tr>
<tr>
<td></td>
<td>(-0.95)</td>
<td>(0.65)</td>
<td>(-0.76)</td>
<td>(-1.02)</td>
<td>(-1.04)</td>
</tr>
<tr>
<td>caste</td>
<td>0.00725</td>
<td>-0.00231</td>
<td>0.00658</td>
<td>-0.00368</td>
<td>-0.00119</td>
</tr>
<tr>
<td></td>
<td>(0.64)</td>
<td>(-0.97)</td>
<td>(0.37)</td>
<td>(-0.25)</td>
<td>(-0.89)</td>
</tr>
<tr>
<td>religion</td>
<td>-0.117</td>
<td>0.163</td>
<td>-0.402</td>
<td>0.153</td>
<td>0.0555</td>
</tr>
<tr>
<td></td>
<td>(-0.53)</td>
<td>(0.94)</td>
<td>(-0.88)</td>
<td>(0.46)</td>
<td>(1.34)</td>
</tr>
<tr>
<td>age_h1</td>
<td>-0.543</td>
<td>0.595</td>
<td>-0.00317</td>
<td>-0.962</td>
<td>-0.0219</td>
</tr>
<tr>
<td></td>
<td>(-0.60)</td>
<td>(1.82)</td>
<td>(-0.02)</td>
<td>(-0.54)</td>
<td>(-1.47)</td>
</tr>
<tr>
<td>age_avg</td>
<td>0.595</td>
<td>-0.615</td>
<td>0.00106</td>
<td>0.931</td>
<td>0.0227</td>
</tr>
<tr>
<td></td>
<td>(0.64)</td>
<td>(-1.85)</td>
<td>(0.01)</td>
<td>(0.52)</td>
<td>(1.22)</td>
</tr>
<tr>
<td>binary_owner</td>
<td>3.247</td>
<td>-3.101***</td>
<td>-9.101</td>
<td>2.937</td>
<td>-1.064***</td>
</tr>
<tr>
<td></td>
<td>(1.31)</td>
<td>(-4.71)</td>
<td>(-1.89)</td>
<td>(0.98)</td>
<td>(-3.03)</td>
</tr>
<tr>
<td>binary_owner_avg</td>
<td>1.363</td>
<td>1.536</td>
<td>6.217</td>
<td>0</td>
<td>1.301**</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(1.40)</td>
<td>(0.95)</td>
<td>(.)</td>
<td>(2.82)</td>
</tr>
<tr>
<td>nbr_crops</td>
<td>0.279</td>
<td>0.415</td>
<td>-0.462</td>
<td>-0.0689</td>
<td>0.0576</td>
</tr>
<tr>
<td></td>
<td>(1.11)</td>
<td>(0.94)</td>
<td>(-0.89)</td>
<td>(-0.16)</td>
<td>(0.47)</td>
</tr>
<tr>
<td>nbcrops_avg</td>
<td>-0.0904</td>
<td>-0.774</td>
<td>0.535</td>
<td>0.509</td>
<td>-0.161</td>
</tr>
<tr>
<td></td>
<td>(-0.21)</td>
<td>(-1.31)</td>
<td>(0.68)</td>
<td>(0.81)</td>
<td>(-1.31)</td>
</tr>
<tr>
<td>lag_log_income</td>
<td>0.180</td>
<td>0.282</td>
<td>-1.683</td>
<td>1.657*</td>
<td>0.0590</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.63)</td>
<td>(-0.82)</td>
<td>(2.03)</td>
<td>(0.38)</td>
</tr>
<tr>
<td>lag_income_avg</td>
<td>-1.497</td>
<td>-0.148</td>
<td>2.201</td>
<td>-1.653</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>(-1.50)</td>
<td>(-0.29)</td>
<td>(0.89)</td>
<td>(-1.68)</td>
<td>(0.62)</td>
</tr>
<tr>
<td>lag_log_borrowings</td>
<td>0.406</td>
<td>-0.0401</td>
<td>1.114</td>
<td>-0.571</td>
<td>0.00229</td>
</tr>
<tr>
<td></td>
<td>(1.05)</td>
<td>(-0.19)</td>
<td>(1.48)</td>
<td>(-0.94)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>lag_borrowings_avg</td>
<td>-0.219</td>
<td>-0.0323</td>
<td>-0.622</td>
<td>-0.0245</td>
<td>0.0613</td>
</tr>
<tr>
<td></td>
<td>(-0.54)</td>
<td>(-0.11)</td>
<td>(0.30)</td>
<td>(-0.04)</td>
<td>(0.54)</td>
</tr>
<tr>
<td>lag_log_gov_benefit</td>
<td>-0.422</td>
<td>-0.11</td>
<td>-0.617</td>
<td>-1.029</td>
<td>-0.0105</td>
</tr>
<tr>
<td></td>
<td>(-0.81)</td>
<td>(-0.51)</td>
<td>(-0.68)</td>
<td>(-1.54)</td>
<td>(-0.12)</td>
</tr>
<tr>
<td>lag_gov_benefit_avg</td>
<td>0.392</td>
<td>-0.453</td>
<td>-0.802</td>
<td>1.453</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>(0.93)</td>
<td>(-1.39)</td>
<td>(-0.69)</td>
<td>(1.57)</td>
<td>(0.92)</td>
</tr>
<tr>
<td>rain</td>
<td>0.442</td>
<td>0.0324</td>
<td>0.526**</td>
<td>-0.298</td>
<td>-0.355***</td>
</tr>
<tr>
<td></td>
<td>(1.08)</td>
<td>(-0.07)</td>
<td>(3.22)</td>
<td>(-1.62)</td>
<td>(-4.70)</td>
</tr>
<tr>
<td>rain_avg</td>
<td>-0.00267</td>
<td>0.00226</td>
<td>-0.00991</td>
<td>-0.0139</td>
<td>0.000373</td>
</tr>
<tr>
<td></td>
<td>(-0.91)</td>
<td>(1.64)</td>
<td>(-0.95)</td>
<td>(-1.05)</td>
<td>(0.54)</td>
</tr>
<tr>
<td>area</td>
<td>-0.0675</td>
<td>0.320</td>
<td>-0.272</td>
<td>-0.500</td>
<td>-0.0882</td>
</tr>
<tr>
<td></td>
<td>(-0.24)</td>
<td>(0.60)</td>
<td>(-0.66)</td>
<td>(-0.93)</td>
<td>(-1.34)</td>
</tr>
<tr>
<td>area_avg</td>
<td>-0.676</td>
<td>-0.621</td>
<td>-0.00872</td>
<td>0.555</td>
<td>0.260</td>
</tr>
<tr>
<td></td>
<td>(-0.84)</td>
<td>(-0.64)</td>
<td>(-0.01)</td>
<td>(0.64)</td>
<td>(1.75)</td>
</tr>
</tbody>
</table>

* t statistics in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
### Table 11.10: Regression - CRE

#### Output rate

<table>
<thead>
<tr>
<th></th>
<th>kharif_cotton</th>
<th>kharif_paddy</th>
<th>kharif_pigeonpea</th>
<th>rabi_sorghum</th>
<th>rabi_wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>irri_percent</td>
<td>0.00343</td>
<td>0.00125</td>
<td>-0.0043</td>
<td>-0.0224</td>
<td>0.00528</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.23)</td>
<td>(0.56)</td>
<td>(-0.95)</td>
<td>(1.03)</td>
</tr>
<tr>
<td>irri_percent_avg</td>
<td>0.0213</td>
<td>0.00399</td>
<td>-0.00332</td>
<td>0.00886</td>
<td>-0.00730</td>
</tr>
<tr>
<td></td>
<td>(0.75)</td>
<td>(0.61)</td>
<td>(-0.07)</td>
<td>(0.30)</td>
<td>(-1.05)</td>
</tr>
<tr>
<td>area_soil_no_degradability</td>
<td>0.552</td>
<td>-0.00656</td>
<td>1.658*</td>
<td>-0.423</td>
<td>2.340**</td>
</tr>
<tr>
<td></td>
<td>(0.95)</td>
<td>(-0.01)</td>
<td>(2.36)</td>
<td>(-0.54)</td>
<td>(2.62)</td>
</tr>
<tr>
<td>area_no_degradability_avg</td>
<td>-0.862</td>
<td>-0.0195</td>
<td>-2.301</td>
<td>0.555</td>
<td>-2.333*</td>
</tr>
<tr>
<td></td>
<td>(-1.03)</td>
<td>(-0.02)</td>
<td>(-1.22)</td>
<td>(0.44)</td>
<td>(-2.30)</td>
</tr>
<tr>
<td>area_soil_erosion</td>
<td>1.047</td>
<td>-0.882</td>
<td>-0.0974</td>
<td>0.283</td>
<td>2.356**</td>
</tr>
<tr>
<td></td>
<td>(1.68)</td>
<td>(-1.04)</td>
<td>(-0.14)</td>
<td>(0.36)</td>
<td>(2.65)</td>
</tr>
<tr>
<td>area_erosion_avg</td>
<td>-1.138</td>
<td>2.129</td>
<td>-0.583</td>
<td>-0.0818</td>
<td>-2.396*</td>
</tr>
<tr>
<td></td>
<td>(-1.27)</td>
<td>(1.58)</td>
<td>(-0.29)</td>
<td>(-0.07)</td>
<td>(-2.32)</td>
</tr>
<tr>
<td>area_soil_nutrient_depletion</td>
<td>-0.124</td>
<td>0</td>
<td>-7.995***</td>
<td>0</td>
<td>1.295</td>
</tr>
<tr>
<td></td>
<td>(-0.07)</td>
<td>()</td>
<td>(-7.37)</td>
<td>()</td>
<td>(1.38)</td>
</tr>
<tr>
<td>area_nutrient_depletion_avg</td>
<td>0</td>
<td>0</td>
<td>8.221*</td>
<td>0</td>
<td>-2.549</td>
</tr>
<tr>
<td></td>
<td>()</td>
<td>()</td>
<td>(2.40)</td>
<td>()</td>
<td>(-1.79)</td>
</tr>
<tr>
<td>area_soil_salinity_acidity</td>
<td>0.233</td>
<td>0.363</td>
<td>0</td>
<td>3.391*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.14)</td>
<td>(0.46)</td>
<td>()</td>
<td>(1.97)</td>
<td></td>
</tr>
<tr>
<td>area_salinity_acidity_avg</td>
<td>0.636</td>
<td>-0.439</td>
<td>0</td>
<td>-3.291</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.99)</td>
<td>(-0.38)</td>
<td>()</td>
<td>(-1.74)</td>
<td></td>
</tr>
<tr>
<td>area_water_logging</td>
<td>-0.678</td>
<td>0</td>
<td>6.955</td>
<td>3.292**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-0.41)</td>
<td>()</td>
<td>(1.49)</td>
<td>()</td>
<td>(2.77)</td>
</tr>
<tr>
<td>area_water_logging_avg</td>
<td>0</td>
<td>0</td>
<td>-5.051</td>
<td>-3.128*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>()</td>
<td>()</td>
<td>(-0.98)</td>
<td>()</td>
<td>(-2.46)</td>
</tr>
<tr>
<td>area_soil_loam</td>
<td>0</td>
<td>-0.339</td>
<td>-7.656***</td>
<td>-0.729</td>
<td>-2.204</td>
</tr>
<tr>
<td></td>
<td>()</td>
<td>(0.19)</td>
<td>(-5.14)</td>
<td>(-0.91)</td>
<td>(-1.76)</td>
</tr>
<tr>
<td>area_loam_avg</td>
<td>0</td>
<td>0</td>
<td>10.72***</td>
<td>-0.8656</td>
<td>3.410*</td>
</tr>
<tr>
<td></td>
<td>()</td>
<td>()</td>
<td>(3.70)</td>
<td>(-0.07)</td>
<td>(2.14)</td>
</tr>
<tr>
<td>area_soil_sandy_loam</td>
<td>-1.245</td>
<td>-1.400</td>
<td>-4.304</td>
<td>-2.271</td>
<td>-0.0948</td>
</tr>
<tr>
<td></td>
<td>(-0.65)</td>
<td>(-1.41)</td>
<td>(-0.78)</td>
<td>(-0.61)</td>
<td>(-0.34)</td>
</tr>
<tr>
<td>area_sandy_loam_avg</td>
<td>-20.65*</td>
<td>0</td>
<td>2.177</td>
<td>2.296</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-2.13)</td>
<td>()</td>
<td>(0.27)</td>
<td>(0.28)</td>
<td>()</td>
</tr>
<tr>
<td>area_problematic_soil</td>
<td>0.121</td>
<td>0.599</td>
<td>0</td>
<td>-1.075</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.70)</td>
<td>()</td>
<td>(-0.69)</td>
<td>()</td>
</tr>
<tr>
<td>area_problematic_soil_avg</td>
<td>-1.610</td>
<td>-2.213</td>
<td>0</td>
<td>-7.684**</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-1.01)</td>
<td>(-1.91)</td>
<td>()</td>
<td>(-2.74)</td>
<td>()</td>
</tr>
<tr>
<td>area_soil_deep_black</td>
<td>-0.800</td>
<td>0.241</td>
<td>-2.281*</td>
<td>-0.365</td>
<td>-2.394**</td>
</tr>
<tr>
<td></td>
<td>(-1.24)</td>
<td>(0.26)</td>
<td>(-1.98)</td>
<td>(-0.43)</td>
<td>(-2.78)</td>
</tr>
<tr>
<td>area_deep_black_avg</td>
<td>1.010</td>
<td>0.830</td>
<td>3.291</td>
<td>0.271</td>
<td>2.357*</td>
</tr>
<tr>
<td></td>
<td>(0.82)</td>
<td>(0.54)</td>
<td>(1.52)</td>
<td>(0.22)</td>
<td>(2.37)</td>
</tr>
<tr>
<td>area_soil_medium_black</td>
<td>-0.521</td>
<td>-0.137</td>
<td>-0.861</td>
<td>0.420</td>
<td>-2.276**</td>
</tr>
<tr>
<td></td>
<td>(-0.89)</td>
<td>(-0.18)</td>
<td>(-1.13)</td>
<td>(0.36)</td>
<td>(-2.58)</td>
</tr>
<tr>
<td>area_medium_black_avg</td>
<td>0.874</td>
<td>-0.347</td>
<td>2.207</td>
<td>-2.985</td>
<td>2.134*</td>
</tr>
<tr>
<td></td>
<td>(0.98)</td>
<td>(-0.35)</td>
<td>(1.04)</td>
<td>(-1.88)</td>
<td>(2.19)</td>
</tr>
<tr>
<td>area_soil_shallow_black_murrum</td>
<td>-0.0169</td>
<td>1.296</td>
<td>-0.548</td>
<td>-1.428</td>
<td>-2.428**</td>
</tr>
<tr>
<td></td>
<td>(-0.03)</td>
<td>(0.38)</td>
<td>(-0.61)</td>
<td>(-1.79)</td>
<td>(-2.60)</td>
</tr>
<tr>
<td>area_shallow_black_murrum_avg</td>
<td>-0.377</td>
<td>1.769</td>
<td>-0.350</td>
<td>1.623</td>
<td>2.256*</td>
</tr>
<tr>
<td></td>
<td>(-0.37)</td>
<td>(0.35)</td>
<td>(-0.16)</td>
<td>(1.30)</td>
<td>(2.20)</td>
</tr>
<tr>
<td>area_soil_sandy</td>
<td>-0.897</td>
<td>-1.216</td>
<td>-4.646</td>
<td>-0.0212</td>
<td>-2.206*</td>
</tr>
<tr>
<td></td>
<td>(-0.98)</td>
<td>(-1.06)</td>
<td>(-1.14)</td>
<td>(-0.03)</td>
<td>(-2.49)</td>
</tr>
<tr>
<td>area_sandy_avg</td>
<td>0.743</td>
<td>1.471</td>
<td>5.510</td>
<td>-0.903</td>
<td>1.980</td>
</tr>
<tr>
<td></td>
<td>(0.69)</td>
<td>(1.02)</td>
<td>(1.16)</td>
<td>(-0.69)</td>
<td>(1.95)</td>
</tr>
</tbody>
</table>

N 444 230 504 214 355

* t statistics in parentheses

* * p < 0.05, ** p < 0.01, *** p < 0.001
Table 11.11: Regression - CRE

Output rate

<table>
<thead>
<tr>
<th></th>
<th>kharif_cotton</th>
<th>kharif_paddy</th>
<th>kharif_pigeonpea</th>
<th>rabi_sorghum</th>
<th>rabi_wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>area_red_soil</td>
<td>-0.131</td>
<td>-0.407</td>
<td>0.702</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.26)</td>
<td>(-0.49)</td>
<td>(0.84)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>area_red_soil_avg</td>
<td>0.626</td>
<td>0.684</td>
<td>-0.467</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.72)</td>
<td>(0.57)</td>
<td>(-0.22)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>nb_plots</td>
<td>0.0649</td>
<td>-0.410</td>
<td>-0.476</td>
<td>2.138*</td>
<td>-0.177</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(-0.90)</td>
<td>(-0.39)</td>
<td>(2.07)</td>
<td>(-1.51)</td>
</tr>
<tr>
<td>nb_plots_avg</td>
<td>0.963</td>
<td>0.334</td>
<td>1.504</td>
<td>-2.070</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>(1.01)</td>
<td>(0.78)</td>
<td>(0.72)</td>
<td>(-1.35)</td>
<td>(0.60)</td>
</tr>
<tr>
<td>qty_fertilizer</td>
<td>-0.000686</td>
<td>0.00115</td>
<td>-0.000414</td>
<td>0.0126</td>
<td>0.000420</td>
</tr>
<tr>
<td></td>
<td>(-0.60)</td>
<td>(0.93)</td>
<td>(-0.06)</td>
<td>(1.58)</td>
<td>(0.73)</td>
</tr>
<tr>
<td>qty_fertilizer_avg</td>
<td>0.00624*</td>
<td>0.00224</td>
<td>0.00392</td>
<td>-0.00491</td>
<td>-0.00141</td>
</tr>
<tr>
<td></td>
<td>(2.37)</td>
<td>(1.05)</td>
<td>(0.27)</td>
<td>(-0.26)</td>
<td>(-1.54)</td>
</tr>
<tr>
<td>qty_pesticide</td>
<td>-0.0343</td>
<td>0.0124</td>
<td>-0.144</td>
<td>0.0242</td>
<td>-0.206</td>
</tr>
<tr>
<td></td>
<td>(-0.84)</td>
<td>(0.82)</td>
<td>(-0.29)</td>
<td>(0.28)</td>
<td>(-1.37)</td>
</tr>
<tr>
<td>qty_pesticide_avg</td>
<td>0.0982</td>
<td>-0.000424</td>
<td>1.027</td>
<td>0.0662</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td>(0.92)</td>
<td>(-0.03)</td>
<td>(1.58)</td>
<td>(0.38)</td>
<td>(1.33)</td>
</tr>
<tr>
<td>area_seed_2</td>
<td>-0.502</td>
<td>0.148</td>
<td>-0.524</td>
<td>0.264</td>
<td>0.0271</td>
</tr>
<tr>
<td></td>
<td>(-1.03)</td>
<td>(0.94)</td>
<td>(-1.42)</td>
<td>(0.61)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>area_seed2_avg</td>
<td>1.461</td>
<td>-0.292</td>
<td>0.113</td>
<td>-0.130</td>
<td>-0.0497</td>
</tr>
<tr>
<td></td>
<td>(1.55)</td>
<td>(-1.09)</td>
<td>(0.16)</td>
<td>(-0.19)</td>
<td>(-0.43)</td>
</tr>
<tr>
<td>area_seed_3</td>
<td>-0.292</td>
<td>-0.195</td>
<td>-1.963</td>
<td>3.835</td>
<td>-0.311</td>
</tr>
<tr>
<td></td>
<td>(-0.74)</td>
<td>(-0.29)</td>
<td>(-1.52)</td>
<td>(0.84)</td>
<td>(-0.89)</td>
</tr>
<tr>
<td>area_seed3_avg</td>
<td>0.864</td>
<td>-0.0396</td>
<td>1.608</td>
<td>-17.22**</td>
<td>0.654</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td>(-0.02)</td>
<td>(0.84)</td>
<td>(-3.06)</td>
<td>(1.00)</td>
</tr>
<tr>
<td>area_seed_4</td>
<td>0.00383</td>
<td>0</td>
<td>-2.484</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(.)</td>
<td>(-1.44)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>area_seed4_avg</td>
<td>0.0388</td>
<td>0</td>
<td>3.555</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(.)</td>
<td>(0.71)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>2011.year</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
<td>(.)</td>
</tr>
<tr>
<td>2012.year</td>
<td>-97.65</td>
<td>4.433</td>
<td>-102.5**</td>
<td>-44.97</td>
<td>60.46***</td>
</tr>
<tr>
<td></td>
<td>(-1.10)</td>
<td>(0.15)</td>
<td>(-2.88)</td>
<td>(-1.77)</td>
<td>(5.00)</td>
</tr>
<tr>
<td>2013.year</td>
<td>-300.7</td>
<td>10.46</td>
<td>-322.1**</td>
<td>42.42</td>
<td>192.9***</td>
</tr>
<tr>
<td></td>
<td>(-1.07)</td>
<td>(0.11)</td>
<td>(-2.89)</td>
<td>(1.47)</td>
<td>(4.75)</td>
</tr>
<tr>
<td></td>
<td>(-1.24)</td>
<td>(0.72)</td>
<td>(-1.43)</td>
<td>(1.52)</td>
<td>(5.06)</td>
</tr>
</tbody>
</table>

N  344  230  504  214  355

t statistics in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
Chapter 12

The dynamics of farm productivity

12.1 Introduction

Rural populations in developing countries have long been characterized by economic stasis thus raising concerns about poverty traps. Agronomists have looked at farm productivity in terms of yield and identified substantial yield gaps. Few has been done is terms of economic productivity. Rather than looking at the gap with crop model predictions, this study focuses on the dynamics of households’ productivity in terms of output rate and total factor productivity. Using crop as economic sector and using panel data for Semi-Arid Tropics (SAT) rural India, results show that the low productive households experience the fastest growing productivity, hence providing legitimate hopes for the poorest.

12.2 Literature review

A poverty trap is characterized by the existence of critical thresholds in the assets distribution that must be exceeded in order to converge towards an upper standard of living stable equilibrium. The inability or impossibility to overcome such thresholds condemns people to remain in their initial poor livelihood and therefore constitutes a self-reinforcing mechanism of chronic poverty [Barrett and Swallow, 2006, Barbier, 2010].

Many research has investigated the concept of poverty trap at the macro level to understand why some countries develop while some others persistently struggle with
poverty [Banerjee and Duflo, 2003]. Less has been done about poverty traps at the micro level. In particular, the reasons why some households succeed in rising out of poverty while others in a similar context fail are not well understood. Research about households’ welfare dynamics and poverty traps at the household level have mostly focused on Sub-Saharan Africa [Barrett and Swallow, 2006, Michelle et al., 2006]. As for India, welfare dynamics has been investigated for a long time [Townsend, 1994, Walker and Ryan, 1990]. However, the hypothesis of household-level poverty trap has attracted attention only recently [Naschold, 2012, Naschold, 2013]. In the case of India, these studies have shown that rural populations are characterized by economic stasis.

Concurrently, another field of research has investigated rural populations’ livelihoods in developing countries via the analysis of farm productivity. One measure of agricultural productivity is the output rate or farm yield. Looking at the difference between the actual farm yields experienced by farmers and potential yields, many case studies have found the existence of large yield gaps. The yield of reference differs within studies. Some consider the potential yield predicted by crop models, either in irrigated or rain fed conditions. In that case, the yield is only limited by biophysical factors such as temperature, water and soil characteristics. Other studies take into account field experiment yields as the baseline. Finally, yield gaps is also measured with respect to the maximum yield attained at the local level. Although these alternative definitions may provide researchers with different measures of yield gaps, differences are usually small [Lobell et al., 2009] and a spread of yield gaps ranging from 20 to 80% is commonly accepted depending on the definition adopted and the location investigated. As an example, considering the yield given by field experiment as the reference, rice yield gaps was estimated to be equal to 52.3% in average for whole India for the period 1990-1998 (FAO, ICAR, 2000). At the state level and for the same period, yield gaps was as low as 15.6% in Tamil Nadul and reached the highest level in Rajasthan with 75.6%.

Substantial efforts have been made to identify the drivers of yield gaps and the potential gains from bridging the gap. The explanatory factors are twofold. A first group of factors consist in biophysics constraints such as low water availability, extreme climatic events such as floods and droughts, pest attack and disease, and adverse soil characteris-
tics, in particular nutrient depletion, salinity, acidity and erosion. A second group of factors gathers socio-economic variables. In particular, lack of knowledge and information, management practices, misuse of inputs such as fertilizers and pesticides, low financial access, inputs and output market price variability have been cited as critical parameters impacting yields.

Research on productivity is not exclusive to agriculture economics and agronomics. Industrial economists have long been studying firms’ productivity. The starting point is the production function initially developed for agricultural production. Historically, a large attention has been given to the determinants of firm productivity in particular via the development of the Solow model for economic growth and productivity. Concerned by economic growth slowdown in recent decades, a huge research effort has been done by economists to study the growth of productivity. Growth in productivity, usually considered to be driven by R&D investment, human capital and knowledge, is expected to be the key solution for both developed and developing nations to pave the way towards sustainable economic welfare. Looking at the growth of technological factor productivity (TFP), recent studies have nevertheless found a dispersion in growth of firm productivity measured by the three-year Cumulative Aggregated Growth Rate (CAGR) \cite{Bahar, 2018}. Looking at different industrial sectors and based on a data set of worldwide firm’s financial information, Bahar found a striking pattern of growth in TFP characterizing a productivity dispersion. In particular, this study shows that firms respectively with the lowest and highest initial total productivity factor experience the fastest growth in TFP. In addition to the productivity dispersion, this research sheds light on the existence of a "middle productivity trap" \cite{Bahar, 2018}. This "U-shaped" relationships holds within country and industrial sectors.

Grounded in these two research areas, namely agricultural economics and industrial growth productivity, our study proposes to combine the developments of both domains in order to investigate the growth of farm productivity for rural households in Semi-Arid Tropics India that were previously characterized by economic stasis.
12.3 Methodology

Following Bahar’s research [Bahar, 2018], our study investigates the dynamics of productivity for rural households in India. By analogy with Bahar’s paper studying firms, the economic unit of analysis is constituted by households, while industrial economic sector are modeled by the different crops. Similarly to Bahar’s analysis, our data set is a panel data. Collected by the International Crop Research Institute for Semi-Arid Tropics (ICRISAT), this panel data set gathers information about households, including family composition, soil and landholding characteristics, farm production, inputs data and financial information.

Analyzed crops are respectively cotton, pigeon pea and wheat. This choice is motivated by three reasons. First, sufficient data are available for those crops. Second, pigeon pea and wheat constitute staple food for many Indian households. Third, these crops are grown only in one season per year, at least for the sample set at hand, thus preventing from confounding the analysis with seasonal variability. In addition, data about production sales do not discriminate sales in function of their period of production. Including production for all seasons would thus lead to compare the growth in productivity for households that produce crops for two seasons and households that produce the same crop only for one period. Cotton and pigeon pea are grown during Kharif season (July-October) while wheat is grown during Rabi season (October-March).

Aiming at studying the dynamics of households’ productivity, we adopt two different measures of productivity. Following agronomic literature, we first look at the growth of output rate. The output rate is defined by the ratio of the production in kilograms (kg) by the cultivated area (in acres). The growth in output rate is taken to be the difference of logarithm of output rate over two consecutive periods:

$$\Delta Q_{i,t} = Q_{i,t} - Q_{i,t-1}$$

(12.1)

where $Q_{i,t}$ is the output for the unit $i$ at time $t$ and $\Delta Q_{i,t}$ is the change in output rate

Second, we measure the productivity via the technological factor productivity (TFP).
For that purpose, we start with a traditional production function modeled by a Cobb-Douglas function:

\[ Y = F^{\beta_F} L^{\beta_L} K^{\beta_K} \]  

(12.2)

where \( F \) is the monetary value of fixed assets, \( L \) is the value of labor inputs, \( K \) is the value of material inputs.

In our study, fixed assets are the farming land value. We estimate the model with two alternative measures: the price of the land and the rental price of farming. Fewer data are available for the price of farm plots and we find less significant coefficients. For that reason, the remaining of this paper reports the results with fixed assets measured by the rental value.

The data set provides the labor inputs effectively employed by households. Labor inputs include family labor, hired labor and exchange labor. All are declined between male, female and child labor. Hired labor also includes Regular Farm Servant (RFS), a traditional employment status in rural India under which workers are tied to a specific farm for a definite period of time. Even if households are not subjected to the costs for family and exchange labor, the equivalent costs are provided by the data set. We thus estimate the production function with a total equivalent labor cost. Material inputs include fertilizers, pesticides, micronutrients, seeds and weedicide.

The dependent variable \( Y \) models the operating revenue. Data provides the total crop production in kg, the production sold and crop sales prices. In practice, only a portion of the total production is usually sold. The remaining part is dedicated to the household’s self-consumption. We thus compute a total equivalent operating revenue by summing the revenue from farm production sales and the equivalent revenue of auto-consumed production. The equivalent price of the production that is not effectively sold is computed by taking the average price of the production sold.
We use three different estimation methods for computing the elasticities of the Cobb-Douglas functions:

- Ordinary Least Squares estimation taking equation 12.2 in logarithm form as suggested in [Bahar, 2018]
- Share of each input value in the final operating revenue at the household level
- Share of each input value in the final operating revenue at the sector level

Once calculated the estimated elasticities $\beta_F$, $\beta_L$ and $\beta_K$, we compute the TFP with the following equation:

$$\log TFP = \log Y - \hat{\beta}_F \log F - \hat{\beta}_L \log L - \hat{\beta}_K \log K$$  \hspace{1cm} (12.3)

We then calculate the growth in TFP by the difference in log TFP between two consecutive periods and estimate the following models:

$$\Delta TFP = \alpha_{10} + \alpha_{11} \log TFP_{t-1} + \alpha_{12} \log TFP^2_{t-1}$$  \hspace{1cm} (12.4)

$$\Delta TFP = \alpha_{20} + \alpha_{21} \log TFP_{t-1} + \alpha_{22} \log TFP^2_{t-1} + \alpha_{23} \log TFP^3_{t-1}$$  \hspace{1cm} (12.5)

We describe below the basic statistics of our data.

**12.3.1 Summary statistics of output rate and change in output rate**

The two following tables summarize the data for the output rate and the change in output rate, for the cotton (kharif season), pigeonpea (kharif season) and wheat (rabi) respectively.
Table 12.1: Output rate statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_output_rate_Kharif_COTTON</td>
<td>783</td>
<td>3.64</td>
<td>0.15</td>
<td>2.56</td>
<td>4.20</td>
</tr>
<tr>
<td>log_output_rate_Kharif_PIGEONPEA</td>
<td>1476</td>
<td>3.62</td>
<td>0.07</td>
<td>1.43</td>
<td>4.31</td>
</tr>
<tr>
<td>log_output_rate_Rabi_WHEAT</td>
<td>1676</td>
<td>2.64</td>
<td>0.21</td>
<td>0.92</td>
<td>3.55</td>
</tr>
</tbody>
</table>

Table 12.2: Change in output rate statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>delta_output_rate_Kharif_COTTON</td>
<td>459</td>
<td>0.045</td>
<td>0.17</td>
<td>−0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>delta_output_rate_Kharif_PIGEONPEA</td>
<td>824</td>
<td>0.016</td>
<td>0.30</td>
<td>−1.124</td>
<td>1.099</td>
</tr>
<tr>
<td>delta_output_rate_Rabi_WHEAT</td>
<td>1071</td>
<td>0.04</td>
<td>0.18</td>
<td>−1.57</td>
<td>1.79</td>
</tr>
</tbody>
</table>

12.3.2 Including land rental value as fixed assets

Summary statistics for each crop

The following tables display the statistics of the data when we include the land rental value as fixed assets. We give the summary tables for each one of the estimation methods (see explanations above).
## i) OLS estimation

### Table 12.3: TFP statistics per sector - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>233</td>
<td>4.68</td>
<td>1.66</td>
<td>1.58</td>
<td>10.18</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>196</td>
<td>0.58</td>
<td>1.70</td>
<td>-2.08</td>
<td>6.37</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>227</td>
<td>12.11</td>
<td>2.62</td>
<td>7.43</td>
<td>16.57</td>
</tr>
</tbody>
</table>

### Table 12.4: Growth TFP statistics per sector - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>83</td>
<td>0.42</td>
<td>1.70</td>
<td>-5.59</td>
<td>3.82</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>30</td>
<td>-0.0018</td>
<td>1.6688</td>
<td>-4.5147</td>
<td>1.988</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>41</td>
<td>0.99</td>
<td>1.86</td>
<td>-4.46</td>
<td>4.54</td>
</tr>
</tbody>
</table>

## ii) Share of input value - household level

### Table 12.5: TFP statistics per sector - Share cost within household estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>233</td>
<td>-22.35</td>
<td>4.01</td>
<td>-33.51</td>
<td>-4.06</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>196</td>
<td>-22.41</td>
<td>6.90</td>
<td>-43.08</td>
<td>-2.75</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>227</td>
<td>-10.36</td>
<td>6.44</td>
<td>-25.33</td>
<td>-0.49</td>
</tr>
</tbody>
</table>
Table 12.6: Growth TFP statistics per sector - Share within household

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>83</td>
<td>0.094</td>
<td>2.885</td>
<td>−9.764</td>
<td>6.771</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>30</td>
<td>−0.28</td>
<td>2.14</td>
<td>−6.58</td>
<td>2.16</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>41</td>
<td>0.82</td>
<td>2.07</td>
<td>−4.93</td>
<td>4.66</td>
</tr>
</tbody>
</table>

iii) Share of input value - sector level

Table 12.7: TFP statistics per sector - Share cost within sector estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>233</td>
<td>−22.51</td>
<td>2.53</td>
<td>−28.70</td>
<td>−13.60</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>196</td>
<td>−21.85</td>
<td>2.99</td>
<td>−31.55</td>
<td>−12.56</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>227</td>
<td>−11.20</td>
<td>2.66</td>
<td>−17.42</td>
<td>−5.86</td>
</tr>
</tbody>
</table>

Table 12.8: Growth TFP statistics per sector - Share within sector

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>83</td>
<td>0.062</td>
<td>2.987</td>
<td>−10.679</td>
<td>6.983</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>30</td>
<td>−0.32</td>
<td>2.30</td>
<td>−7.34</td>
<td>2.12</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>41</td>
<td>0.88</td>
<td>2.08</td>
<td>−4.88</td>
<td>4.71</td>
</tr>
</tbody>
</table>

Summary statistics for each dependent and independent variable

We regress the dependent variable - respectively the change in output rate and the growth of TFP - all sectors together and then by sector. When we keep sectors all together, we
have to delete the households that grow at least two of the studied crops, because we use a panel data estimator with fixed effects for the household and year. Effectively, observations for a given year are not independent in this case.

i) Statistics of the output rate

Table 12.9: Output rate statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_output_rate</td>
<td>1987</td>
<td>3.13</td>
<td>0.56</td>
<td>0.92</td>
<td>4.32</td>
</tr>
<tr>
<td>log_output_rate_square</td>
<td>1987</td>
<td>9.83</td>
<td>3.49</td>
<td>0.84</td>
<td>18.64</td>
</tr>
<tr>
<td>log_output_rate_cube</td>
<td>1987</td>
<td>30.83</td>
<td>16.73</td>
<td>0.77</td>
<td>80.48</td>
</tr>
<tr>
<td>delta_output_rate</td>
<td>1150</td>
<td>0.035</td>
<td>0.22</td>
<td>-1.57</td>
<td>1.79</td>
</tr>
</tbody>
</table>

ii) Statistics of the TFP - OLS estimation

Table 12.10: TFP statistics - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>280</td>
<td>-1.00</td>
<td>2.19</td>
<td>-5.25</td>
<td>4.47</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>280</td>
<td>5.78</td>
<td>4.86</td>
<td>0.0002</td>
<td>27.59</td>
</tr>
<tr>
<td>log_output_rate_cube</td>
<td>280</td>
<td>-10</td>
<td>25.5</td>
<td>-144.9</td>
<td>89.3</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>67</td>
<td>0.071</td>
<td>1.93</td>
<td>-6.408</td>
<td>5.006</td>
</tr>
</tbody>
</table>
iii) Statistics of the TFP - Share of input value at household level

Table 12.11: TFP statistics - Share within household

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>280</td>
<td>−16.40</td>
<td>8.47</td>
<td>−42.21</td>
<td>−2.33</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>280</td>
<td>340.63</td>
<td>270.11</td>
<td>5.42</td>
<td>1781.65</td>
</tr>
<tr>
<td>log_TFP_cube</td>
<td>280</td>
<td>−7785</td>
<td>8455.7</td>
<td>−75203</td>
<td>−12.6</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>67</td>
<td>−0.045</td>
<td>2.585</td>
<td>−8.380</td>
<td>6.771</td>
</tr>
</tbody>
</table>

iv) Statistics of the TFP - Share of input value at sector level

Table 12.12: TFP statistics - Share within sector

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>280</td>
<td>−16.87</td>
<td>5.98</td>
<td>−26.94</td>
<td>−6.01</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>280</td>
<td>320.2</td>
<td>190.7</td>
<td>36.1</td>
<td>725.8</td>
</tr>
<tr>
<td>log_TFP_cube</td>
<td>280</td>
<td>−6527.17</td>
<td>4976.90</td>
<td>−19554.75</td>
<td>−216.70</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>67</td>
<td>0.51</td>
<td>3.72</td>
<td>−8.93</td>
<td>10.14</td>
</tr>
</tbody>
</table>

12.3.3 Without fixed assets

We know provide similar statistics summary when we do not include fixed assets.
Summary statistics for each crop

i) Statistics of the TFP - OLS estimation

Table 12.13: TFP statistics per sector - without fixed assets - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>596</td>
<td>7.483</td>
<td>1.494</td>
<td>4.322</td>
<td>12.669</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>493</td>
<td>1.391</td>
<td>1.811</td>
<td>−2.757</td>
<td>7.687</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>574</td>
<td>9.35</td>
<td>2.40</td>
<td>6.33</td>
<td>15.61</td>
</tr>
</tbody>
</table>

Table 12.14: Growth TFP statistics per sector - without fixed assets - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>332</td>
<td>0.406</td>
<td>1.503</td>
<td>−5.600</td>
<td>4.869</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>178</td>
<td>−0.225</td>
<td>1.368</td>
<td>−4.786</td>
<td>6.539</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>248</td>
<td>0.28</td>
<td>2.08</td>
<td>−5.79</td>
<td>5.64</td>
</tr>
</tbody>
</table>

ii) Statistics of the TFP - Share of input value at household level

Table 12.15: TFP statistics per sector - without fixed assets - Share cost within household estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>596</td>
<td>−16.35</td>
<td>3.76</td>
<td>−27.75</td>
<td>0.46</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>493</td>
<td>−15.22</td>
<td>6.72</td>
<td>−39.44</td>
<td>−3.62</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>574</td>
<td>−6.64</td>
<td>5.54</td>
<td>−17.26</td>
<td>5.56</td>
</tr>
</tbody>
</table>
Table 12.16: Growth TFP statistics per sector - without fixed assets - Share within household estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>332</td>
<td>0.12</td>
<td>2.50</td>
<td>−10.29</td>
<td>8.34</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>178</td>
<td>−0.27</td>
<td>2.13</td>
<td>−7.22</td>
<td>6.05</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>248</td>
<td>0.173</td>
<td>2.475</td>
<td>−7.186</td>
<td>6.169</td>
</tr>
</tbody>
</table>

### iii) Statistics of the TFP - Share of input value at sector level

Table 12.17: TFP statistics per sector - without fixed assets - Share cost within sector estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP_Kharif_COTTON</td>
<td>596</td>
<td>−16.298</td>
<td>2.401</td>
<td>−22.439</td>
<td>−7.806</td>
</tr>
<tr>
<td>log_TFP_Kharif_PIGEONPEA</td>
<td>493</td>
<td>−14.99</td>
<td>3.16</td>
<td>−24.32</td>
<td>−4.80</td>
</tr>
<tr>
<td>log_TFP_Rabi_WHEAT</td>
<td>574</td>
<td>−6.54</td>
<td>2.74</td>
<td>−11.89</td>
<td>1.55</td>
</tr>
</tbody>
</table>

Table 12.18: Growth TFP statistics per sector - without fixed assets - Share within sector estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_TFP_Kharif_COTTON</td>
<td>332</td>
<td>0.13</td>
<td>2.55</td>
<td>−10.64</td>
<td>9.39</td>
</tr>
<tr>
<td>growth_TFP_Kharif_PIGEONPEA</td>
<td>178</td>
<td>−0.28</td>
<td>2.23</td>
<td>−7.01</td>
<td>5.89</td>
</tr>
<tr>
<td>growth_TFP_Rabi_WHEAT</td>
<td>248</td>
<td>0.183</td>
<td>2.439</td>
<td>−7.208</td>
<td>6.121</td>
</tr>
</tbody>
</table>
Summary statistics for each dependent and independent variable

i) Statistics of the TFP - OLS estimation

Table 12.19: TFP statistics - without fixed assets - OLS estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>499</td>
<td>5.50</td>
<td>2.48</td>
<td>0.42</td>
<td>10.45</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>499</td>
<td>36.40</td>
<td>31.39</td>
<td>0.18</td>
<td>109.16</td>
</tr>
<tr>
<td>log_TFP_cube</td>
<td>499</td>
<td>277.247</td>
<td>325.142</td>
<td>0.075</td>
<td>1140.473</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>192</td>
<td>0.012</td>
<td>2.300</td>
<td>−6.956</td>
<td>6.816</td>
</tr>
</tbody>
</table>

ii) Statistics of the TFP - Share of input value at household level

Table 12.20: TFP statistics - without fixed assets - share within household

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>499</td>
<td>−10.66</td>
<td>7.26</td>
<td>−35.31</td>
<td>4.17</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>499</td>
<td>166.14</td>
<td>144.83</td>
<td>0.00069</td>
<td>1246.82</td>
</tr>
<tr>
<td>log_TFP_cube</td>
<td>499</td>
<td>−2729.6</td>
<td>3501.0</td>
<td>−44025.9</td>
<td>72.2</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>192</td>
<td>−0.025</td>
<td>2.836</td>
<td>−8.845</td>
<td>7.108</td>
</tr>
</tbody>
</table>
iii) Statistics of the TFP - Share of input value at sector level

Table 12.21: TFP statistics - without fixed assets - share within sector estimation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>499</td>
<td>-11.03</td>
<td>5.28</td>
<td>-20.54</td>
<td>-0.90</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>499</td>
<td>149.46</td>
<td>109.31</td>
<td>0.81</td>
<td>421.84</td>
</tr>
<tr>
<td>log_TFP_cube</td>
<td>499</td>
<td>-2210.33</td>
<td>2024.45</td>
<td>-8664.00</td>
<td>-0.72</td>
</tr>
<tr>
<td>growth_TFP</td>
<td>192</td>
<td>-0.12</td>
<td>3.25</td>
<td>-9.19</td>
<td>9.71</td>
</tr>
</tbody>
</table>

12.4 Results

As for the change in output rate, when including the logarithm of the output rate lagged by one period and its square equivalent, only the former is significant except when all sector are confounded. The correlation is systematically negative, with a coefficient respectively equal to -1.300, -1.900 and 1.100 for cotton, pigeon pea and wheat. These coefficients are statistically significative at level 0.001. When we gather all "sectors", the coefficient for the linear variable equals -0.490, the coefficient for the squared variable equals -0.058, respectively at significance level 0.1% and 5%.

When the model also include the cube term of the logarithm of the lagged output rate, all three covariates are significant at 0.1% for each separated sector and all sectors considered. In all cases, we find a negative correlation with the linear term and the cube term, and a positive correlation with the squared term.

12.4.1 Change in output rate

We show below the graph representation of the change in output rate for each crop. Graph 12-1a provide a graph for all crops together. Graphs 12-1b, 12-1c and 12-1d show the rela-
tion for individual crop, more specifically for cotton, pigeonpea and wheat respectively.

Figure 12-1: Change in output rate over two consecutive periods

(a) Change in output rate over two consecutive periods
(b) Change in output rate - Cotton
(c) Change in output rate - Pigeonpea
(d) Change in output rate - Wheat
We then give the results of the regression given by the model in equations 12.4 and 12.5.

**Equation model 12.4 (square term)**

Table 12.22: Regression change in output rate

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_output_rate</td>
<td>-0.490^***</td>
<td>-1.300^*</td>
<td>-1.900^***</td>
<td>-1.100^***</td>
</tr>
<tr>
<td></td>
<td>(0.170)</td>
<td>(0.740)</td>
<td>(0.670)</td>
<td>(0.260)</td>
</tr>
<tr>
<td>log_output_rate_square</td>
<td>-0.058^*</td>
<td>0.018</td>
<td>0.091</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.100)</td>
<td>(0.097)</td>
<td>(0.049)</td>
</tr>
<tr>
<td>N</td>
<td>1,146</td>
<td>459</td>
<td>824</td>
<td>1,071</td>
</tr>
<tr>
<td>R²</td>
<td>0.470</td>
<td>0.800</td>
<td>0.520</td>
<td>0.450</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.160</td>
<td>0.670</td>
<td>0.230</td>
<td>0.150</td>
</tr>
<tr>
<td>F Statistic</td>
<td>322.000***</td>
<td>561.000***</td>
<td>280.000***</td>
<td>282.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 2; 723)</td>
<td>(df = 2; 276)</td>
<td>(df = 2; 507)</td>
<td>(df = 2; 693)</td>
</tr>
</tbody>
</table>

`t` statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Equation model [12.5] (cube term)

Table 12.23: Regression change in output rate

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_output_rate, shift = 1)</td>
<td>-3.300***</td>
<td>-28.000***</td>
<td>-18.000**</td>
<td>-6.800***</td>
</tr>
<tr>
<td></td>
<td>(0.630)</td>
<td>(8.200)</td>
<td>(7.600)</td>
<td>(0.790)</td>
</tr>
<tr>
<td>Lag(log_output_rate_square, shift = 1)</td>
<td>1.000***</td>
<td>7.400***</td>
<td>5.000**</td>
<td>2.600***</td>
</tr>
<tr>
<td></td>
<td>(0.230)</td>
<td>(2.300)</td>
<td>(2.300)</td>
<td>(0.340)</td>
</tr>
<tr>
<td>Lag(log_output_rate_cube, shift = 1)</td>
<td>-0.130***</td>
<td>-0.690***</td>
<td>-0.490**</td>
<td>-0.360***</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.210)</td>
<td>(0.230)</td>
<td>(0.048)</td>
</tr>
<tr>
<td>N</td>
<td>1,146</td>
<td>459</td>
<td>824</td>
<td>1,071</td>
</tr>
<tr>
<td>R²</td>
<td>0.490</td>
<td>0.810</td>
<td>0.530</td>
<td>0.490</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.190</td>
<td>0.680</td>
<td>0.230</td>
<td>0.210</td>
</tr>
<tr>
<td>F Statistic</td>
<td>228.000***</td>
<td>390.000***</td>
<td>190.000***</td>
<td>222.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 722)</td>
<td>(df = 3; 275)</td>
<td>(df = 3; 506)</td>
<td>(df = 3; 692)</td>
</tr>
</tbody>
</table>

`t` statistics in parentheses

*p < 0.05, **p < 0.01, ***p < 0.001

12.4.2 Growth in TFP - Fixed assets included in the productivity function

We now give the result of the regression when the dependent variable is the growth of the TFP and when we include fixed assets in the Cobb-Douglas productivity function. Again, we give the results for the three estimation methods of the coefficients of the Cobb-Douglas productivity function.
i) OLS estimation

Using OLS estimation, the coefficients of the production function are given by the following table:

Table 12.24: Coefficients production function - OLS estimation

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed assets log_rental_value</td>
<td>1.100^***</td>
<td>0.920^***</td>
<td>0.560</td>
<td>-0.760</td>
</tr>
<tr>
<td></td>
<td>(0.390)</td>
<td>(0.270)</td>
<td>(0.400)</td>
<td>(0.720)</td>
</tr>
<tr>
<td>Labor inputs value</td>
<td>-0.100</td>
<td>-0.360^**</td>
<td>-0.023</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>(0.220)</td>
<td>(0.170)</td>
<td>(0.200)</td>
<td>(0.260)</td>
</tr>
<tr>
<td>Material inputs value</td>
<td>0.480</td>
<td>0.310</td>
<td>0.460</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td>(0.340)</td>
<td>(0.240)</td>
<td>(0.380)</td>
<td>(0.680)</td>
</tr>
<tr>
<td>N</td>
<td>280</td>
<td>233</td>
<td>194</td>
<td>227</td>
</tr>
<tr>
<td>R²</td>
<td>0.180</td>
<td>0.150</td>
<td>0.140</td>
<td>0.035</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-2.000</td>
<td>-0.880</td>
<td>-2.100</td>
<td>-2.900</td>
</tr>
<tr>
<td>F Statistic</td>
<td>5.700***</td>
<td>6.000***</td>
<td>2.900**</td>
<td>0.670</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 77)</td>
<td>(df = 3; 105)</td>
<td>(df = 3; 53)</td>
<td>(df = 3; 56)</td>
</tr>
</tbody>
</table>

\( t \) statistics in parentheses

\* \( p < 0.05 \), \** \( p < 0.01 \), \*** \( p < 0.001 \)
Equation model \[12.4\] (square term)

Table 12.25: Coefficients growth in TFP  
OLS estimation - Square term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>-0.800^***</td>
<td>-2.700^**</td>
<td>-0.960</td>
<td>-3.000</td>
</tr>
<tr>
<td></td>
<td>(0.220)</td>
<td>(1.100)</td>
<td>(0.670)</td>
<td>(2.600)</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>0.120^*</td>
<td>0.160^*</td>
<td>1.000</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
<td>(0.085)</td>
<td>(0.720)</td>
<td>(0.110)</td>
</tr>
<tr>
<td>N</td>
<td>65</td>
<td>82</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>R^2</td>
<td>0.440</td>
<td>0.340</td>
<td>0.700</td>
<td>0.700</td>
</tr>
<tr>
<td>Adjusted R^2</td>
<td>-1.100</td>
<td>-1.100</td>
<td>-0.950</td>
<td></td>
</tr>
<tr>
<td>F Statistic</td>
<td>6.700***</td>
<td>9.000***</td>
<td>4.700*</td>
<td>7.000**</td>
</tr>
<tr>
<td>(df = 2; 17)</td>
<td>(df = 2; 36)</td>
<td>4.700*</td>
<td>(df = 2; 6)</td>
<td></td>
</tr>
</tbody>
</table>

*t statistics in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
Equation model [12.5](cube term)

**Table 12.26: Coefficients growth in TFP**

**OLS elasticities estimation - Cube term**

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log TFP, shift = 1)</td>
<td>-1.300^***</td>
<td>1.700</td>
<td>0.068</td>
<td>4.000</td>
</tr>
<tr>
<td></td>
<td>(0.320)</td>
<td>(5.600)</td>
<td>(1.300)</td>
<td>(26.000)</td>
</tr>
<tr>
<td>Lag(log TFP_square, shift = 1)</td>
<td>0.012</td>
<td>-0.600</td>
<td>1.400</td>
<td>-0.520</td>
</tr>
<tr>
<td></td>
<td>(0.084)</td>
<td>(0.960)</td>
<td>(0.840)</td>
<td>(2.200)</td>
</tr>
<tr>
<td>Lag(log TFP_cube, shift = 1)</td>
<td>0.056^*</td>
<td>0.042</td>
<td>-0.870</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.053)</td>
<td>(0.930)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>N</td>
<td>65</td>
<td>82</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>R²</td>
<td>0.560</td>
<td>0.350</td>
<td>0.770</td>
<td>0.700</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-0.770</td>
<td>-0.510</td>
<td>-1.200</td>
<td>-1.300</td>
</tr>
<tr>
<td>F Statistic</td>
<td>6.700***</td>
<td>6.200***</td>
<td>3.300</td>
<td>4.000*</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 16)</td>
<td>(df = 3; 35)</td>
<td>(df = 3; 3)</td>
<td>(df = 3; 5)</td>
</tr>
</tbody>
</table>

*t statistics in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
We now show the plots of the growth of TFP (y-axis) as a function of the TFP at the previous period $t - 1$ (x-axis). Similarly to the figure 1, the plot in Figure 12-2a represents all crops together. Figures 12-2b, 12-2c, and 12-2d are for cotton, pigeonpea, and wheat respectively.

Figure 12-2: Growth in TFP - OLS estimation

(a) Growth in TFP

(b) Growth in TFP - Cotton

(c) Growth in TFP - Pigeonpea

(d) Growth in TFP - Wheat
ii) Estimation via the share of input value at household level

We show below the results for the regression model given in equation 12.4 (i.e. square term) and 12.5 (i.e. cube term).

**Equation model 12.4 (square term)**

Table 12.27: Coefficients growth in TFP
Share per household elasticities estimation - Square term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>-0.008</td>
<td>-0.640</td>
<td>1.200</td>
<td>-0.880</td>
</tr>
<tr>
<td></td>
<td>(0.500)</td>
<td>(1.300)</td>
<td>(3.300)</td>
<td>(2.300)</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>0.033*</td>
<td>-0.006</td>
<td>-0.016</td>
<td>-0.014</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.028)</td>
<td>(0.063)</td>
<td>(0.091)</td>
</tr>
</tbody>
</table>

|    |    |    |    |    |
|----|----|----|----|
| N  | 65 | 82 | 29 | 40 |
| R² | 0.4620 | 0.150 | 0.120 | 0.440 |
| Adjusted R² | -0.910 | -5.200 | -2.600 |
| F Statistic | 14.000*** | 3.200* | 0.280 | 2.400 |
| F Statistic | (df = 2; 17) | (df = 2; 36) | (df = 2; 4) | (df = 2; 6) |

*t statistics in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
Equation model [12.5](cube term)

Table 12.28: Coefficients growth in TFP
Share per household elasticities estimation - Cube term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>0.590</td>
<td>21.000**</td>
<td>-64.000</td>
<td>26.000*</td>
</tr>
<tr>
<td></td>
<td>(1.700)</td>
<td>(9.400)</td>
<td>(46.000)</td>
<td>(13.000)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>0.082</td>
<td>0.930**</td>
<td>-2.800</td>
<td>2.000*</td>
</tr>
<tr>
<td></td>
<td>(0.130)</td>
<td>(0.410)</td>
<td>(2.000)</td>
<td>(0.950)</td>
</tr>
<tr>
<td>Lag(log_TFP_cube, shift = 1)</td>
<td>0.001</td>
<td>0.013**</td>
<td>-0.039</td>
<td>0.048*</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.006)</td>
<td>(0.028)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>N</td>
<td>65</td>
<td>82</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>R^2</td>
<td>0.620</td>
<td>0.260</td>
<td>0.470</td>
<td>0.700</td>
</tr>
<tr>
<td>Adjusted R^2</td>
<td>-0.500</td>
<td>-0.700</td>
<td>-4.000</td>
<td>-1.300</td>
</tr>
<tr>
<td>F Statistic</td>
<td>8.900***</td>
<td>4.200**</td>
<td>0.880</td>
<td>4.000*</td>
</tr>
<tr>
<td>F Statistic</td>
<td>(df = 3; 16)</td>
<td>(df = 3; 35)</td>
<td>(df = 3; 3)</td>
<td>(df = 3; 5)</td>
</tr>
</tbody>
</table>

_t_ statistics in parentheses

* _p_ < 0.05, ** _p_ < 0.01, *** _p_ < 0.001
We show the plots of the growth of TFP (y-axis) as a function of the TFP at the previous period \( t - 1 \) (x-axis). The plot in Figure 12-3a represents all crops together. Figures 12-3b, 12-3c and 12-3d are for cotton, pigeonpea and wheat taken individually.

Figure 12-3: Growth in TFP - Share within household estimation

(a) Growth in TFP

(b) Growth in TFP - Cotton

(c) Growth in TFP - Pigeonpea

(d) Growth in TFP - Wheat
iii) Estimation share per sector

We now provide the results when the coefficients of the Cobb-Douglas production function are computed by using the share of the input value in the final output within sector. Table 12.29 gives the coefficients of the Cobb-Douglas function for each sector.

Table 12.29: Coefficients production function - Share per sector estimation - with fixed assets

<table>
<thead>
<tr>
<th></th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor inputs value</td>
<td>1.2</td>
<td>1.2</td>
<td>0.95</td>
</tr>
<tr>
<td>Material inputs value</td>
<td>0.98</td>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td>N</td>
<td>215</td>
<td>250</td>
<td>280</td>
</tr>
</tbody>
</table>
We know show the results of the regression for the model [12.4] and model [12.5].

**Equation model 12.4 (square term)**

Table 12.30: Coefficients growth in TFP
Share per sector elasticities estimation - Square term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>0.900</td>
<td>4.500^*</td>
<td>-12.000</td>
<td>4.800</td>
</tr>
<tr>
<td></td>
<td>(1.100)</td>
<td>(2.600)</td>
<td>(9.900)</td>
<td>(4.700)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>0.051</td>
<td>0.110^*</td>
<td>-0.240</td>
<td>0.220</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.057)</td>
<td>(0.210)</td>
<td>(0.180)</td>
</tr>
<tr>
<td>N</td>
<td>65</td>
<td>82</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>R^2</td>
<td>0.320</td>
<td>0.320</td>
<td>0.450</td>
<td>0.600</td>
</tr>
<tr>
<td>Adjusted R^2</td>
<td>-0.540</td>
<td>-2.900</td>
<td>-1.600</td>
<td></td>
</tr>
<tr>
<td>F Statistic</td>
<td>4.000**</td>
<td>8.400***</td>
<td>1.600</td>
<td>4.400*</td>
</tr>
<tr>
<td></td>
<td>(df = 2; 17)</td>
<td>(df = 2; 36)</td>
<td>(df = 2; 4)</td>
<td>(df = 2; 6)</td>
</tr>
</tbody>
</table>

*t statistics in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
Equation model [12.5] (cube term)

Table 12.31: Coefficients growth in TFP
Share per sector elasticities estimation - Cube term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>-4.600</td>
<td>3.800</td>
<td>386.000</td>
<td>-76.000</td>
</tr>
<tr>
<td></td>
<td>(6.400)</td>
<td>(26.000)</td>
<td>(403.000)</td>
<td>(87.000)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>-0.300</td>
<td>0.081</td>
<td>17.000</td>
<td>-6.200</td>
</tr>
<tr>
<td></td>
<td>(0.400)</td>
<td>(1.200)</td>
<td>(18.000)</td>
<td>(6.900)</td>
</tr>
<tr>
<td>Lag(log_TFP_cube, shift = 1)</td>
<td>-0.007</td>
<td>-0.0005</td>
<td>0.260</td>
<td>-0.170</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.017)</td>
<td>(0.260)</td>
<td>(0.180)</td>
</tr>
</tbody>
</table>

| N                        | 65       | 82       | 29        | 40       |
| R²                       | 0.350    | 0.320    | 0.580     | 0.660    |
| Adjusted R²              | -1.600   | -0.580   | -2.900    | 0.660    |
| F Statistic              | 2.900*   | 5.400*** | 1.400     | 3.200    |

(df = 3; 16) (df = 3; 35) (df = 3; 3) (df = 3; 5)

\( t \) statistics in parentheses

* \( p < 0.05, ** p < 0.01, *** p < 0.001 \)
We show below the graphs representing the growth in TFP (y-axis) versus the TFP at the previous period (x-axis).

Figure 12-4: Growth in TFP - Share within sector estimation

(a)  

(b)  

(c)  

(d)
12.4.3 Growth in TFP - Without fixed assets

i) OLS estimation

When we do not include the fixed assets in the Cobb-Douglas productivity function, the coefficients of this function are given by the following table.

Table 12.32: Coefficients production function - OLS estimation - without fixed assets

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor inputs value</td>
<td>0.120</td>
<td>0.052</td>
<td>0.280^***</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>(0.160)</td>
<td>(0.092)</td>
<td>(0.082)</td>
<td>(0.120)</td>
</tr>
<tr>
<td>Material inputs value</td>
<td>0.450^**</td>
<td>0.280^**</td>
<td>0.530^***</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.210)</td>
<td>(0.130)</td>
<td>(0.110)</td>
<td>(0.190)</td>
</tr>
<tr>
<td>N</td>
<td>499</td>
<td>595</td>
<td>489</td>
<td>574</td>
</tr>
<tr>
<td>R²</td>
<td>0.046</td>
<td>0.023</td>
<td>0.270</td>
<td>0.007</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-1.200</td>
<td>-0.530</td>
<td>-0.510</td>
<td>-0.950</td>
</tr>
<tr>
<td>F Statistic</td>
<td>5.300***</td>
<td>4.500**</td>
<td>43.000***</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>(df = 2; 218)</td>
<td>(df = 2; 379)</td>
<td>(df = 2; 237)</td>
<td>(df = 2; 292)</td>
</tr>
</tbody>
</table>

_t statistics in parentheses

* _p < 0.05, ** _p < 0.01, *** _p < 0.001
We now provide the results of the regression for the model [12.4] and model [12.5] when we do not include the fixed assets in the productivity function.

**Equation model [12.4] (square term)**

Table 12.33: Coefficients growth in TFP
OLS estimation - without fixed assets - Square term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_TFP</td>
<td>-1.00</td>
<td>-0.048</td>
<td>-0.820***</td>
<td>0.430</td>
</tr>
<tr>
<td></td>
<td>(0.70)</td>
<td>(0.94)</td>
<td>(0.20)</td>
<td>(1.10)</td>
</tr>
<tr>
<td>log_TFP_square</td>
<td>-0.014</td>
<td>-0.006</td>
<td>-0.150***</td>
<td>-0.059</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>N</td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td>R²</td>
<td>0.560</td>
<td>0.017</td>
<td>0.640</td>
<td>0.390</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-0.710</td>
<td>0.200</td>
<td>-0.200</td>
<td></td>
</tr>
<tr>
<td>F Statistic</td>
<td>50.000***</td>
<td>1.700</td>
<td>71.000***</td>
<td>40.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 2; 77)</td>
<td>(df = 2; 190)</td>
<td>(df = 2; 79)</td>
<td>(df = 2; 126)</td>
</tr>
</tbody>
</table>

*t statistics in parentheses

* *p < 0.05, ** *p < 0.01, *** *p < 0.001
Equation model 12.5 (cube term)

Table 12.34: Coefficients growth in TFP
OLS elasticities estimation - without fixed assets - Cube term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>2.100</td>
<td>-0.570</td>
<td>-0.920^***</td>
<td>8.500</td>
</tr>
<tr>
<td></td>
<td>(1.800)</td>
<td>(7.000)</td>
<td>(0.200)</td>
<td>(7.400)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>-0.620^*</td>
<td>0.057</td>
<td>0.029</td>
<td>-0.790</td>
</tr>
<tr>
<td></td>
<td>(0.320)</td>
<td>(0.830)</td>
<td>(0.110)</td>
<td>(0.670)</td>
</tr>
<tr>
<td>Lag(log_TFP_cube, shift = 1)</td>
<td>0.035^*</td>
<td>-0.002</td>
<td>-0.025^*</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.032)</td>
<td>(0.014)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>N</td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td>R²</td>
<td>0.580</td>
<td>0.017</td>
<td>0.660</td>
<td>0.390</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-0.013</td>
<td>-0.720</td>
<td>-0.220</td>
<td>-0.190</td>
</tr>
<tr>
<td>F Statistic</td>
<td>36.000***</td>
<td>1.100</td>
<td>50.000***</td>
<td>27.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 76)</td>
<td>(df = 3; 189)</td>
<td>(df = 3; 78)</td>
<td>(df = 3; 125)</td>
</tr>
</tbody>
</table>

*t statistics in parentheses

*p < 0.05, **p < 0.01, ***p < 0.001
We now show the plots of the growth of the TFP versus the TFP at the previous period.

Figure 12-5: Growth in TFP - OLS estimation
ii) Estimation via the share of input value at household level

**Equation model 12.4 (square term)**

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lag(log_TFP, shift = 1)</strong></td>
<td>-0.570^**</td>
<td>-1.500^***</td>
<td>-1.400^***</td>
<td>-0.410^**</td>
</tr>
<tr>
<td></td>
<td>(0.240)</td>
<td>(0.450)</td>
<td>(0.250)</td>
<td>(0.200)</td>
</tr>
<tr>
<td><strong>log_TFP_square</strong></td>
<td>0.032^***</td>
<td>-0.039^***</td>
<td>-0.030^***</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.014)</td>
<td>(0.007)</td>
<td>(0.015)</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td><strong>R²</strong></td>
<td>0.630</td>
<td>0.110</td>
<td>0.280</td>
<td>0.190</td>
</tr>
<tr>
<td><strong>Adjusted R² 0.100</strong></td>
<td>-0.540</td>
<td>-0.620</td>
<td>-0.570</td>
<td></td>
</tr>
<tr>
<td><strong>F Statistic</strong></td>
<td>65.000***</td>
<td>12.000***</td>
<td>15.000***</td>
<td>15.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 2; 77)</td>
<td>(df = 2; 190)</td>
<td>(df = 2; 79)</td>
<td>(df = 2; 126)</td>
</tr>
</tbody>
</table>

*t statistics in parentheses*

*p < 0.05, **p < 0.01, ***p < 0.001
### Table 12.36: Coefficients growth in TFP
Share per household elasticities estimation - Cube term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag(log_TFP, shift = 1)</td>
<td>0.240</td>
<td>-3.700^**</td>
<td>-1.000^**</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>(0.390)</td>
<td>(1.700)</td>
<td>(0.500)</td>
<td>(0.270)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>0.140^***</td>
<td>-0.180^*</td>
<td>-0.005</td>
<td>0.160^***</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.100)</td>
<td>(0.031)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>Lag(log_TFP_cube, shift = 1)</td>
<td>0.003^**</td>
<td>-0.003</td>
<td>0.0005</td>
<td>0.008^***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>N</td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td>R^2</td>
<td>0.660</td>
<td>0.120</td>
<td>0.280</td>
<td>0.270</td>
</tr>
<tr>
<td>Adjusted R^2</td>
<td>0.170</td>
<td>-0.540</td>
<td>0.280</td>
<td>0.270</td>
</tr>
<tr>
<td>F Statistic</td>
<td>49.000^***</td>
<td>8.700^***</td>
<td>10.000***</td>
<td>15.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 76)</td>
<td>(df = 3; 189)</td>
<td>(df = 3; 78)</td>
<td>(df = 3; 125)</td>
</tr>
</tbody>
</table>

`t` statistics in parentheses

* *p < 0.05, **p < 0.01, ***p < 0.001
We show below the graphs representing the growth in TFP as a function of the TFP at the previous period.

Figure 12-6: Growth in TFP - Share within household estimation - without fixed assets
iii) Estimation via the share of input value at sector level

When we compute the Cobb-Douglas function by using the share of the input value at the sector level, the coefficients are given by the following table.

Table 12.37: Coefficients production function - Share per sector estimation - without fixed assets

<table>
<thead>
<tr>
<th></th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor inputs value</td>
<td>1.2</td>
<td>1.2</td>
<td>0.95</td>
</tr>
<tr>
<td>Material inputs value</td>
<td>0.98</td>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td>(N)</td>
<td>215</td>
<td>250</td>
<td>280</td>
</tr>
</tbody>
</table>

The results of the regression are shown in tables 12.38 and 12.39.

Equation model 12.4 (square term)

Table 12.38: Coefficients growth in TFP
Share per household elasticities estimation - Square term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Lag(log}_\text{TFP, shift = 1)})</td>
<td>-0.340</td>
<td>0.018</td>
<td>-3.400^{***}</td>
<td>-0.140</td>
</tr>
<tr>
<td>(N)</td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.540</td>
<td>0.150</td>
<td>0.400</td>
<td>0.340</td>
</tr>
<tr>
<td>Adjusted (R^2) -0.095</td>
<td>-0.480</td>
<td>-0.340</td>
<td>-0.290</td>
<td></td>
</tr>
<tr>
<td>(F) Statistic</td>
<td>46.000^{***}</td>
<td>17.000^{***}</td>
<td>26.000^{***}</td>
<td>33.000^{***}</td>
</tr>
<tr>
<td>(\text{(df = 2; 77)})</td>
<td></td>
<td>(\text{(df = 2; 190)})</td>
<td>(\text{(df = 2; 79)})</td>
<td>(\text{(df = 2; 126)})</td>
</tr>
</tbody>
</table>

\(t\) statistics in parentheses

\(p < 0.05, \, ** \, p < 0.01, \, *** \, p < 0.001\)
### Equation model [12.5](cube term)

#### Table 12.39: Coefficients growth in TFP

Share per household elasticities estimation - Cube term

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>cotton</th>
<th>pigeonpea</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.064</td>
<td>1.100</td>
<td>-1.400</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>(1.100)</td>
<td>(6.400)</td>
<td>(2.300)</td>
<td>(1.600)</td>
</tr>
<tr>
<td>Lag(log_TFP_square, shift = 1)</td>
<td>0.082</td>
<td>0.087</td>
<td>0.054</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>(0.110)</td>
<td>(0.400)</td>
<td>(0.150)</td>
<td>(0.260)</td>
</tr>
<tr>
<td>Lag(log_TFP_cube, shift = 1)</td>
<td>0.001</td>
<td>0.001</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.008)</td>
<td>(0.003)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>N</td>
<td>186</td>
<td>331</td>
<td>177</td>
<td>247</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.540</td>
<td>0.150</td>
<td>0.410</td>
<td>0.340</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>-0.110</td>
<td>-0.490</td>
<td>-0.340</td>
<td>-0.300</td>
</tr>
<tr>
<td>F Statistic</td>
<td>30.000***</td>
<td>11.000***</td>
<td>18.000***</td>
<td>22.000***</td>
</tr>
<tr>
<td></td>
<td>(df = 3; 76)</td>
<td>(df = 3; 189)</td>
<td>(df = 3; 78)</td>
<td>(df = 3; 125)</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The figures below display the growth in TFP as a function of the TFP at the previous period.

Figure 12-7: Growth in TFP - Share within sector estimation - without fixed assets
Conclusion

This section was initially motivated by a paper by Dany Bahar [Bahar, 2018] showing the "middle productivity trap" at the firm-level in the US. In other words, this research showed that fast productivity growth was experienced by both the highest and lowest productive firm, leaving the middle in a situation of stagnation. This result is illustrated by a U-shape relationship between the growth in productivity and the current level of productivity.

For the case of farmers in Semi-Arid Tropic India, we find that the a negative relationship between the growth in productivity and the current level of productivity. This result suggest two conclusions.

First, unsurprisingly farmers cannot grow in productivity exponentially, in particular due to natural constraints. Second, the lowest productive farmers are not necessarily trapped in their current state. It does not allow us to conclude anything about the persistence of poverty traps. Nevertheless, it shows some evidence that tend to mitigate the concept of "low-productivity trap".
Chapter 13

Conclusion

We started this research project aiming at identifying the main variables that drive yield gaps in Semi-Arid Tropic India. We refer the reader to intermediate conclusion enclosed in each chapter in order to grasp into details the results that we found. We provide in this conclusion the “take-home” messages of our investigation and indicate possible policy recommendations. We notice for the reader that recommendations and interpretations that go beyond the quantitative analysis provided by this thesis are the personal thoughts of the author and should be taken with cautious. In particular, more advanced analysis should be required.

13.1 Conclusions from the second part

1. Farmers in semi-arid tropic India have low educational level

Data shows that 34% of first heads in households and 43% of second heads in households have no education (see 10). In this analysis, we do not specifically analyze the impact of education on output production. We nevertheless include this variable in our fixed effects model in chapter II and we do not find significant correlation. Deeper analysis focusing on the impact of education on farming production is necessary. In particular, investigation on the field via interviews and focus groups could pave the way towards more insight about the role of education. We could hypothesize that higher background in elementary calculus could help farmers to manage the use of fertilizers and pesticides.

2. Most farmers lose money from farming. One of the most striking result of this study is
the observation that most farmers lose money from farming \cite{10, 12}. Nevertheless, total revenue of farmers in Semi-Arid Tropics in rural India has increased over the last 20 years because most of them diversify their activities. In particular, the non-farm sector in rural areas has dramatically increased since the 2000’s. From a policy perspective, it might be worth to study the relevance of governmental subsidies (on fertilizer for instance). Indeed, government benefits could potentially be used to help farmers to diversify their economic activities. Alternatively, subsidies could be allocated to customize advice to farmers instead of providing material input without any guidance. Both suggestions are not grounded on empirical experiments and should be carefully assessed with further studies.

3. The non-farming sector progressively replaces farming in rural areas As mentioned above, the non-farming sector has grown since the 2000’s and constitutes a critical means of subsistence. Such result is confirmed by ICRISAT’s scientists. In particular, in Rural Non-Farm Employment and Rural Transformation in India \cite{Reddy et al., 2014}, Dr. Amarendra Reddy finds that “Currently, non-farm sector is no longer a residual sector, but an emerging driver of rural development and transformation, contributing 65% to the rural Net Domestic Product in 2010 […]”. Thus labor scarcity has emerged as one of the major constraints to increasing agricultural production in India […]. Though the macro-level picture may conceal more than what it reveals, the larger patterns may be helpful in raising appropriate questions, and in designing appropriate a priori hypotheses that could be subjected to more rigorous analysis based on micro-level data available from sources like the ICRISAT Village Level Studies […]. “[…] It is likely that any single source of income is not sufficient to meet rural individual or household needs. […]” One of the significant changes in the rural production structure is the growing share of the non-farm sector, which increased from 37% in 1980-81 to 65% in 2009-10 […] and thus shows that in terms of value of production, rural is no longer merely agricultural. This provides much justification for the observation that “the old vision of rural economies purely focused on agriculture no longer fully reflects the reality” \cite{Haggblade et al., 2010}. […]”

4. Within villages, farmers are pretty homogeneous in terms of productivity. To our surprise, we find that within villages, farming outputs are very close to each other. There is usually a high frequency at a single value and another smaller one at zero production. This observation suggests that either farmers produce crop yield and in this case, they are quite
homogeneous in terms of production or their crop totally fails due to exogenous events such as droughts, floods or pest attacks.

5. Within villages, farmers are pretty homogeneous in terms of income and wealth.

6. The lowest productive are more likely to increase their yield. In chapter 12 we found the lowest productive farmers are more likely to increase their yield in the following year, while the most productive are more likely to stagnate or decrease. This result confirms the homogeneity of farmers mentioned in point 5. It also confirms that farmers are constrained by the physical limitations of land.

7. It is difficult to classify in a category in terms of productivity because of large fluctuations from one season to the other. At some point in our analysis, we tried to classify farmers in terms of productivity in order to identify the lowest and the most productive farmers. We nevertheless didn’t find consistent and satisfying groups over time, because the classification fluctuates from one year to the other.

8. Econometrics does not identify systematic variables showing significant correlation with farm productivity. In chapter 11 we use a fixed effects model in order to identify the variables that are correlated with the output rate and yield gaps. We use three different models and four econometric methods (fixed effects, random effects, correlated random effects and quantile regression). Even though some variables turn out to be significant for some crops, we do not find any variables that are systematically significant for every crop that we study (kharif cotton, kharif paddy, kharif pigeonpea, rabi sorghum and rabi wheat). We might hypothesize that other hidden variables have a key influence in farming production and yield gaps.

9. Soil seems to matter but social factors appear to be critical too (cf. the homogeneity within villages that could be explained by the spread of information in village network).
13.2 High-level interpretation of the challenges around agriculture in India

One of the most striking results of our research is that almost all farmers are losing money from their farming activities. In other words, despite governmental subsidies, agriculture as it is practiced today would not be profitable under the current physical constraints (soil, temperature, water) in Semi-Arid Tropics India. This result sheds light on the tremendous challenges that India has to face on the way towards socio-economic development. As reminded in introduction, around 52% of the population live from farming. Discussions with local farmers during our fieldwork in August 2017 and January 2018 provide some evidence that most new generations go to cities in order to find a job as a taxi-driver or employee in hotels for instance. Such examples do not represent the diversity of individual trajectories. Nevertheless, they are symptomatic of a general trend, namely the mutation of the socio-economic model in India.

In our view, crisis in agriculture are only the tipping point of the disruption of the socio-economic paradigm. This mutation should be addressed in a holistic manner that includes education policy, urban planning and environmental policy. The progressive democratization of the education will very likely increase the disaffect for field work, in particular in regions where whether conditions add uncertainty to the harshness of the labor. Paradoxically, it will also very likely increase the inequalities between educated and uneducated people. As more and more young people leave the countryside, the cities are growing, very often in an anarchic manner. Many Indian megalopolis already have to face congestion, air pollution and shortage of housing among others issues. These challenges are not limited to India or developing countries. Nevertheless, the still rapid growth of their population makes these problems harder to overcome. Such situations have already been resolved in the past, in particular in Europe. Nevertheless, never before countries had to face with both economic development and global warming. In our opinion, we are therefore at a “new frontier” where policymaker should worry for both individual’s wellbeing and environmental sustainability. It is an unknown direction for which neither economists nor politicians seem to be ready to confront. The ability of policymakers to satisfy both the economic and environmental constraints will determine whether or not developing nations can reach the standard of living of western countries, if only such a situation is physically attainable. We hypothesize that the definition of economic development in terms of GDP itself should be very
soon outdated, at the risk otherwise of leading the humanity towards dramatic and uncontrolled events.

13.3 Limitations

In this subsection, we highlight some limitations to our analysis.

First, even though the data is of good quality and quite rare for developing nations, the data collection would benefit from asking more details to farmers, in particular about their management practices (fertilizer application, pesticide for instance). Furthermore, our analysis would gain from combining this panel data with more accurate soil data. We realize that information about soil – type of soil, degradation, slope – were sometimes quite approximate and thus decide not to include some of this items (e.g. slope).

Second, we only analyze the data from the campaign 2009-2014. It would be worth to do further analysis by integrating data from 1975-1985 and 2001-2009. This task will require some demanding computational work in order to homogenize the questionnaires and answers. However, such an analysis would provide a larger time frame that very few research papers have analyzed so far. Optimally, more recent data (after 2014) should be included. From our knowledge, such data do not exist.

Third, our econometric model could still be improved. Our methodology has more to do with quantitative methods used in social sciences than economics since we keep a large set of independent variables in our model. Reducing further the number of variables as economists usually do could slightly modify some of our results.

Finally, our study would benefit from including qualitative data. Specifically, our econometric analysis could be associated with a qualitative analysis (focus group, interviews) conducted on the field in order to confirm the results of the quantitative data.
13.4 Directions for further research

We now propose some direction for further investigation. First, for the identification of the variables that influence yield gaps, econometric models have been largely used in the literature. We propose a fixed effects models applied on a panel data. This type of data is rarely investigated in agricultural economics because collecting such a data is expensive and challenging in terms of logistics. Nevertheless, our study does not enable us to identify causal relationships. Randomized controlled trials already used in applied economics will very likely be a key avenue in the near future in order to tackle the question of yield gaps.

Second, further research about the correlated random effects model would be useful. This model proposed by Wooldridge in 2019 [Wooldridge, 2019] to the community of econometricians is not widely spread and still needs further analysis in order to grasp its advantages and limitations. Alternatively, the Correlated Random Coefficient (CRC) model proposed by [Suri, 2011] would be worth to be tested on this dataset. As a further research question, this model could possibly help to assess whether the use of technologies (mechanization, improved seeds) modifies the result that most farmers lose money from farming.

Finally, we suggest to further investigate the networks of farmers in villages. Based on chapter 3, we could hypothesize that farming practices and behaviors spread within villages, thus possibly explaining the observed homogeneity between farmers. Further research is necessary to validate this hypothesis. In this respect, the study by [Cai et al., 2015] about insurance in farming villages is a starting point in this direction.
Bibliography


232


[Giusti et al., 2016] Giusti, C., Ghrist, R., and Bassett, D. S. (2016). Two’s company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data.


