UNIVERSIDAD
EAFIT®

The role of regions in economic growth

Daniel E. Restrepo Montoya

TESIS DOCTORAL

The role of regions in economic growth

Daniel E. Restrepo Montoya

UNIVERSIDAD
EAFIT

2020

The role of regions in economic growth

DOCTORADO EN ECONOMÍA

DEPARTAMENTO DE ECONOMÍA
ESCUELA DE ECONOMÍA Y FINANZAS

Juan Carlos Duque Cardona
Director

Daniel Eduardo Restrepo Montoya
Estudiante

Medellín, 2020

UNIVERSIDAD
EAFIT®

Juan Carlos Duque Cardona, profesor del Departamento de Matemáticas de Eafit.

INFORMA:

Que la tesis titulada *The role of regions in economic growth* presentada por *Daniel Eduardo Restrepo Montoya* para optar al título de Doctor en Economía por la Universidad de EAFIT, ha sido realizada bajo mi dirección en la Escuela de Economía y Finanzas, y considerándola finalizada, autorizo su presentación para ser juzgada por el tribunal correspondiente.

Y para que así conste, firmo la presente, en Medellín el 19 de octubre de 2020.

Juan Carlos

Contents

List of Figures

List of Tables

Abstract

The aim of this work is to develop a novel strategy to encourage selective innovation or diversification patterns in a country through the design of regions for industrial cooperation. More precisely, we propose a framework in which the recent developments in evolutionary economic geography and complexity theory are combined with the techniques of region design to determine suitable regional divisions of a country for the application of economic policies that forest the industrial cooperation within them. We regard these regions as *innovation ecosystems* in which the selective interaction between industries has as ultimate goal the development of complex activities and, therefore, the economic growth of the country. This work is divided in three parts that corresponds with the three articles that constitute the thesis. In the first part, we explore, from a strategic point of view, the implications of encouraging the formation of industrial ecosystems by relocating systematically sets of firms to a given region through subsidies or similar economic incentives. The second part is devoted to the formulation of the optimization model that determines the innovation ecosystems and to the corresponding development of tools to interpret the output of the model. Lastly, in the third part, based on the model proposed in the second part we construct two diversification strategies, (1) a free dynamic model in which the country fosters economic interactions seeking to generate new activities that maximize the total economic complexity improvement of the country, and (2) a guided model in which the main goal consists in developing a target set of activities defined by the economic or political interests of the country. The contributions of this work orbit around the formulation of the p-innovation ecosystems model developed in the second part of the thesis. From a theoretical point of view, this model constitutes the first rigorous link between complexity theory and region design. Additionally, using game theory tools we find out that the strategy of designing regions is more robust with respect to information asymmetries that may face the government bargaining with firms than other policies such as firm relocation. From an applied perspective, we are able to identify various innovation ecosystems divisions for the Colombian economy for application of industrial policy, i.e., we identify a suitable aggregation of the Colombian municipalities using information of its exports for 2014 and we provide diagnostics about the key existing industrial associations and the innovation opportunities; we find two other regionalizations of dynamic nature that allow us to identify diversification strategies and routes for the Colombian economy taking 2010 as starting point. Based on the various models introduced in this work, we assess the current diversification agenda proposed by the current government in its development plan.

Introduction

Since the seminal papers of Robert Solow, the mainstream economic growth analyses have used aggregate growth models (Solow, 1957; Solow, 1956) as their principal workhorses. Abstract quantities like technological progress or positive externalities, such as spillovers or scale effects, usually play a central role in the classic growth theory since they are the primary mechanisms that encourage firm productivity and, therefore, generate economic growth. Nevertheless, given that the aggregative models address the interaction between different goods as a black box, they cannot endogenize the positive externalities that come from the strategic production of complementary goods that leads to the emergence of new economic activities (Hidalgo et al., 2007a). Nowadays, the general consensus is that the diversification of the productive structure (i.e., selective innovation) stands out as a fundamental goal to foster economic growth. Governments like Saudi Arabia Arabia, 2016, Indonesia, Peru and Chile (see Alshamsi et al., 2018) have established diversification as explicit goals in their economic policy agenda.

Understanding the relatedness structure among economic activities thus becomes a key part of the analysis of the diversification and agglomeration of firms. Hidalgo et al. (2007a) found a suitable framework in network theory to understand the impacts of the productive structure of countries and regions regarding their diversification patterns (see also Neffke et al., 2011; Hausmann and Hidalgo, 2011; Hausmann et al., 2014). This approach have shown to be effective to gauge the natural restrictions the economies face to develop complex and desirable economic activities (Hidalgo et al., 2007b). Besides the relatedness restrictions imposed by the natural economic forces, the spatial constraints intrinsic in the diffusion of knowledge or, in general, the capabilities required for the creation of new products have also shaped significantly the industrial ecosystems (e.g., industrial clusters) (Hausmann et al., 2019). For example, Bahar et al. (2014) considered the geographic nature of the capabilities, showing that the spatial proximity among countries explains significantly their current basket of exports and, moreover, evidencing that there is a high and significant spatial correlation between the diffusion of Revealed Comparative Advantage (RCA) of neighboring countries. This geographic effect is justified partially in Bahar et al. (2014) by the strong effect of the distance in the, arguably, more important capability: knowledge (see Thompson and Fox-Kean, 2005, Bottazzi and Peri, 2003; Bahar et al., 2014). Moreover, works like Neffke (2009) have pointed out that knowledge or know-how and other capabilities face strong mobility restrictions even within countries and then the economic regions should be regarded as the spatial unit in which the emergence of capabilities (and then innovation) is more likely to occur.

However, several geographic, institutional and historical factors such as the distances between the firms, the topography, administrative borders, differential taxes among states, productivity, labor availability and so forth are important constraints that may preclude the formation of optimal industrial clusters, i.e., the idiosyncratic factors of the territory can encourage the agglomeration of industries producing less complex industries. This clustering process, although suboptimal, ends up reinforcing itself by the impact of the agglomeration economies that, despite all the possible drawbacks, such as congestion or elevated factor costs, leads the firms to locate close to each other (Diodato et al., 2018 and Rosenthal and Strange, 2004). This mismatch between conditions of spatial nature and initial industrial endowments seems to explain why the classical predictions of urban and regional economics does not hold or are inconclusive in many cases (Martin and Sunley, 1996; De Groot et al., 2016) . More explicitly, it can be argued that the lack of conclusive empirical validation of the aforementioned theories is partially explained by the trade-off between diversity and similarity (see Neffke et al., 2011); labor, knowledge, and knowhow diffuse easily among similar firms, but only the interactions between firms that are sufficiently distinct can induce the generation of new ideas. More precisely, this idea (taken from cognitive theory) explains how a nondirected diversification can either lead to big dissimilarities among firms (cognitive distance), which hamper interactions or lead to an excessive overlapping of skills, which amounts to cognitive lock-in (Neffke et al., 2011).

In summary, we face a setting in which (1) there are (quantifiable) target interactions among firms that enhance the innovation processes, (2) these key interactions could be hindered by or neglected due to incompatible local industrial policies that are inherent to the regional divisions and, (3) these regional divisions are structural and, in most cases, unmodifiable. With this scenario in mind our main goal in this work consists in developing a novel strategy to encourage selective innovation or diversification patterns in a country through the design of regions for industrial cooperation. More precisely, we propose a framework in which the recent developments of evolutionary economic geography and economic complexity theory are combined with the techniques of region design to determine suitable regional divisions of a country for the application of economic policies that forest the industrial cooperation within them. We regard these regions as *innovation ecosystems* in which the selective interaction between industries has as ultimate goal the development of complex activities and, therefore, the economic growth of the country. We justify the creation of innovation ecosystems through region design showing that other strategies to encourage the industrial collaboration, like the systematic relocation of industries using subsidies or similar incentives, leads to undesirable economic outputs such as rent extractions. Additionally, based on the definition of the innovation ecosystems, we propose a series of dynamic models and metrics to encourage and to describe selective diversification process. With this family of models at hand we propose a framework to evaluate two types of innovation strategies, (1) free evolution models in which the industrial ecosystems adapt gradually seeking to maximize the emergence of complex activities, and (2) guided models that direct the economy to the development of products established by the policy makers.

This work is divided in three main chapters corresponding to the three articles that constitute the thesis, and another chapter containing the overall conclusions.

In the first chapter, we discuss the economic consequences of implementing some other strategies different from redesigning the industrial regions. In particular, we analyze the practice of relocating firms using economic incentives, which have been extensively used since the middle of the last century. Our approach proves that this strategy has some inherent problems of it (e.g. rent extraction from the firms to the government), most of them explained by the information asymmetries between the government and the firms. Given the strategic situations that involves the implementation of a relocation strategy, specially the bargaining between the government and the firms to define the economic incentive to encourage the firm to change its location, we propose an alternating offers model with asymmetric information à la Rubinstein. We show that under certain assumptions this strategy always leads to undesirable outcomes like rent extraction. Additionally, using the generality of our model we address and analyze other related situations that involves rent extraction like expropriations of firms by a predatory governments.

In the second chapter, we introduce the main model of the thesis: *the p-innovation ecosystems* (p-IE) model. This model is proposed as a spatially constrained clustering problem belonging to the family of p-regions problems. Our formulation is motivated by the recent developments of economic complexity regarding the evolution of the economic output through key interactions among industries within economic regions. The objective of this model is to aggregate a set of geographic areas into a prescribed number of regions (so-called innovation ecosystems) such that the resulting regions preserve the most relevant interactions among industries. We formulate the p-IE model as a mixed-integer programming (MIP) problem and propose a heuristic solution approach. This theoretical framework enables to draw a first link between regional designing and economic growth. In addition to this we explore a case involving the municipalities of Colombia to illustrate how such a model can be applied and used for policy and regional development. We also introduce a series of metrics and analysis that are relevant by themselves, not only as complements of the model, but as diagnostics of regions in terms of their innovation potential and economic complexity.

Lastly, the third chapter addresses the dynamic issues related with the implementation of diversification strategies in the framework of the p-IE model. We introduce two different long-run diversification strategies that capitalize the initial endowment of activities in a country, to either maximize the level of economic complexity of the country or to pursue specific targeted innovations. We formulate our models as two-step optimization problems, where the first step determines optimal diversification routes for the country, while the second step applies the p-IE model to find an optimal number of innovation ecosystems. We illustrate both strategies analyzing diversification opportunities for Colombia at its municipalities. In this context, we assess a guided diversification strategy proposed by the current Colombian administration (Economía naranja) comparing it with a model that let the industries interact and evolve freely over the time. We show that this guided diversification policy could reinforce the divergence process of the regional economies of the country, boosting the emergence of complex industries in regions with a significantly large initial endowment of capabilities and, at the same time, diminishing the overall complexity of the economy of certain less developed regions.

Chapter 1

Industrial regions or firm relocation? *A game theory approach*

Abstract

In this paper we analyze some economic consequences of the implementation of centralized policies intended to relocate firms. We address this issue with a strategic bargaining model that describes how a government and a firm negotiate subsidies (e.g., tax abatements) that would guarantee the relocation of the firm. Our model is a simplification of the classical alternating offers bargaining model with two variations: a chance of breaking the negotiations with certain probability at the end of each period, and the introduction of imperfect information in the model (uninformed government). This approach allows us to show that the presence of asymmetric information in the bargaining process to relocate the firms induces socially undesirable outcomes, like rents extraction of the firms to the government. We also analyze a slight variation of our model to study the implications of a more drastic action that the government can take to ensure the firms relocation, the expropriation.

1.1 Introduction

The geographical distribution of firms within regions or countries has a clear impact on the dynamics of economic activity. Industrial distribution phenomena such as the formation of clusters and agglomeration economies have played a central role in theoretical issues such as the study of spillovers and the accumulation of human capital, and hence in the study of economic growth (Krugman, 1991a). Policy makers have also been interested in understanding these types of phenomena in order to take advantage of them when pursuing certain public policy goals. In fact, the strategy of using subsidies to relocate firms to desired locations has been used in most industrialized countries since the 1950s, mainly to reduce interregional inequalities in income and employment (Brouwer et al., 2004).

Tax abatements, construction bonds, relaxation of environmental policies and other types of indirect relocation subsidies have been used extensively to change the geographic locations of firms between and within regions, with the goal of creating a better distribution of economic activity. For instance, in the 1950s and 1960s, rapid demographic and economic growth in the cities of many industrialized countries induced a core-periphery phenomenon where the cities prospered at the expense of the peripheral or rural regions, which lagged behind. Local and regional policies addressed this problem, relocating firms in an effort to encourage economic activity in the lagging regions in order to reduce disparities (Pellenbarg et al., 2002). A more recent example of this type of policy can be found in the European Union (EU), which at the beginning of the 2000s spent about one third of its budget on regional support in order to reduce wide disparities among zones of the community (see e.g., Commission of the European Communities, 2001). Mainly, the EU policies consisted of attracting firms to poor regions, mostly to peripheral regions in order to encourage productivity and economic growth in those areas (Ulltveit-Moe, 2007).

Other types of strategies implemented to encourage economic and productivity growth focus on mitigating the boundary effects that result from within-country administrative divisions. These types of strategies need not entail any changes in the relative locations of firms. For example, the New West Partnership Trade Agreement between the Canadian provinces of Alberta, British Columbia, and Saskatchewan seeks to end the constant trade and labor mobility disputes between these provinces (BC-SK-AB, 2016). Another example can be found in Ecuador, which in 2017 modified its administrative divisions to improve the allocation and administration of resources (Maya, 2013). An example of combining the relocation of firms with a regional design strategy can be found in China during the process known as the Third Front Movement, especially during the period 1978-1993. During this time, the Chinese government encouraged the movement of existing factories to the coastal region, and, at the same time, created a series of special economic zones to foster the development of new industries by means of foreign direct investment (Wei, 2007; Wang, 2004).

Despite the success of some countries in applying policies that encourage firms to relocate, several studies have shown that these policies have not been effective at reducing productivity disparities among regions, as is the case in the EU (see Midelfart-Knarvik and Overman, 2002 and Ulltveit-Moe, 2007). In fact, such policies can even increase disparities in cases where the difference in capital ownership among regions is substantial (Dupont and Martin, 2005). If local policies intended to support industrial redistribution also induce competition among regions as they try to attract firms, then regions may become vulnerable if firms threaten to relocate to a different region if no concessions or incentives are granted. This situation could actually worsen the situations of poor regions with regard to welfare (Vogel, 2000).

Since the 1990s, various alternatives have replaced the firm relocation strategy as the principal means of implementing regional policies. These ideas, based principally in concepts like scale effects, agglomeration economies, knowledge spillovers and learning regions, have introduced new strategies to encourage economic growth at a regional level. Examples include regional funds to create local conditions and incentives for innovation as well as the creation of new economic activities (Pellenbarg et al., 2002). However, empirically, it seems that some of these new strategies have not yielded the expected outcomes. For instance, Midelfart-Knarvik and Overman (2002) shows that the policy of Regional Funds in the EU generates inefficient allocation in the realm of industrial activity, as the industrial location pattern induced by this policy is different from the pattern produced by economic forces.

More recently, several theoretical developments in economic geography and regional science have taken new approaches to provide a better understanding of the main problems of the above-mentioned strategies and, at the same time, have opened up new possibilities for implementing policies more coherent with the structure of the regional economies. Certain concepts, such as the capacity of a region to incorporate technologies or knowledge from abroad (absorptive capacity), have turned out to be useful in explaining disparities in productivity among economic regions Jung and López-Bazo (2017a). Moreover, some of these studies, such as (Caragliu and Nijkamp, 2016) have found that certain types of poverty traps are generated when regions with low absorptive capacity are proximate. This problem occurs because the benefits of the investments intended to help the regions accumulate knowledge are not maximized given the lack of absorptive capacity of the local firms. Following this logic, the strategy of relocating firms can be understood now as method to introduce new goods or technologies in regions that were not already producing them, thus mitigating the poverty trap by generating absorptive capacity clubs (Caragliu and Nijkamp, 2016).

In this order of ideas, this type of approach suggests that relocating firms should be effective whenever the firms and the incentives used to relocate them are coherent with the absorptive capacity of the regions and its neighbors and other characteristics of the regional economies that guarantee an effective introduction of the new economic activity. However, as shown by Neffke et al. (2014) in Sweden or by Hausmann and Neffke (2019) in Germany, the pioneers who produce a good in a region are, in most cases, people who already have the required knowhow for producing the good, such as people who had already worked in a related industry in another region. Thus, creating a new industry in a place that lacks workers with the required knowhow also involves the relocation of experienced workers.

In summary, it is still not clear which of these strategies are better at achieving these types of policy goals, since the outcome of each strategy may depend of the characteristics of the country. For instance, countries like China, with a strong and centralized government, can relatively easily implement a program to relocate some firms; whereas in other types of countries that have opted for free-market policies, regional or local policies could, as discussed above, generate competition among the regions to attract firms, leading to undesirable outcomes.

In this paper, we analyze, from a strategic perspective, some economic implications for a government that relocates firms into strategic zones. Explicitly, we represent this situation using a variant of the bargaining model of alternating offers with incomplete information presented in Osborne and Rubinstein (1990). We show that the presence of asymmetric information in the bargaining process to relocate the firms induces socially undesirable outcomes such as rent extraction from the firms by the government. We also use some slight variations of this approach to study the implications of the different types of threats that the government can use to ensure relocation of the firms, for instance, expropriations.

Most studies on firms' relocation, from a game theory perspective, have focused on studying the case in which two regions compete to attract firms. These studies, to the best of our knowledge, have addressed the problem using cooperative bargaining (Nash axiomatic method). Within this context, the main contribution of this paper is to provide a new approximation of the firms relocation problem by means of a non-cooperative bargaining approach.

The rest of this paper is organized as follows. Section 2 provides a literature review. In Section 3 we present the model, and in Section 4 we present the conclusions.

1.2 Literature review

Theoretical and empirical studies in firm relocation have evolved simultaneously over time. Many studies have focused on understanding the determinants of firm relocation, seeking to contribute both information and analytical tools to use in formulating regional planning policies(Brouwer et al., 2004). Most studies in this field were initiated by policy makers who were aiming to solve current economic issues or who were formulating policy agendas based on regional planning (Pellenbarg et al., 2002).

The literature on firm relocation is scarce, given that this phenomenon is usually understood as a particular case of the firm location process. The principal difference between firm location and firm relocation is that the latter takes into account the histories of the firms, seeking to understand the balance between the factors that encourage firms to leave a region and the factors that attract them to another region; these are the so-called push-pull factors (Brouwer et al., 2004). Nonetheless, neither location nor relocation studies have had a clear paradigm, leading to the emergence of multiple approaches and theories that seek to determine the principal factors that induce firms to relocate. A complete survey of the principal location theories used in relocation analysis can be found in Pellenbarg et al. (2002) and in the references therein.

In the 1990s, the ambiguity in the effects of locational or relocational policies as well as the emergence of new policy-making concerns such as environmental policy or urban renewal policy fostered the search for a new methodology to model regional planning and regional development (Moulaert and Sekia, 2003). In some of these studies, attention was centered on the capacity of the regions to create conditions for the development of new technologies, relying on concepts like agglomeration effects, knowledge spillovers or learning regions, all of them condensed in the so-called Territorial Innovation Model (Lagendijk, 2006). These ideas, principally implemented in the EU, have evolved over time into concepts such as the absorptive capacity of regions (already discussed in the introduction) or into strategies such as smart specialization of R&D in countries seeking to take advantage of the productivity potentials of each region (Camagni and Capello, 2013).

Another important theoretical approach that emerged in the 1990s is the so-called New Economic

Geography (NEG), based on the seminal contribution of Krugman (1991c) and developed in the books Krugman (1997) and Fujita et al. (1999). All the models proposed by the NEG have in common the extensive use of agglomeration effects and positive externalities (mostly, scale effects). NEG explicitly shows how some types of externalities arise naturally in the presence of transaction costs and monopolistic competition (Ottaviano, 2003). In other words "*The key contribution of the new economic geography is a framework in which standard building blocks of mainstream economics (especially rational decision making and simple general equilibrium models) are used to model the trade between dispersal and centripetal forces*" (Neary, 2001, p. 536). In fact, NEG models are a useful approach when analyzing how the interaction between firms and pecuniary externalities induces changes in regional patterns of growth as well as in the locations of firms and workers.

The extensive use of aggregative models, such as Dixit and Stiglitz's model of monopolistic competition (see Dixit and Stiglitz, 1977), makes the NEG approach useless to explain the behavior of individual firms, as explained in the introduction (see also Neary, 2001). Furthermore, as most NEG models use schemes of monopolistic competition, the firms modeled are nearly homogeneous, playing the almost passive role of profit-maximizing entities that cannot influence the terms or the regional policies under which they exist, i.e. they have no bargaining power (see e.g., Okubo, 2012). However, the simplicity of NEG models makes them a versatile tool with which to analyze policies that involve change in the global incentives of the firms, like general tax abatements or regional transfers of funds (see e.g., Ulltveit-Moe, 2007; Dupont and Martin, 2005).

Early investigations, such as Vogel (2000), have pointed out that the strategic behavior of the agents involved in the implementation of a regional policy may generate undesirable outcomes. For instance, when two regions compete to attract a firm, the firm gains bargaining power, which enables it to impose conditions to secure more profitable deals with the regions. In the same spirit, empirical works such as Ciabuschi et al. (2012) or Chino (2016) have shown the direct relation between bargaining power and rent-seeking behavior, the former for subsidiaries of multinational corporations and the latter for labor unions in firms. Another type of result is obtained by Greaker (2003). This paper presents a model of strategic interaction between two regions and two firms where the regions define the level of environmental regulation to be implemented and the firms threaten to relocate if the environmental policies are too strict. In this particular case, the relocation threat (an expression of bargaining power) combined with the costs of relocation yield higher welfare for the region and stronger levels of environmental regulation than in the case where the firms are not mobile. In any case, it is clear that the outcome of the policy is highly dependent on the bargaining power of the regions and firms involved.

Situations where governments or regions within a country are involved in strategic schemes to secure optimal policies have been extensively studied. One example is a situation in which governments or regions compete to attract firms by using differentiated subsidies (see e.g., Burbidge et al., 2006; Baldwin and Okubo, 2014; Haufler and Stähler, 2013 and the references therein). To the best of our knowledge, the most recent work that address this problem using theoretical bargaining methods is Han and Leach (2008). In this work, the bargaining is modeled using Nash's axiomatic bargaining solution (see e.g., Binmore et al., 1986), in which the agreement is reached by the maximization of the joint gains of bargaining. For this particular case, given that all the parties have symmetric power of bargaining, the outcome must be equally distributed between all of them. Related examples of Nash's axiomatic bargaining solution are, for example, modeling optimal locations for bilateral monopolists (see e.g., Brekke and Straume, 2004) or modeling the interaction among different institutional or regional parties of a government to decide public policies (see e.g., Luelfesmann et al., 2015).

Non-cooperative approaches have also been applied to analyze the competition among regions or countries to establish local economic policies. A recent example of this can be found in Dong et al. (2017). In this paper, the authors study a situation where a host country and an investing country bargain over the degree of openness with respect to international capital flows. Dong et al. (2017) use an alternating-offers model with a two-sided uncertainty (see e.g., Rubinstein, 1985). So far, to the best of our knowledge, there is no work that analyzes, using a non-cooperative approach, the bargaining process between a government and a firm to decide subsidies (tax abatements, etc.) that would guarantee the relocation of the firm.

1.3 The model

In this section, we propose a bargaining model of alternating offers; our model is closely related to the model discussed in Chapter 5 of Osborne and Rubinstein (1990). The alternating offers model was initially studied in Rubinstein (1982), which proved the existence and uniqueness of a subgame perfect equilibrium under certain conditions. Several extensions of the model have been proposed, one of which changes the classical time preference by adding a chance of breaking the negotiations with certain probability at the end of each period (see e.g., Chapter 7 Osborne and Rubinstein, 1994); another of these variations introduces imperfect information in the model (see e.g., Rubinstein, 1985 and Gul and Sonnenschein, 1988) in order to analyze the possible delays in reaching an agreement between the parties due to the presence of uncertainty. Essentially, our model combines these two variations to study the bargaining of subsidies between a firm and an imperfectly informed government.

We propose a game between two players, Player 1 is the government and Player 2 is a firm. As discussed above, we are interested in analyzing the situation where both parties are involved in a negotiation to determine the amount of subsidies, tax breaks, that the government has to give to the firms in order to achieve its policy goals.

The general assumptions about the environment of the game are summarized as follows:

- The government is willing to offer a subsidy of at most $V > 0$ to the firm.
- The firm has some associated costs in moving from its original location, namely $\gamma \in [0, V]$.

The next assumption marks the principal difference between our model and the classical bargaining models, which include the impatience of the players (or cost of delay).

- After the rejection of an offer by either of the two parties, a chance c moves, finishing the bargaining with a fixed probability $\alpha \in (0, 1)$, and continuing the game with probability $1 - \alpha$.

In each situation that we will model, the main goal of the government is to achieve its long-run policy objective, such as relocating the firm or defining new environmental policies. Thus, the government's payoff from the bargaining is not penalized by delays in reaching the agreement. Because the firm does not change its current activities during the bargaining period, the precise time when the agreement is achieved does not matter. However, the credibility of the government depends heavily on its capacity to attain its policy goals; thus, each time that one of the players rejects an offer, the government loses political credibility and people's trust. Thus, in these models, α stands for the political pressure to implement the policy.

We also follow the standard assumption that both the firm and the government are risk-neutral. This assumption relies on the fact that these types of agents can pool risk over many projects, given the diversification of their incomes. However, in the case of the government, Arrow and Lind (1970) showed that the governments are actually risk-neutral because they have the capability to distribute the risk associated with any investment among a large number of people. For a recent discussion of this subject (see Randall, 2014).

1.3.1 Description of the game

This first model analyzes a situation where the government is bargaining with the firm over a subsidy to incentivize the firm to move from its actual location to a new region. In this case, the political pressure (α) comes from the fact that the government can implement an alternative policy such as a redesigned regional policy or any other of the strategies already discussed, which partially fulfills the policy goals. The existence of this alternative is translated into political pressure as discussed above, and it also provides the government with a credible threat (a significantly big value of α) to ensure the cooperation of the firm.

In this case, Bernoulli's utility of the government is given by $u_1(S) = V - S$ and Bernoulli's utility of a firm with relocation cost γ is given by $u_{2_\gamma}(S) = S - \gamma$ for any $S \in [0, V]$. Let us suppose in this case that there are two types of firms, a firm with low relocation costs l and a firm with high relocation costs h, as well as two types of subsidies that can be offered in the negotiation, a low subsidy S_l and a high subsidy S_h such that

$$0 < l < S_l < h < S_h < V.$$

A complete information alternating offers game between the government and a firm 2_γ consists of an extensive game with perfect information (see e.g., Osborne and Rubinstein, 1994) between two players with the following dynamic: The government moves first, making a proposal that can be either S_l or S_h, after which the firm 2_γ accepts or rejects the offer. If the firm accepts the offer, the game ends and each player receives the corresponding payoffs defined by their utility function; otherwise, a random variable (a chance) ends the game with probability α. If the game ends, we say that the government decided to implement the alternative policy, expending the entire budget

V, and the firm does not receive any subsidy, implying that both players obtain a 0 payoff in this case. In the other case, if the game does not end, it leads to the next period, where the firm 2_γ makes a proposal that the government accepts or rejects. Again, the acceptance finishes the game and rejection leads to a movement of the chance. In the case where the proposal is accepted, the outcome is a lottery between the outcome obtained by the bargaining and a payoff of 0, given the possibility of a breakdown in the negotiations; the probabilities of this lottery are determined by the number of times where the chance c intervened in the game. In the case where the proposal is rejected, if the chance does not finish the game, the game reaches the next period, where it is Player 1's turn to make the proposals again. The game continues in this fashion, only finishing if one player accepts a proposal or if the chance c ends it.

Precisely, this game is defined as an extensive game with perfect information with chance movement. Namely, a tuple $\Gamma = \langle N, H, P, f_c, (\succeq_i)_{i=1,2_\gamma} \rangle$ (see Definition 89.1, Osborne and Rubinstein, 1994), where the set of players N is formed by the government and the firm 2_γ and the preferences $\succeq_i$ for $i = 1, 2_\gamma$ are represented by the utility functions of the players defined above. If we represent accept by A, reject by R, breakdown by B, continue by C and a generic offer in the period n by $x_n \in \{S_l, S_h\}$, then all the possible histories H are of the form

1. $\emptyset$ (initial story).
2. $(x_0, R, C, x_1, \cdots, C, x_n)$.
3. $(x_0, R, C, x_1, \cdots, C, x_n, R)$.
4. $(x_0, R, C, x_1, \cdots, C, x_n, R, C)$.
5. $(x_0, R, C, x_1, \cdots, C, x_n, A)$.
6. $(x_0, R, C, x_1, \cdots, C, x_n, R, B)$.

It is worth noting that, despite this game having an infinite horizon, only histories with finitely many periods can occur with positive probability; hence, we do not include histories with infinite length. By our previous description, we have that the histories of types 5 and 6 are terminal, and we also have that the player whose turn it is to move chooses either S_l or S_h after the histories of types 1 and 4, and that player chooses between A and R after histories of type 2. In this case, the chance c determines whether the game continues or not, as we explained above, and in each case, the function f_c assigns (independently in each case), α to B and $1 - \alpha$ to C. The player function P is defined by $P(\mathbf{h}) = 1$ for histories $\mathbf{h}$ of the form 1, of the form 2 if n is odd and of the form 3 if n is even; $P(h) = 2$ if $\mathbf{h}$ is of the form 2 if n is even and of the form 3 if n is odd and $P(\mathbf{h}) = c$ if $\mathbf{h}$ is of the form 3.

In this case, a strategic profile is defined as a function that assigns a possible action to each nonterminal history; i.e., A or R after histories of type 2 or S_l or S_h after histories of type 4, in each case for the player whose turn it is to move. On the other hand, given the uncertainty introduced by the chance c the outcomes are lotteries over the terminal histories, explicitly, a lottery between the breakdown B and the subsidy decided at the end of the negotiation. Hence, given our assumptions about the players, a Nash equilibrium is a strategy profile that induces an expected outcome that cannot be improved by any of the players unilaterally changing his own strategy. Analogously, a subgame perfect equilibrium is defined as a strategy profile that induces a Nash equilibrium in each subgame of Γ.

1.3.2 Subgame perfect equilibria for the perfect information case

As pointed out by Rubinstein (1982), a Nash equilibrium is not a suitable equilibrium concept for this kind of game because many strategy profiles that admit incredible threats arise as Nash equilibria; given this, the subgame perfect equilibrium refinement becomes a good alternative to discard unreasonable Nash equilibria. Indeed, we will show that in each case, this refinement isolates a unique Nash equilibrium, similarly to the original model proposed by Rubinstein.

Clearly, if the government plays an alternating offers game with 2_h, the unique subgame perfect equilibrium would be given by the case where player 1 only offers S_h and only accepts S_l and where player 2_h only accepts and offers S_h. In this case, the agreement is reached immediately with payoffs $(V - S_h, S_h - h)$. In the case where the government faces a firm with low relocation costs, the bargaining process is a little more interesting. In order to analyze this situation, we introduce some required notation.

Definition 1. *For each positive integer n we define the (possibly empty) intervals*

$$I_n := \left(1 - \left[\frac{V - S_h}{V - S_l}\right]^{\frac{1}{n+1}}, 1 - \left[\frac{S_l - l}{S_h - l}\right]^{\frac{1}{n}}\right), \qquad J_n := \left(1 - \left[\frac{S_l - l}{S_h - l}\right]^{\frac{1}{n+1}}, 1 - \left[\frac{V - S_h}{V - S_l}\right]^{\frac{1}{n}}\right).$$

Proposition 2. *For an alternating offers model between the government and a firm with low cost, we have that:*

1. *If $\alpha > 1 - \frac{S_l - l}{S_h - l}$, then in any subgame perfect equilibrium and any subgame where Player 1 moves first, he will always offer only S_l and Player 2_l will accept any type of subsidy.*

2. *If $\alpha > 1 - \frac{V - S_h}{V - S_l}$, then in any subgame perfect equilibrium and any subgame where Player 2_l moves first, he will always offer only S_h and Player 1 will accept any type of subsidy.*

3. *For $\alpha \in I_n$, if n is odd, we have that in any subgame perfect equilibrium and any subgame where Player 1 moves first, he will always offer S_h and Player 2_l will only accept S_h. On the other hand, if n is even, we have that in any subgame perfect equilibrium and any subgame where Player 2_l moves first, he will always offer S_h and Player 1 will accept any type of subsidy.*

4. *For $\alpha \in J_n$, if n is even, we have that in any subgame perfect equilibrium and any subgame where player 1 moves first, he will always offer S_l, and Player 2_l will accept any offer. On the other hand, if n is odd, we have that in any subgame perfect equilibrium and any subgame where Player 2_l moves first, he will always offer S_l and player 1 will accept only S_l.*

Remark 3. *Let us notice that for any value of n, at most one of the intervals, I_n or J_{n+1}, can be empty. Moreover, given $\alpha < 1 - \frac{S_l - l}{S_h - l}$, which does not coincide with an endpoint of any interval I_n or J_n, we have that there exists a maximal n such that one of the following conditions holds*

$$n \text{ is odd and } \quad \alpha < 1 - \left[\frac{S_l - l}{S_h - l}\right]^{\frac{1}{n}} \qquad \text{or} \qquad n \text{ is even and } \quad \alpha < 1 - \left[\frac{V - S_h}{V - S_l}\right]^{\frac{1}{n}}.$$

Assuming (without loss of generality) the first case, the maximality of n implies that

$$\alpha > 1 - \left[\frac{V - S_h}{V - S_l}\right]^{\frac{1}{n+1}}$$

implying that $\alpha \in I_n$. This implies that Proposition 2 gives a complete classification of the subgame perfect equilibria of this game for any value of $\alpha \in (0, 1)$, except for the cases when α coincides with a boundary point of the intervals I_n or J_n, where one of the agents becomes indifferent between two different options.

Proof.

1. Consider any subgame where player 1 moves first, since $\alpha > 1 - \frac{S_l - l}{S_h - l}$, the expected value of any agreement obtained after the first period for the player 2_l is lower than the payoff obtained by accepting S_l in the first period. Hence, in any subgame perfect equilibrium, player 2_l will always accept S_l. Given this, player 1 will always offer S_l, which gives him the highest possible payoff.

2. If player 2_l moves first, the condition $\alpha > 1 - \frac{V - S_h}{V - S_l}$ ensures that the expected value for Player 1 of any agreement reached after the first movement of Player 2_l will be lower than the payoff obtained by accepting S_h in the first period. Given this, Player 2_l chooses to always offer S_h knowing that it will not be rejected.

3. Consider any subgame where Player 1 moves first, if $\alpha \in I_n$ and n is even, it follows that the expected payoff of any agreement reached after (at least) $n + 1$ periods is lower than the payoff obtained by accepting S_h in the first period. At the same time, these conditions imply that an agreement reached in the n-th period, where Player 2_l offers S_h and player 1 accepts, is still more profitable for Player 2_l than accepting S_l in the first period. Hence, it follows that Player 1 does not have a credible threat to continue the game after the n-th period, but since Player 2_l knows this, he will only accept and offer S_h. Finally, to avoid delaying in the agreement, Player 1 offers S_h from the beginning of the game. The analysis for the case where 2_l moves first and n is odd follows analogously.

4. In this case, if $\alpha \in J_n$ and Player 1 moves first, it follows that Player 1 has incentives to offer S_l up to the n-th period. However, for Player 2_l, it is better to accept S_l in the first period than continue the bargaining for more than n periods. The rest of this proof proceeds, mutatis mutandis, like the previous one.

□

The following corollary summarizes the possible outcomes of the negotiation between Player 1 and Player 2_l.

Corollary 4. *In each case analyzed in Proposition 2, the agreement is reached immediately (in the first period). All the possible outcomes of these agreements are summarized in Table 1.1.*

α	Player 1's outcome	Player 2_l's outcome
$\left(1 - \frac{S_l - l}{S_h - l}, 1\right)$	$V - S_l$	$S_l - l$
I_n, n odd	$V - S_h$	$S_h - l$
J_n, n even	$V - S_l$	$S_l - l$

Table 1.1: Payoffs of the bargaining between Player 1 and Player 2_l.

The three types of subgame perfect equilibria summarized in Table 1.1, though apparently different, follow the same rationality. The first one tells us that when α is big enough, for Player 2_l, the game has essentially one period, because from the second period onwards, all the payoffs are strictly dominated by the worst possible payoff that Player 2_l can receive in the first period. In the same way, the second and third types of equilibria also reduce (essentially) the bargaining to a sequential game with a finite horizon, given that some of the two players will not have incentives to continue formulating offers from some moment onwards. Hence, a kind of backward induction rationality is implemented by the players to achieve the subgame perfect equilibrium. This rationality implies that the player with the longest bargaining horizon can force the other party to accept (or even propose) his most preferred alternative.

1.3.3 Relocation bargaining model with asymmetric information

Let us suppose now that Player 1 cannot recognize whether he is facing Player 2_l or 2_h. We model this situation as a Bayesian extensive game with observable actions and chance movement (see Definition 231.1 in Osborne and Rubinstein, 1994). There are two possible types of Player 2, 2_l and 2_h, that can be selected with a given probability $\pi_h \in (0, 1)$ and $1 - \pi_h$, respectively.

In period 0, Player 1 formulates a proposal $x_0 \in \{S_l, S_h\}$, which players 2_h and 2_l either accept or reject; if the proposal is effectively accepted, the game ends with a subsidy of x_0 for the respective type of Player 2 and for Player 1 in this situation. If the proposal is rejected, the chance c moves exactly as in the complete information case, giving a payoff of 0 for each player if the game ends, and leading to the next period in the other case. In the new period, Player 2 makes a counteroffer $x_1 \in \{S_l, S_h\}$ that is observed by player 1, even though he can not determine if the counteroffer was proposed by 2_l or by 2_h. If the proposal is accepted, the outcome will be a lottery, as explained in the complete information case; the only difference is that, in this case, the lottery received by Player

1 also involves the probability π_h because he is facing this new kind of uncertainty. Again, if Player 1 rejects the offer and if the chance c does not induce a breakdown, the game reaches the second period, where it is Player 1's turn to make the proposals again. The game continues in this fashion, only finishing if one player accepts a proposal or if the chance c ends it.

This game is defined as a tuple $\Gamma(\pi_h) = \langle N, H, P, f_c, (\Theta_i)_{i=1,2}, \pi_h, (U_i)_{i=1,2}\rangle$, where $N = \{1, 2_l, 2_h\}$ is the set of players, H, P and f_c are respectively the set of histories, the player function and the density function of the chance movement, defined as in the previous section. The sets $(\Theta_i)_{i=1,2}$ are the sets of possible types of players with $\Theta_1 = \{1\}$ and with $\Theta_2 = \{2_h, 2_l\}$ and $\pi_h \in (0, 1)$ determine the (a priori) probability of finding a player of type 2_h which is independent from the movements of the chance c. Finally, the functions $(U_i)_{i=1,2_l,2_h}$ are the von Neumann-Morgenstern utility functions over the terminal histories associated with the Bernoulli utility functions $(u_i)_{i=1,2}$, previously defined.

As in all Bayesian extensive games with observable actions, this game also has an equivalent representation as an extensive game with imperfect information. However, the latter representation does not have any proper subgame whereby the equilibrium concept (subgame perfect equilibrium) is not useful to analyze this situation. Hence, it is necessary to use a different kind of refinement that precludes the implementation of incredible threats; to this end, we use perfect Bayesian equilibrium as our equilibrium concept (Definition 232.1 in Osborne and Rubinstein, 1994). This equilibrium concept, in addition to establishing a strategy for each player, also provides a sequence of probability measures for each stage of the game for player 1, enabling him to update his beliefs about the typology of player 2. The following definition provides a precise description of a system of beliefs for the game $\Gamma(\pi_h)$

Definition 5. *A system of beliefs for* $\Gamma(\pi_h)$ *is a function*

$$b_h : \{\boldsymbol{h} \in H | P(\boldsymbol{h}) = 1\} \longrightarrow [0, 1],$$

i. e. b_h *assigns to any history* $\boldsymbol{h}$*, after which Player 1 has to move, the probability that he gives to the event that his opponent is* 2_h*.*

In this paper, we are interested only in pure strategies, and therefore we do not introduce any type of strategies, such as behavioral strategies, where the players use random devices to select their moves. In this case, a strategy profile is triplet of strategies, one for Player 1, and two for Player 2 (one for each type), where the strategies are defined as in the previous section. The following definitions specify the conditions that we will impose on a strategy profile and on a belief system in order to establish a perfect Bayesian equilibrium.

Definition 6 (Outcome function)**.** *Given a history* $\boldsymbol{h}$*, a strategy profile* $s = (s_1, s_l, s_h)$ *that generates at least one history that contains* $\boldsymbol{h}$ *and a belief system* b_h *of the game* $\Gamma(\pi_h)$*, we define the compound lottery* $L(s, b_h|\boldsymbol{h})$ *on the set of terminal histories of*$\Gamma(\pi_h)$ *in which the players use the strategy profile s, Player 1 updates his beliefs according to* b_h *and that history* $\boldsymbol{h}$ *has occurred. This compound lottery assigns probability* $b_h(\boldsymbol{h})$ *to the lottery induced by* (s_1, s_h) *over the terminal histories and* $1 - b_h(\boldsymbol{h})$ *to the lottery induced by* (s_1, s_l) *over the terminal histories.*

Definition 7 (Sequential rationality). *We say that a strategy profile $s = (s_1, s_l, s_h)$ of the game $\Gamma(\pi_h)$ is sequentially rational given a system of beliefs b_h*

- *If for every $\boldsymbol{h}$ that can be generated by s such that $P(\boldsymbol{h}) = 1$, does not exist a strategy s'_1 that can generate h such that $L(s'_1, s_l, s_h, b_h|\boldsymbol{h})$ is strictly preferred to $L(s_1, s_l, s_h, b_h|\boldsymbol{h})$ under U_1.*
- *If for $I = h, l$, Player 2_I cannot improve its utility U_{2_I}, changing his strategy unilaterally.*

For the consistency property of the beliefs, we follow the condition introduced in chapter 5 of Osborne and Rubinstein (1990), which plays the role of the consistency in the beliefs originally introduced by Kreps and Wilson (1982).

Definition 8 (Consistency). *We say that a system of beliefs b_h is consistent with a strategy profile $s = (s_1, s_l, s_h)$ of the game $\Gamma(\pi_h)$ if the following conditions hold.*

- *The initial belief corresponds to π_h, i.e. $b_h(\emptyset) = \pi_h$.*
- *Let $\boldsymbol{h} = (x_0, R, C, x_1, \cdots, C, x_n, R)$ and $\boldsymbol{h'} = (x_0, R, C, x_1, \cdots, C, x_{n+1}, R, C, x_{n+2})$ with n odd. If both strategies s_l and s_h call for the two types of Player 2 to reject x_{n+1} and to propose x_{n+2} then $b_h(\boldsymbol{h}) = b_h(\boldsymbol{h'})$. If $b_h(\boldsymbol{h}) \neq 1$ and only s_l rejects x_{n+1} and counteroffers x_{n+2} then $b_h(\boldsymbol{h}) = 0$; on the other hand, if $b_h(\boldsymbol{h}) \neq 0$ and only s_h rejects x_{n+1} and counteroffers x_{n+2} then $b_h(\boldsymbol{h}) = 1$.*

Notice that the second condition of the previous definition corresponds to the requirement that Player 1 uses, whenever possible, the Bayes rule to update his beliefs about Player 2. On the other hand, this condition also precludes the possibility that Player 1 changes his own beliefs by his own actions. Finally, notice that we do not impose any restriction on the beliefs of Player 1 in situations where he observes a history inconsistent with the strategies s_l and s_h; in this case, Player 1 is free to form new beliefs.

With these ingredients, we are in a position to define our solution concept to analyze $\Gamma(\pi_h)$.

Definition 9. *A perfect Bayesian equilibrium for $\Gamma(\pi_h)$ consists of a strategy profile s and a system of beliefs that are sequentially rational and consistent.*

Notice that our definitions of sequential rationality and consistency imply the 4 conditions of Definition 232.1 in Osborne and Rubinstein (1994).

1.3.4 Subgame perfect equilibria for the perfect information case

In subsection 3.1, we showed that Player 2_h plays a passive role in the alternating offers model, since he always offers and accepts only s_h. In the case when Player 1 is imperfectly informed, this situation does not change at all; hence, our analysis is simplified again to consider only the possible strategies of Player 1 and Player 2_l. The following proposition characterizes all the possible perfect Bayesian equilibria of $\Gamma(\pi_h)$.

Proposition 10. *For any perfect Bayesian equilibrium of $\Gamma(\pi_h)$ we have that:*

1. *If $\alpha > \left(\frac{1-\pi_h}{\pi_h}\right)\frac{S_h-S_l}{V-S_h}$, then in any subgame of $\Gamma(\pi_h)$ in which the prior belief is π_h and where Player 1 moves first, he always offers S_h, which Player 2 accepts independently of his type.*

2. *For $\alpha < 1 - \left[\frac{S_l-l}{S_h-l}\right]^{\frac{1}{n}}$ such that $V - S_h > (1-\alpha)^{n+1}(1-\pi_h)(V-s_l) + (1-\alpha)^{n+2}\pi_h(V-s_h)$, if n is odd we have that in any subgame in which the prior belief is π_h and where Player 1 moves first, he will always offer S_h and player 2_l will only accept S_h. On the other hand, if n is even, we have that in any subgame in which the prior belief is π_h and where Player 2 moves first, Player 2_l will always offer S_h and Player 1 will accept any type of subsidy.*

3. *For $\alpha > 1 - \left[\frac{S_l-l}{S_h-l}\right]^{\frac{1}{n+1}}$ such that $V - S_h < (1-\alpha)^{n}(1-\pi_h)(V-s_l) + (1-\alpha)^{n+1}\pi_h(V-s_h)$, if n is even (or 0), we have that in any subgame in which the prior belief is π_h and where Player 1 moves first, he always will offer S_l and Player 2_l will accept any offer. On the other hand, if n is odd, we have that in any subgame in which the prior belief is π_h and where Player 2 moves first, Player 2_l will always offer S_l and player 1 will accept only S_l.*

Remark 11. *By an argument analogous to the one presented in Remark 3, it can be shown that the previous proposition covers all possible cases except for those that generate indifferences in at least one of the players. Unfortunately, in this case, the conditions for Player 1 cannot be written in a compact form given that these involve finding the roots of a polynomial of degree n.*

Proof. 1. Given that Player 2_h only offers and accepts S_h, the best possible outcome that Player 1 can get offering S_l in the first turn will be $(V-S_l)(1-\pi_h) + (1-\alpha)\pi_h(V-S_h)$. On the other hand, given that $\alpha > \left(\frac{1-\pi_h}{\pi_h}\right)\frac{S_h-S_l}{V-S_h}$, we have that this payoff is strictly lower than the corresponding payoff to offer S_h in the first turn $(V-S_h)$, which is immediately accepted by 2_l and 2_h.

2. As proven in Proposition 2,the former condition implies that Player 2_l has incentives to offer S_l at least until the n-th period, whereas the latter condition implies that it is better for Player 1 to offer S_h in period 0 than face a lottery in which he agrees to a subsidy of S_l after $n+1$ periods with Player 2_l with probability $(1-\alpha)^{n+1}(1-\pi_h)$, or he agrees to a subsidy of S_h after $n+2$ periods with Player 2_h with probability $(1-\alpha)^{n+2}\pi_h$, or he gets 0 in any other case. These considerations imply that Player 1 does not have a credible threat to continue the game for more than n periods, whereas for both types of Player 2, it is still profitable to offer S_h until that period, hence, the sequential rationality implies that Player 1 must offer S_h immediately and that Players 2_h and 2_l should accept only S_h.

   The other case follows from the same reasoning; thus, we omit its proof.

3. From Proposition 2, we know that the former condition implies that Player 2_l prefers a payoff of S_l in the first payoff than getting S_h with a delay of n periods, whereas the latter condition implies that Player 1 prefers the lottery explained in the last part to a certain payoff of $V-S_h$ reached in period 0. These considerations imply that Player 2_l does not have a credible threat to continue the game for more than n periods, whereas, if Player 2_l always offers and accepts

S_h, the consistency implies that Player 1 does not change his prior beliefs and therefore it is still profitable for him to offer S_l until the n-th period. Thus, the sequential rationality implies that Player 2_l must accept S_l immediately, implying by consistency that the player who rejects S_l in period 0 of the subgame must be 2_h with probability 1, hence, in the next period, Player 1 accepts his offer, which should be S_h.

The other case follows from applying an analogous argument.

□

For further reference, we summarize the possible outcomes of the game $\Gamma(\pi_h)$ in the following corollary.

Corollary 12. *There exist two types of perfect Bayesian equilibria for the game* $\Gamma(\pi_h)$*:*

1. ***Pooling equilibrium:*** *Player* 1 *offers* S_h *in period 0, which both Player* 2_h *and* 2_l *accept immediately.*
2. ***Separating equilibrium:*** *Player* 1 *offers* S_l *in period 0, which Player* 2_l *accepts and Player* 2_h *rejects; the latter (if the game continues) offers* S_h *in the next period, which Player* 1 *accepts.*

Moreover, each of the cases analyzed in the Proposition 10 leads to one of these types of equilibrium according to Table 1.2.

Condition on α	**Type of equilibrium**
$P_0(\alpha) < V - S_h$	Pooling
$P_{n+1}(\alpha) < V - S_h$ $\alpha < 1 - \left[\frac{S_l - l}{S_h - l}\right]^{\frac{1}{n}}$ n odd	Pooling
$P_n(\alpha) > V - S_h$ $\alpha > 1 - \left[\frac{S_l - l}{S_h - l}\right]^{\frac{1}{n+1}}$ n even or 0	Separating

Table 1.2: Classification of perfect Bayesian equilibria of $\Gamma(\pi_h)$.

With $P_n(\alpha) := (1-\alpha)^n(1-\pi_h)(V-s_l) + (1-\alpha)^{n+1}\pi_h(V-s_h)$.

The previous corollary shows us that some basic features of the alternating offers game between Player 1 and Player 2_l are preserved in this case. For instance, from the results presented in Table 1.2, we conclude in this case that the player who has incentives to continue the game for more periods can induce the other player to agree to his most preferred subsidy. However, some substantial changes are evident. First of all, if π_h is not too low, when α is big enough, Player 1 has no incentives to offer the low subsidy; moreover, Player 2 induces Player 1 to offer the high subsidy S_h independently of Player 2's type. Second, Player 1 only identifies the typology of Player 2 when his bargaining horizon is longer, i.e., in a separating equilibrium. Finally, in this case, the presence of imperfect information may lead to a delay of one period in reaching the agreement between the parties, in the separating equilibria.

1.3.5 A variant of the model

We propose a slight variation of the model where the government strengthens its threat, i.e., changes the alternative strategy into a more extreme one. Explicitly, the only aspect that changes with respect to the model already discussed is that when the chance c ends the game with probability α, the government decides to expropriate the firm and then relocate it. In order to introduce this variation, let us make the following assumption.

A_1) When the change c ends the game, the players 2_h and 2_l receive a payoff of $-h$ and $-l$, respectively, and Player 1 receives a payoff of $-T$ for some fixed $T > 0$.

In this paper we will use the concept of expropriation in a broad sense, covering from the expropriation of the assets of the company to an intervention in its financial decisions. Following this log, we argue, following Wydick (2007), that when property rights are fuzzy, firms are discouraged from improving their economic performance because they are uncertain about their right to appropriate returns from their own assets. This implies that once the expropriation is executed, negative incentives are generated, inducing outcomes contrary to the government's policy goals. This discussion is reflected in the new payoff assumed for Player 1 in case of breakdown, whereas the payoffs for Player 2_h and 2_l simplify the analysis assuming that their market values and their relocating costs are equal.

Making the obvious changes, we can define an extensive game with perfect information Γ_E from Γ by simply changing the outcomes after the movements of the chance c according to the assumption A_1. Analogously, we can define a Bayesian extensive game with observable action and chance movement $\Gamma_E(\pi_h)$ based on $\Gamma(\pi_h)$. The arguments required to analyze the set of equilibria of Γ_E and $\Gamma_E(\pi)$ are, mutatis mutandis, the same used in sections 3.1 and 3.3. Therefore we will simply summarize, without proof, the principal results of the equilibria analysis of these two games in the following proposition.

Proposition 13. *1. In every subgame perfect equilibria of Γ_E the agreement between Player 1 and Player 2_l is reached in period 0, and all of the possible outcomes of these agreements are summarized in Table 1.3.*

α	Player 1's outcome	Player 2_l's outcome
$\left(1-\frac{S_l}{S_h},1\right)$	$V-S_l$	S_l-l
$\left(1-\left[\frac{V+T-S_h}{V+T-S_l}\right]^{\frac{1}{n+1}},1-\left[\frac{S_l}{S_h}\right]^{\frac{1}{n}}\right)$, n odd	$V-S_h$	S_h-l
$\left(1-\left[\frac{S_l}{S_h}\right]^{\frac{1}{n+1}},1-\left[\frac{V+T-S_h}{V+T-S_l}\right]^{\frac{1}{n}}\right)$, n even	$V-S_l$	S_l-l

Table 1.3: Payoffs of the game Γ_E.

2. *Only the two types of perfect Bayesian equilibria introduced in Corollary 12 can hold for the game $\Gamma_E(\pi_h)$ and, in these cases, they are classified according to Table 1.4.*

Condition on α	**Type of equilibrium**
$Q_0(\alpha) < V-S_h$	Pooling
$Q_{n+1}(\alpha) < V-S_h$ $\alpha < 1-\left[\frac{S_l}{S_h}\right]^{\frac{1}{n}}$ n odd	Pooling
$Q_n(\alpha) > V-S_h$ $\alpha > 1-\left[\frac{S_l}{S_h}\right]^{\frac{1}{n+1}}$ n even or 0	Separating

Table 1.4: Classification of perfect Bayesian equilibria of $\Gamma_E(\pi_h)$.

With $Q_n(\alpha) := (1-\alpha)^n(1-\pi_h)(V-S_l)+(1-\alpha)^{n+1}\pi_h(V+T-S_h)$.

In this case, the agents achieve the equilibria by means of the same basic rationality analyzed in Proposition 2 and Proposition 10. However, it is worth noting some implications of the introduction of the assumption A_1. In the perfect information case, the expropriation threat enlarges the size of the interval considered in the first row of Table 1.3, implying that, in this game, the threshold for α for which Player 2_l is willing to accept the low subsidy S_l without further negotiations is lowered. On the other hand, an interesting trade-off arises given the two different costs associated with the implementation of the government's political pressure threats. Explicitly, given that $\frac{V+T-S_h}{V+T-S_h}$

is increasing in T and that $\frac{S_l - l}{S_h - l}$ is decreasing in l, both players have fewer incentives to continue the game for many periods, implying that the player with more resistance to the new costs of the breakdown can increase his chances to induce his most preferred payoff.

In the presence of information asymmetries, contrary to the previous case, the expropriation threat improves the chances of 2_l getting a high subsidy when α is big. In short, this proves that when the government is imperfectly informed and faces high political pressure to achieve its policy goals, the implementation of a more aggressive alternative strategy more easily leads to scenarios where small firms get high subsidies, i.e., it induces rent extraction. Also, the trade-off induced for the different costs associated with the breakdown situation is preserved in this situation.

1.4 Conclusions

In this paper, we analyzed some strategic implications of the implementation of a relocation policy for firms. Historically, the idea of firm relocation as political strategy has been abandoned because it has not provided the expected results. This failure is explained to some extent by the lack of an objective criterion to determine whether a firms relocation helps to reduce regional disparities or improve regional productivity. The availability of both new data and solid new theoretical approaches (e. g., the theory of Hausmann and Hidalgo) creates the possibility to use a new, better understanding of the actual and potential productive structure of the country to redirect new policies and re-implement old strategies. However, the welfare and social consequences found in our model demonstrate the existence of structural problems associated with the implementation of this strategy regardless of the nature of the types of firms to be relocated, which implies that the inefficiencies of the relocational strategy go beyond the lack of information.

We modeled a situation where the government and a firm bargain over a subsidy level to induce the relocation of the firm. We showed that for any alternative policy used by the government, if the political pressure to implement the policy is high and the government is asymmetrically informed about the typology of the firm, then a high subsidy is agreed upon independently of the firm's type. This first result lets us conclude that there is a rent extraction risk inherent to the implementation of this policy.

We also show that, in any case, the information asymmetries lead to only two types of perfect Bayesian equilibria. In addition to the equilibria already described, we find other equilibria where the government is capable of recognizing the firm's type and of giving it the corresponding subsidy. However, if the firm requires a high subsidy (is of the type 2_h), the equilibrium in this case will be reached after one period of delay, giving a positive probability to the breakdown of the bargaining between the firm and the government.

In brief, independently of the type of equilibrium attained in this game, there exists a positive probability that either the policy is not implemented due to a delay in reaching an agreement or the negotiations lead to a case where a firm can extract a higher subsidy than that required to change its location.

In the implementation of the alternative model, we found that when a government uses a more aggressive alternative that threatens the property rights of the firms, a high political pressure to carry out the relocation of the firms more easily induces a rent extraction situation. Our finding illustrates a new mechanism that relates the influence of a predatory or intrusive government to the rent-seeking behavior of the firms. The understanding of these types of mechanisms plays a central role in transitional economies such as China or Russia, where property rights are still not consolidated (Du and Mickiewicz, 2016).

The model proposed in this paper is quite simple and has several limitations that suggest possible extensions. One of these possible extensions, as a future work, would be a model of two-sided uncertainty where the government could be either aggressive or predatory, i.e., by maintaining the expropriation alternative, or passive, i.e., using alternative strategies like the one proposed in the first model. Another possible extension would consist of modeling the relocation bargaining while keeping in mind the possible temporal dynamics of the firm, i.e., giving incentives to the firm to relocate and to stay in the desired relocation. We expect that this latter extension would lead to (at least) the same strategic issues illustrated in this work.

Chapter 2

The p-innovation ecosystems problem

Abstract

In this paper [1], we propose a spatially constrained clustering problem belonging to the family of p-regions problems. Our formulation is motivated by the recent developments of economic complexity regarding the evolution of the economic output through key interactions among industries within economic regions. The objective of this model is to aggregate a set of geographic areas into a prescribed number of regions (so-called innovation ecosystems) such that the resulting regions preserve the most relevant interactions among industries. We formulate the p-innovation ecosystems model as a mixed-integer programming (MIP) problem and propose a heuristic solution approach. We explore a case involving the municipalities of Colombia to illustrate how such a model can be applied and used for policy and regional development.

2.1 Introduction

Since the seminal papers of Robert Solow, the mainstream economic growth analyses have used aggregate growth models (Solow, 1957; Solow, 1956) as their principal workhorses. Abstract quantities such as technological progress or positive externalities, such as spillovers or scale effects, usually play a central role in the classic growth theory since they are the primary mechanisms that encourage firm productivity and, therefore, generate economic growth. Nevertheless, given that the aggregative models address the interaction between different goods as a black box, they cannot endogenize the positive externalities that come from the strategic production of complementary goods, such as the scale effects or the spillover effects mentioned above (Hidalgo et al., 2007a).

Regardless of the approach (aggregative or not), the general consensus is that the diversification of the productive structure (i.e., selective innovation) stands out as a fundamental goal to foster economic growth. Thus, many theoretical efforts have focused on understanding and identifying suitable environments for innovation such as cities or industrial clusters (see e.g. Krugman, 1991b). Classic examples of such ideas within the context of urban economics are the so-called Jacobs'

[1] Joint work with Richard Church of UC Santa Barbara.

externalities (Jacobs, 2016) and the agglomeration and location externalities of the new economic geography (NEG). The approaches of Jacobs and the NEG intended to provide more detailed analyses of the positive effects of industrial interaction, advocating for the diversification of the production and the division of labor as concomitant processes that boost innovation opportunities.

Understanding the relatedness structure among economic activities thus becomes a key part of the analysis of the diversification and agglomeration of firms. Several studies following the pioneering ideas introduced by Hausmann and Hidalgo have found a suitable framework in network theory to understand the impacts of the productive structure of countries and regions regarding their diversification patterns (Hidalgo et al., 2007a; Neffke et al., 2011; Hausmann and Hidalgo, 2011; Hausmann et al., 2014). Hidalgo et al. (2007a) introduced the product space (PS), which is an undirected network that captures the relatedness structure of the world tradable goods market. This network revealed a core-periphery distribution, where the more complex goods ,i.e., the ones requiring more know-how, are highly connected whereas goods requiring less knowhow and skills are more likely to be the leaves of the graph. Complex goods are present in countries with a diverse export basket and according to Hidalgo et al. (2007a) and Hausmann and Hidalgo (2011) arise naturally through the interaction of existing industries. Therefore, enhancing the production of complex goods and fostering industrial interactions that lead to their emergence are natural policy goals to pursue. In fact, Hausmann et al. (2014) shows that the amount of complexity reached by a country is a strong predictor of its economic growth.

However, the geographic, institutional and historical factors such as the distances between the firms, the topography, administrative borders, differential taxes among states, productivity, labor availability and so forth are important constraints that may preclude the formation of optimal industrial clusters, i.e., the idiosyncratic factors of the territory can encourage the agglomeration of industries producing peripheral, or less valuable, industries in terms of the PS. This clustering process, although suboptimal, ends up reinforcing itself by the impact of the agglomeration economies that, despite all the possible drawbacks, such as congestion or elevated factor costs, leads the firms to locate close to each other (Diodato et al., 2018 and Rosenthal and Strange, 2004). Additionally, the empirical studies have shown that the NEG predictions (Martin and Sunley, 1996) are not robust and that the effects of Jacobs externalities are (at most) inconclusive (De Groot et al., 2016). It can be argued that the lack of conclusive empirical validation of the aforementioned theories is partially explained by the trade-off between diversity and similarity (see Neffke et al., 2011); labor, knowledge, and knowhow diffuse easily among similar firms, but only the interactions between firms that are sufficiently distinct can induce the generation of new ideas. More precisely, this idea (taken from cognitive theory) explains how a nondirected diversification can either lead to big dissimilarities among firms (cognitive distance), which hamper interactions or lead to an excessive overlapping of skills, which amounts to cognitive lock-in (Neffke et al., 2011). In the case of regions, this trade-off can be put in terms of the capacity of a region to incorporate technologies or knowledge from other region (absorptive capacity) (Jung and López-Bazo, 2017b). This concept has been found to be useful to explain disparities in the productivity levels among economic regions and, analogously, to explain the cognitive lock-in among firms. It has been noticed that some types of poverty traps are generated when regions with low absorptive capacity cluster together. This obeys the fact that the benefits of the investments directed to accumulate knowledge within the region are not maximized due to the lack of the absorptive capacity of the surrounding regions (see e.g. Caragliu and Nijkamp, 2016).

In summary, we face a setting in which (1) there are (quantifiable) target interactions among firms that enhance the innovation processes, (2) these key interactions could be hindered by or neglected due to incompatible local industrial policies that are inherent to the regional divisions, (3) these regional divisions are structural and, in most cases, unmodifiable, and (4) it is not sustainable or impossible to relocate firms from one region to another. With this scenario in mind, in this article we propose a quantitative model, the p-innovation ecosystems (p-IE) model, that involves the following three main goals: first, to identify the most relevant interactions among the existing industries in the country; second, to identify feasible sets of target goods that can improve the economic complexity of the country; and third, to find optimal regional configurations (formed by spatially contiguous administrative units) that maximize the most relevant interactions among the key industries (according to the previous two elements of the model) within each region. Thus, the p-IE model is intended to enable policymakers to apply suitable regional industrial policies that foster innovation processes according to the productive structure of each region. Roughly speaking, in terms of the literature of economic complexity, the p-IE model seeks to design regions that improve the local cohesiveness of the regional productive structure according to its position in the PS and, simultaneously, looks for an optimal diffusion by branching into the most complex neighboring industries in the PS.

The p-IE model is formulated as a mixed-integer programming (MIP) model that belongs to the family of the p-regions models devised by Duque et al. (2011). In brief, the p-IE involves the aggregation of n areas into p spatially contiguous regions that (1) maximizes the number of links (l,m) in the PS, in which industries l and m are within the same region; and (2) identifies strategic relationships between the industries within the regions that maximizes the probability of activating innovation processes that allow the region to "jump" to more complex neighboring nodes in the PS. Since the p-IE model is related to a family of problems that are classified as computationally nondeterministic polynomial-time hard (NP-hard) (Cliff et al., 1975; Keane, 1975) problems, we propose a heuristic solution to effectively compute a near-optimal, if not optimal solution.

From a practical point of view, this approach offers a new way to study the role of regions within countries (e.g., the states in the United States, the departments in Colombia, autonomous communities in Spain, etc.) as either boosters or as obstructions for the key industrial interactions that lead to the improvement of the complexity of the country. More precisely, the existence of regions may affect the evolution of the productive structure in the following two ways: on the one hand, regional borders between two industries may impair their interaction (as nodes in the PS) diminishing the positive externalities predicted from classic economic theory; on the other hand, the incentives generated by the local administrations could shape the innovation processes of different regions in rather distinct ways enhancing some industrial interactions and discouraging others.

The rest of the paper is presented as follows: Section 3.2 contains a literature review with the main conceptual elements upon which our model is built. Section 2.3 contains a step-by-step construction of the model, the exact formulation, and some illustrative examples. Section 2.4 presents the heuristic method to solve large instances of this model. Section 2.5 presents a case study for Colombia. Finally, Section 2.6 presents our conclusions and possible future research lines.

2.2 Literature review

The main ideas of this paper are built upon the following three theoretical bases: (1) the complexity theory of Hausmann and Hidalgo, (2) evolutionary economics, and (3) the family of p-regions models. This section is primarly devoted to discussing the features of these three theories that are relevant for our work.

In Hausmann and Hidalgo (2011), Hausmann and Hidalgo present their theory based upon the following two main principles: (1) products are the combination of capabilities (of all sorts: inputs, technology, know-how, institutions), and (2) a country can produce a product if and only if it has all the required capabilities to produce the product. Since capabilities are not observable they are determined a posteriori based on the following two main consequences of their assumptions: (a) the diversity of the products manufactured by a country is related to the diversity and number of capabilities in the country, thus, more capabilities lead to more products; (b) the ubiquity of a product (how many countries produce that product) is a measure of the number of capabilities required by the product.

Although capabilities are quite reminiscent of a general input in the standard production theory in economics, it is important to point out that Hausmann and Hidalgo use them in a more instrumental way. That is, the function that turns capabilities into products is binary (a characteristic function), i.e., it only determines whether the product is produced or not given the availability of the required capabilities within the country (see Hausmann and Hidalgo, 2011). Their approach in Hausmann and Hidalgo, 2011 is also blind with respect to the intensity of the factors, i.e., it only captures the requirement of a factor and not the amount or the way in which the factor is required for production. Hence, their assignation function that relates capabilities with products cannot be thought of as a production function but rather as a vector of characteristics of each existent good in the global economy (more related to the international trade theory of Heckscher-Ohlin-Vanek, (see Hausmann et al., 2019).

The practical work of Hausmann and Hidalgo is divided into two main blocks,i.e., complexity theory and relatedness theory, both of which are underpinned by the concept of capabilities. The concept of economic complexity is derived from a recursive analysis on the duality between diversity (number of exports of a country) and ubiquity (number of countries exporting a product). Diversity per se is not a good measure of the amount of capabilities a country has, since the types of products exported among countries may have substantially different capability requirements depending on the productive characteristics of the exporter. The number of capabilities required by any product is correlated with the scarcity of the product, i.e., if a product requires many rare capabilities only a few countries would be able to produce it. Thus, the interplay of these two concepts, i.e., diversity and ubiquity, can determine (as a posterior measure) the amount and the type of existing capabilities within the countries and required by the products. Additionally, this method helps to determine the degree of complexity embodied in a product. The exact recursive procedure proposed in Hausmann et al. (2014) will be further discussed in Section 2.3.1

It is important to note that the economic complexity of a country is determined by the complexity of the products produced by the country. With this in mind, and according to the findings in Hausmann et al. (2014), complex products can be regarded as the target products that may be reached (produced) by a country to improve its overall complexity and, therefore, its economic growth. The question of how these complex goods can be reached by specific countries is addressed from the relatedness theory that corresponds to the dynamic component of Hausmann and Hidalgo's theory. The dynamic component of Hausmann and Hidalgo's theory appears in Hidalgo et al. (2007a) when they seek to understand how countries diversify over time (which, in their terms, is equivalent to accumulating capabilities). They do this with using network theory (the PS): where the nodes are products and the strength of the link is determined by the number of capabilities shared by them, i.e., the link between two products is high if they require a similar set of capabilities to be produced. As such, measuring the relatedness between two goods requires using an observable quantity that captures the capabilities overlapping among goods. However, there is still not a clear consensus on which is the optimal way to measure the relatedness among a set of products. In fact, the study of the structure of relatedness of products within countries and geographic units is still an open field of research in economic geography, mostly in evolutionary economic theory (see Frenken and Boschma, 2007; Boschma, 2017).

For the sake of brevity and since it is the only one to be used in our applications; we consider only the PS approach proposed in Hidalgo et al. (2007a). Similar constructions of networks that measure the structure of the relatedness of geographic areas can be found in Neffke et al. (2011), O'Clery et al. (2019), and Diodato et al. (2018). All these approaches are designed so that they measure the cooccurrence (coproduction, coexportation) of goods within the same geographic unit weighted by some quantity (e.g. input requirements, total output) that reflects the extent to which specific economic activities are carried out inside each geographic unit. Explicitly, in Hidalgo et al. (2007a) a more agnostic or outcome-based approach is proposed. Using world trade data, they measure the relatedness using the Revealed Comparative Advantage (RCA) index (Balassa, 1964, see also Section 3.1). Based on this index, the PS is defined as the network of available goods in the world economy with links weighted by the probability of having a comparative advantage among each pair of goods (see Hidalgo et al., 2007a). Thus, if the weighted links that define the network are regarded as probabilities to "jump' between the nodes that join them, then the PS not only describes the present output of the country but the actual constraints that the country faces to diversify towards other goods in the network. In fact, if we think of the position of a country on the PS as the set of products in which it has an RCA greater or equal to one, then the innovation dynamics of the countries can be analyzed as diffusion processes on the PS (Hidalgo et al., 2007a; Alshamsi et al., 2018; O'Clery et al., 2018).

Based on their theoretical framework, moving in the PS requires acquiring missing capabilities that are combined with the existing ones, which allows a country to produce a new product and, therefore, move to more complex products in the PS. Thus, the interaction between industries with similar (but not equal) sets of capabilities plays an important role in the acquisition of the new capabilities required to produce complex products and move towards better positions in the PS.

Together, the complexity embodied in the products and the relatedness structure of the economic output of the countries provides a tractable framework for longstanding ideas of evolutionary economic theory. Theories such as the process of creative destruction proposed by Schumpeter in Schumpeter et al. (1939) can be quantitatively underpinned by the path-dependency inherent to the dynamics on the PS. More explicitly, the evolution led by the natural forces of the economy or the diffusion of the RCA, in terms of the PS, has two main implications, as follows: the disappearance of old industries poorly connected in the PS with the local industry and the appearance of new industries naturally associated, according to the PS, with the local industry. This issue regarding the selective evolution of the economic output has been studied empirically in the context of countries (Hidalgo et al., 2007a; Hausmann and Hidalgo, 2011; Hausmann et al., 2014), economic regions (Hausmann and Neffke, 2019; Neffke et al., 2011) and cities (O'Clery et al., 2019; Hausmann et al., 2019) and shows that this phenomenon is transversal to any geographic unit in which industry clustering is admissible.

A remarkable empirical work that sheds light on how the process of creative destruction takes place in a geographic (regional) framework is presented in Neffke et al. (2011). In this work, the authors study the evolution of the economic output within 70 industrial regions in Sweden along a period of approximately 30 years (1969-2002) and; work, they measure the evolution of the economic cohesion of the Swedish regions over time. As a measure of cohesion, they use the average of the closeness (in the PS) of each product to the existing portfolio in the region. Intuitively a region is cohesive if there are no incentives for moving, i.e., if the industries within the region are more related each other rather than with industries belonging to other regions. In Neffke et al. (2011), the authors find that the cohesiveness of the regions has a significant tendency to increase through the years following a process or creative destruction. This work provides a solid confirmation of an idea already hinted by many other studies in complexity theory, as follows (Bustos et al., 2012; O'Clery et al., 2019: Diodato et al., 2018): there is a tendency of the geographical clusters (cities, regions, metropolitan areas) to reproduce the industrial clusters suggested by the PS. In principle, these two types of clusters (geographic and industrial) may disagree because of historical, geographic, cultural or political reasons that encouraged certain types of firms to establish in the regions, thus generating an initial endowment of productive activities that shaped the evolution of the economic landscape of the region in the subsequent history. Additionally, for methodological reasons, it is worth recalling that by construction the PS, at least the one built upon world trade data, captures industrial clusters independently from the spatial features of countries and regions.

Another important implication that can be inferred from the works previously discussed is that certain policies that seek to boost local economies by restructuring the economic landscape (such as firm relocation or selective innovation) may be misleading and may have structural issues due to the lack of closeness to the existing portfolio of the target region. In terms of the work of Neffke et al. (2011), for example, the introduction of economic activities weakly linked to existing industries of the region have a higher conditional probability to fail and exit the region.

In light of the contributions of complexity theory and evolutionary economics it seems plausible to take advantage of the information encoded in the PS to foster a selective improvement of the productivity (or, more precisely, complexity) as long as the strategy to follow harmonizes with the natural constraints imposed by the topology of the PS. In this order of ideas, and following the arguments presented in the introduction, we opted, in this paper, to use a strategy that involves the design of industrial regions that fosters the creation and development of industrial ecosystems. These regions are intended to be spaces in which the interactions among industries are fostered by suitable public policies so that the capabilities required to create new and more complex goods appear naturally according to the dynamics of the PS. This is, the so-called industrial or innovation ecosystems (see also O'Clery et al., 2018).

There is a wide and broad literature about regional planning and region designing (see e.g. Fischer, 1980; Duque et al., 2007); this area has focused on providing objective criteria to define a suitable regionalization that will help in the implementation of specific public policies. The main idea of this approach consists of decomposing the global policy objectives that are applied into large geographic units into local tasks in specific regions that integrates in a harmonic way and, finally, to promote the achievement of the policy goal (Fischer, 1980).

Among the formulations for this problem are linear optimization models (Glover, 1977; Zoltners and Sinha, 1983), mixed-integer programming (Duque et al., 2011) and nonlinear models (Macmillan and Pierce, 1994), and the use of techniques such as the implicit enumeration of feasible solutions (Glover, 1989) and column generation (Mehrotra et al., 1998). In particular, our model follows the lines of Duque et al. (2011) which is formulated a mixed integer programming (MIP) problem, which is called the p-regions problem, for aggregating n areas into p spatially contiguous regions, using three different strategies to ensure continuity. One of the main advantages of the p-regions type models is their flexibility to incorporate spatial contiguity without adding constraints on the shape of the resulting regions. This, in turn, allows them to capture a diverse range of spatial patter (compact, elongated, among others). For example, based on the p-regions model there have been different MIP formulations created by adding additional features, such as maximizing the number of regions being added (Duque et al., 2012), generating p regions while maximizing their compactness (Li et al., 2014), defining p-functional regions (Kim et al., 2015), among others. Such formulations ensure that we can obtain an optimal solution. However, they are computationally intensive to solve, and it is sometimes better to use heuristic approaches for resolving large problem instances. A good heuristic should be capable of solving a problem of region delineation, within a shorter amount of computation time, with the goal of identifying feasible solutions as close as possible to the optimum (Duque et al., 2007).

In this work, we take advantage of the versatility and robustness of the economic complexity theory of Hausmann and Hidalgo and embed it into a region designing model. Thus, we propose drawing a conceptual and theoretical link between these two subjects, thus providing a new model with a solid economic foundation to the p-regions family, which is based on Hausmann and Hidalgo's theory and evolutionary economics. This also allows us to incorporate new elements into Hausmann and Hidalgo's theory such as the role of the regions in the innovation process followed by the countries in the PS. Additionally, a proved relationship between the regional division of a country and its economic growth will open the possibilities for developing new economic-based models to allow policymakers to: (1) identify how far the actual regional configuration of its economy is from the optimal one that would potentially foster economic growth, and (2) to design industrial regions, or intraregional agreements between administrative units, to minimize the boundary effects and ease the interaction among the key regions to promote the generation of potential innovation ecosystems.

2.3 The model

In this section we present the elements and assemble the objective function of the p-IE model step-by-step. For the sake of clarity, we divide this section in 4 parts. The first part provides an explanation of the inputs of the model, i.e., elements of the PS (economic) and spatial elements (geography); additionally, it shows how these elements are systematically incorporated into the objective function. The second part provides a summarized statement of the optimization problem. Part three contains the final expression that will be maximized. Finally, in part four we illustrate the properties of the solution of the p-IE model by means of some suitable toy examples.

2.3.1 Conceptual framework

The complexity theory, in this work, plays the role of giving a quantitative framework to the process of identifying key industries. There are two types of key industries, as follows: (1) the ones that already exist and whose interaction is central in the given economy, and (2) complex industries that are yet to appear in the economy of interest. Our measurement tool in this case is the PS, which describes the relatedness structure between all the goods in the world (or in another macrogeographic reference, e.g., a continent, a set of countries, etc.) and, therefore, provides information about the classes of key industries.

Mathematically, we will view the PS together with the complexity theory as a weighted network of products, where the weight in the nodes gives a measure of the relevance of each good (1 represents the most relevant good and 0 the least relevant), and the weight in the edges gives a normalized notion of the association between any pair of goods. It is important to stress the fact that even if we consider a weighted network built upon Hausmann and Hidalgo's theory (Hausmann et al., 2014; Hidalgo et al., 2007a), any network with the properties specified above would work for our purposes.

First term: relatedness structure and key industrial links

The PS proposed in (Hausmann et al., 2014) measures the relatedness by means of the Revealed Comparative Advantage (RCA) index introduced by Balassa (Balassa, 1964) which is given by the following

$$\mathbf{RCA}_{c,p} = \frac{x_{cp}/\sum_p x_{cp}}{\sum_c x_{cp}/\sum_{c,p} x_{cp}}$$

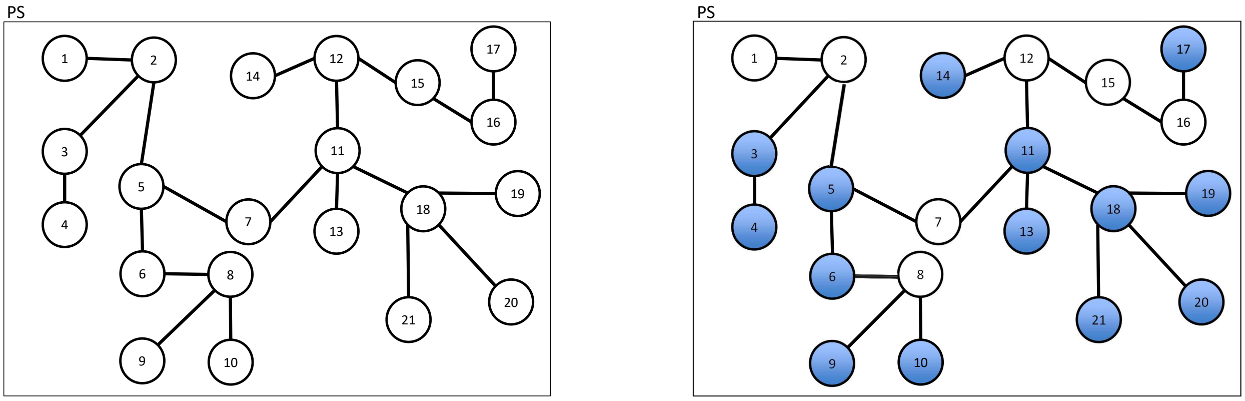
Where x_{cp} is the amount of the good p exported by the country c and the indices of the sums include all the possible goods a countries. Given these indices for each good and each country, Hausmann et al. (2014) defines the degree of relatedness between goods k and l as follows:

$$y_{lm} = \min\{\mathbb{P}(\mathbf{RCA}_{m,p} > 1|\mathbf{RCA}_{l,p} > 1), \mathbb{P}(\mathbf{RCA}_{l,p} > 1|\mathbf{RCA}_{m,p} > 1)\} \tag{2.1}$$

That is, the minimum of the conditional probability of obtaining RCA in a good given that the country already has RCA in the other good (the minimum is taken so that the weight is undirected).

Hence, the PS is defined as the undirected network (see Figure 2.1a) of available goods in the world economy, where the links are weighted by the numbers y_{lm} whenever these weights exceed a certain threshold (see Hausmann et al., 2014, or Section 2.5 for a brief discussion regarding the determination of the threshold).

Following the ideas of Hausmann et al. (2014) we can assess the overall production of a country from its position on the PS. The idea includes determining the set of goods in which the country has revealed a comparative advantage, i.e., an RCA greater than 1. We regard this set as the initial position of the country on the graph (on the PS) and we will denote it by IG (Initial goods), analogously we denote by IL the set of initial links induced in the PS by IG, i.e., the links of the PS connecting the nodes belonging to IG (see Figure 2.1b).

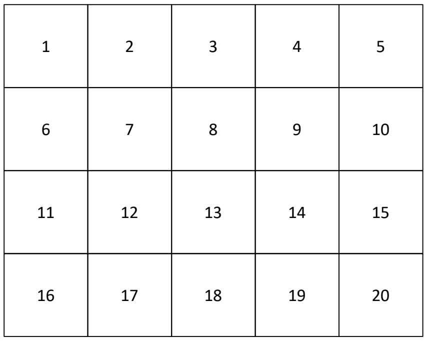


(a) Layout of the PS of an abstract economy with 21 goods.

(b) Available (blue) goods in the economy IG determine the position of the country on the PS.

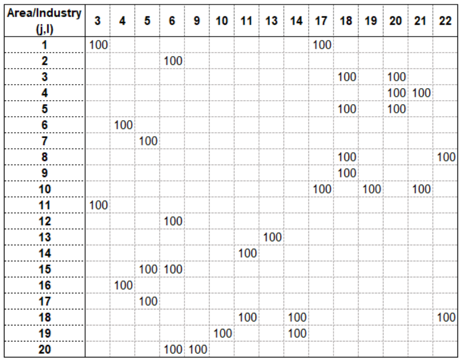
Figure 2.1: The first part shows the PS faced by the economy, whilst the second part shows in blue the goods in which the economy has comparative advantage.

We introduce geography in this model by means of the following two effects: the boundary effect and distance decay. Since production and, subsequently, the spillover effects measured by Hausmann and Hidalgo's theory take place in very concrete spatial units, we incorporate the boundary effect by measuring the interactions between industries that have an active presence within the same region. These measures are derived from two spatial inputs: the division of the country (or the general spatial framework of reference) into areas (see Figure 2.2a) and the frequency matrix $f = (f_{li})$ of the industries among the areas (see Figure 2.2b) (value of exports in this case, since we follow Hausmann et al., 2014).



1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20

(a) Country with 20 areas.



Area/Industry (j,l)	3	4	5	6	9	10	11	13	14	17	18	19	20	21	22
1	100									100					
2				100											
3											100		100		
4													100	100	
5											100		100		
6		100													
7			100												
8											100				100
9											100				
10										100		100		100	
11	100														
12				100											
13								100							
14							100								
15			100	100											
16		100													
17			100												
18							100		100						100
19						100			100						
20				100	100										

(b) Distribution matrix $f = (f_{lj})$ for an economy with 22 industries ($l = 1, \cdots, 22$) and 20 areas ($j = 1, \cdots, 20$).

Figure 2.2: The first part depicts the geographic distribution of the 20 areas in which each area is given by a square of side length 1. The second part shows the distribution of 22 industries in the country. Notice that we omit the columns corresponding to industries 1 and 2 since the country does not produce any of them.

From the geographic distribution of the firms we compute the relative relevance of area i producing good l (that is the share of an area in the total production, computed from Figure 2.2b), as follows

$$r_{li} = \frac{f_{il}}{\sum_i f_{il}} \tag{2.2}$$

We also incorporate a distance decay factor that penalizes the interactions of industries located far from each other. This geographic effect is justified theoretically in Hausmann et al. (2019) by the strong effect of distance in the, arguably, more important capability, i.e., knowledge. We propose a decay function $f : [0, \infty) \mapsto [-1, 1]$ of the form:

$$f(x) = 1 - 2\left(\frac{x}{M}\right)^{\alpha}, \tag{2.3}$$

where M is the maximal distance between two areas in the country and α is the distance decay parameter. Notice that $f(0) = 1$ and that $f(M) = -1$. The parameter α determines how fast the decay function becomes negative, decaying faster when α is greater than 1 and slower whenever α approaches 0 (see Section 2.3.4 and Section 2.5 for general rules to establish the value of α). Thus, we capture the interaction between the goods l and m produced in the areas i and j, respectively, using the following expression

$$t_{ij}^{lm,k} r_{li} r_{mj} f(d_{ij}), \tag{2.4}$$

where d_{ij} is the distance between area i and j, $t_{ij}^{lm,k}$ is an indicator function that is 1 whenever i and j produce goods l and m, respectively, and both belong to the same region k (see Figure 2.3). Hence, in a setting with n areas and $p < n$ regions, the first term of our objective function would have the following form

$$FT(S_p) = \sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=i}^{n} \sum_{(l,m)\in LI} t_{ij}^{lm,k} r_{li} r_{mj} f(d_{ij}) \tag{2.5}$$

where S_p is any partition of the n areas into p regions which in turn determine the value of the variables $t_{ij}^{lm,k}$.

We stress the relevance of two features in the structure of the first term of the objective. First, notice that we require the inverse function of the distance f to assign negative values to penalize long distances with negative values (see Section 2.3.4 for a further discussion). Second, we capture the interaction among areas in a multiplicative way to neglect all the cases in which at least one of the intensities is 0. This idea and the role of the indicator functions $t_{ij}^{lm,k}$ are fully explained in Figure 2.3. We describe five possible scenarios of interaction as follows: in the first one, both, areas i and j produce two industries linked in the PS (l and m), yielding 2 self-interactions within each area ($r_{li}r_{mi}$ and $r_{lj}r_{mj}$) and two unidirectional interactions between industries l and m across areas ($r_{li}r_{mj}$ and $r_{mi}r_{lj}$); the second scenario exhibits one loop interaction within area i between industries l and m ($r_{li}r_{mi}$) and the interaction between industry l in area j and industry m in area i ($r_{mi}r_{lj}$); the third scenario presents one unidirectional interaction; scenarios four and five show that neither the presence of only one industry in one region nor the presence of the same industry in several areas yields any interaction.

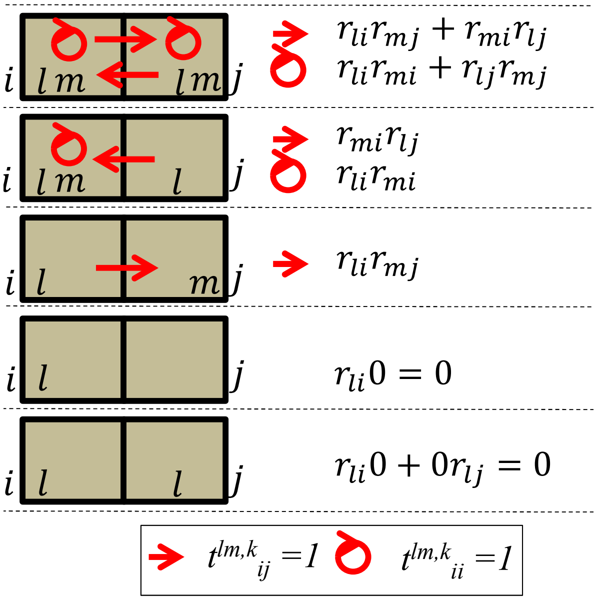
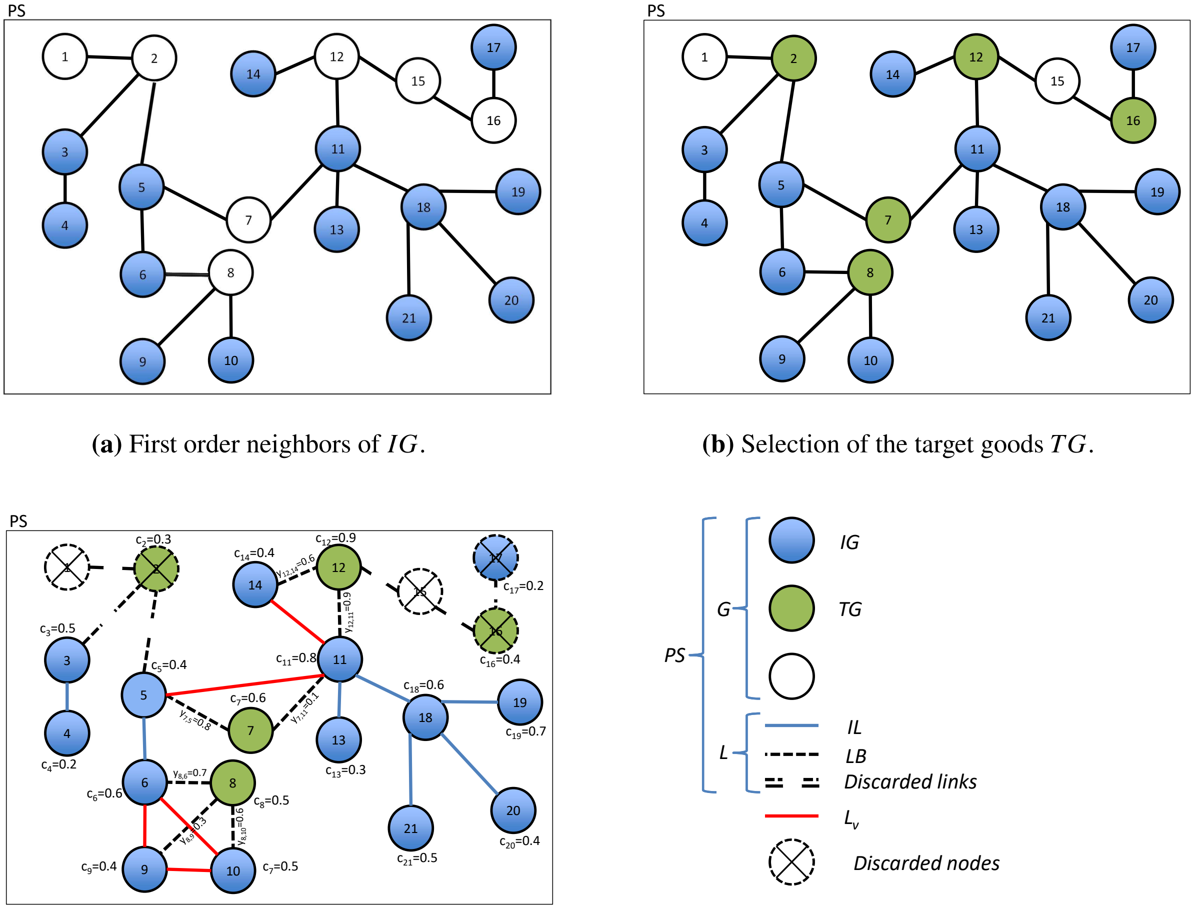


Figure 2.3: Fixing a region k, the indicator functions $t_{ij}^{lm,k}$ capture the self-interactions within areas (loops) and the interactions between pairs of areas (i, j)) (arrows) producing linked goods (l, m) in the PS.

Second term: complexity and target products

While the first component enhances the interaction among the goods that already exist in the economy, the second component is designed to maximize the probability to reach new products in the PS that are not yet produced by the country. The construction of this second part relies in two basic principles, as follows: (1) nodes with larger weights (more complex) are more desirable for the economy and (2) the pairwise interaction of products with high association to a third nonexistent good in the economy increases the probability of this new good arising whenever the interaction is given within the same region. More generally, the problem addressed by the second term of the objective functions lies in the general framework of a directed complex contagion problem, i.e., a diffusion process in the PS in which the process can diffuse from several nodes to others depending on their complexity. Such models predict a nonlinear diffusion, more explicitly, the probability of reaching a new node follows a power law distribution increasing in the number of first order neighbors of the node (see Alshamsi et al., 2018). Nonetheless, in Alshamsi et al. (2018), the contagion power for the PS was found to be conveniently close to 1 (1.03, more precisely), enabling us to consider the sum of the pairwise interactions. Notice also that the study of contagion as a pairwise interaction of nodes is quite reminiscent to the basic concept of clustering by triplets on networks (see e.g. the definition of the clustering coefficient in Barabási et al., 2016). For our subsequent analysis we introduce the notation L_v for the set of (undirected links) of neighbors of v in the PS. Notice that two nodes linked in L_v may not be (but are usually) linked in the PS.

Since the nonexistent goods may be far away from the set IG in the PS, we restrict our analysis to products that can be reached after one step of diffusion, in an effort to simplify the study of more complex dynamics. Additionally, since this component of the objective function is intended to improve the complexity, then we define the target goods as the ones that are more complex than at least one of their neighbors in the PS already produced by the country. Further, the spatial nature of this model allows us to only consider the complex goods whose arising requires the interaction of two or more industries, i.e., we discard the leaves of the network. We thoroughly depict these processes in Figure 2.4a. This set of nodes in the graph constitutes the set of target goods denoted by TG. Although many complexity measures would work for our purposes, we want to stress that its introduction in this model is motivated by the definition of the Economic Complexity Index (ECI) introduced in Hausmann et al. (2014). In our case of study (see Section 2.5), we use a normalized version in the unit interval of this index as a measure of the complexity.



(a) First order neighbors of IG.

(b) Selection of the target goods TG.

(c) Final input network.

(d) Nomenclature.

Figure 2.4: The first three parts depict how the target nodes (TG) are chosen. In Figure 2.4a we select the first order neighbors of the blue nodes IG. In Figure 2.4b we discard the higher order neighbors (nodes 1 and 15), the leaves of the network that do not require spatial interaction to be produced (node 16), and the green nodes that do not improve the complexity of the system (node 2). Figure 2.4c depicts the final version of the PS and Figure 2.4d summarizes the nomenclature used for for this selection process.

Thus, arguing as in the formulation of the first part we have that the second component should have the form

$$ST(S_p) = \sum_{k=1}^{p} \sum_{v \in TG} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{(l,m) \in L_v} t_{ij}^{lm,k} (y_{lv} r_{li} + y_{mv} r_{mj}) f(d_{ij}) c_v \tag{2.6}$$

where c_v is the complexity of the good $v \in TG$, and y_{lv} is the weight of the link between the goods l and v in the PS (equation (2.1)).

Assembling the objective function

So far, we have set up both components of our objective function without specifying the interplay between them. By construction, the first term enables us to capitalize on the key industrial interactions already formed by the natural economic forces of the country of interest. Since the second term tries to reach as many complex goods as possible it has the tendency to group multiple areas and create large regions while introducing volatility into the model. Thus, the first term plays the role of guaranteeing a reasonable size and shape of the economic regions so that key interactions are preserved, whereas the second term is intended to "perturb" or modify the configurations suggested by the first term to maximize the complexity without making abrupt changes to the regional configurations. Hence, the next step consists of finding suitable weights to guarantee the predominance of the first term over the second one.

Notice that the terms in equations (2.5) and (2.6) are bounded from above by 2. Hence, to bound this term, it is necessary to determine the maximum number of ways to assign n areas into p regions.

This problem can be phrased in terms of a maximization problem. Let z_i be the number of areas in region i. From Figure 2.3 we have that each area can interact with all the areas, including itself. Therefore, if we count the number of interactions we have that the first area can interact with z_i areas, the second with $z_i - 1$ and so forth. Thus, we face the following:

$$\begin{aligned} \text{Maximize} \, \frac{1}{2} \sum_{i=1}^{p} z_i (Z_i - 1), \\ \text{s.t.} \sum_{i=1}^{p} z_i = p, \\ z_i \geq 1. \end{aligned}$$

Here, the objective function measures the number of interactions between all the areas in each region, i.e., the number of $t_{ij}^{lm,k}$ that can be 1 assuming that all goods are produced in all the areas and that two areas can be joined in only one region (where this latter restriction will be guaranteed by the constraints of the model). A direct computation shows us that the solution of this maximization problem is $\frac{n}{2p(n-p)}$. Thus, the first component of the objective function is bounded from above by

$\frac{n}{2p(n-p)}|IL|$, where $|IL|$ is the number of elements in IL. A similar reasoning applied to the second term shows that this term is bounded from above by $h = \frac{n}{2p}(n-p)\sum_{v\in TG}|L_v|(|L_v|-1)$.

Hence, a possible factor that guarantees the dominance of the first term of the objective function over the second term would be $W = 10^{1+\lfloor \log(h)\rfloor}$ (where $\lfloor\cdot\rfloor$ is the standard floor function.) We can also normalize the weights by setting $W_1 = \frac{W}{W+1}$ and $W_2 = 1 - W_1$. Finally, given a p-regionalization S_p (or a partition of n areas into p regions) we define the objective function of the p-IE as follows

$$Z(S_p) = W_1 FT(S_p) + W_2 ST(S_p) \tag{2.7}$$

2.3.2 Problem statement

Let $I = \{1,\cdots,n\}$ be the set of areas of a country with n areas and let $p \in \{2,\cdots,n-1\}$ be the prescribed number of regions. Let Π be the set of feasible p-regionalizations given by contiguous partitions of I, i.e., collections of areas the form $P_p = \{R_1,\cdots,R_p\}$ such that

- $R \neq \emptyset$ for $R \in P_p$
- $R \cap R' = \emptyset$ for any pair of distinct elements $R, R' \in P_p$.
- $\bigcup_{R\in P_p} R = I$.
- Each region $R \in P_p$ is connected.

The p-IE problem may be formulated as follows:

$$\text{Determine } P_p^* \in \Pi \text{ such that: } Z(P_p^*) \geq Z(P_p),\ \forall P_p \in \Pi.$$

2.3.3 The exact formulation of the p-Innovation Ecosystems problem

Now we are in position to formulate the exact model.

Parameters:

** The geography*

$A =$ set of areas, $A = \{1,\cdots,n\}$;
$i, j =$ the indices used to refer to specific areas, where $i, j \in A$;
$n_{ij} = \begin{cases} 1, \text{ if areas } i \text{ and } j \text{ share a border, with } i \neq j; \\ 0, \text{ otherwise;} \end{cases}$
$N_i =$ $\{j \mid n_{ij} = 1\}$, the set of areas that are adjacent to area i;
$d_{ij} =$ the distance between areas i and j, where $i < j$;
$f =$ a decreasing function in the distance between areas i and j, whose range lies on $[-1, 1]$;
$k, K =$ the index and set of regions, $K = \{1,\cdots,p\}$;
$o =$ the index used to refer to the contiguity order, $o \in \{0,\cdots,q\}$, with $q := n - p$.

** The Product Space*

$PS =$ any instance of the Product Space (PS);

$g, G =$ the index and set of goods/nodes in the PS;

$l, m, v =$ the indices used to refer to specific goods, where $l, m, v \in G$;

$c_v =$ the complexity of good v ;

$L, \{l, m\} =$ the set L in which each element (l, m) indicates a link connecting goods $l \in G$ and $m \in G$ in the PS;

$y_{lm} =$ the weight in the PS between the goods l and m;

$IG =$ set of available goods for the country under study, which are connected in the PS given certain threshold in the weights of the links;

$TG =$ set of goods/nodes $g \in G$ that the country under study defines as target goods. $GT \subseteq G$ (see equation 2.6). The set TG satisfies the following properties:
$TG \cap IG = \varnothing$;
$TG \cup IG \subseteq G$;
each good in TG is a first order neighbor of, at least, two goods in IG;
i.e., for each good $l \in TG$ there are links l, m in PS such that $m \in IG$;
each good v in TG, has a c_v that is greater than the minimal complexity of their neighboring goods in IG (there is an improvement in the complexity);

$IL =$ the set of links $\{l, m\}$, such that $l \in IG$ and $m \in IG$. $IL \subseteq L$;

$LB =$ the set of links $\{l, m\}$, such that $l \in IG$ and $m \in TG$. $LB \subseteq L$;

$L_v =$ the set of links (l, m) with the pairwise connections between the goods $l, m \in IG$ such that both are connected in the PS to $v \in TG$. These links are not of the same nature as the links in the PS.

** The relationship between the geography and the Product Space*

$r_{li} =$ The relative relevance of the area i producing the good l, see equation (2.2);

$$b_{li} = \begin{cases} 1, \text{ if } r_{l,i} > 0 \\ 0, \text{otherwise.} \end{cases}$$

** Instrumental parameters*

$h =$ $h = \frac{n}{2p}(n-p)\sum_{v \in TG} |L_v|(|L_v| - 1)$;

$W =$ $10^{1+\lfloor \log(h) \rfloor}$;

$W_1 =$ $\frac{1+W}{W}$;

$W_2 =$ $1 - W_1$.

Decision variables:

$$t_{ij}^{lm,k} = \begin{cases} 1, \text{ if areas } i \text{ and } j \text{ belong to region } k, \text{ area } i \text{ produces good } l \\ \quad \text{and area } j \text{ produces good } m, \\ \text{and 0 otherwise;} \end{cases}$$

$$x_i^{ko} = \begin{cases} 1, \text{ if areas } i \text{ is assigned to region } k \text{ in order } o, \\ \text{and 0 otherwise.} \end{cases}$$

The model can now be formulated as follows:

Maximize:

$$\begin{aligned} Z =& W_1 \sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=i}^{n} \sum_{\{l,m\} \in LI} t_{ij}^{lm,k} \left(r_{li} r_{mj} y_{lm} \right) f\left(d_{ij} \right) + \\ & W_2 \sum_{k=1}^{p} \sum_{v \in GT} \sum_{i=1}^{n} \sum_{j=i}^{n} \sum_{\{l,m\} \in L(L_v)} t_{ij}^{lm,k} \left(y_{lv} r_{li} + y_{mv} r_{mj} \right) c_v f\left(d_{ij} \right). \end{aligned} \tag{2.8}$$

Subject to:

$$\sum_{i=1}^{n} x_i^{k0} = 1 \qquad \forall k = 1, \cdots, p, \tag{2.9}$$

$$\sum_{k=1}^{p} \sum_{o=0}^{q} x_i^{ko} = 1 \qquad \forall i = 1, \cdots, n, \tag{2.10}$$

$$x_i^{ko} \leq \sum_{j \in N_i} x_j^{k(o-1)} \qquad \forall i = 1, \cdots, n; \forall k = 1, \cdots, p; \forall o = 1, \cdots, q, \tag{2.11}$$

$$2 t_{ij}^{lm,k} \leq \sum_{o=0}^{q} x_i^{ko} b_{li} + \sum_{o=0}^{q} x_j^{ko} b_{mj} \qquad \forall i, j = 1, \cdots, n; \forall k = 1, \cdots, p; \forall \{l, m\} \in LI \cup L_v, \tag{2.12}$$

$$t_{ij}^{lm,k} \geq \sum_{o=0}^{q} x_i^{ko} b_{li} + \sum_{o=0}^{q} x_j^{ko} b_{mj} - 1 \qquad \forall i, j = 1, \cdots, n; \forall k = 1, \cdots, p; \forall \{l, m\} \in LI \cup L_v; f(d_{ij}) < 0, \tag{2.13}$$

$$x_i^{ko} \in \{0, 1\} \qquad \forall i = 1, \cdots, n; \forall k = 1, \cdots, p; \forall o = 0, \cdots, q, \tag{2.14}$$

$$t_{ij}^{lm,k} \in \{0, 1\} \qquad \forall i, j = 1, \cdots, n; \forall k = 1, \cdots, p; \forall \{l, m\} \in LI \cup L_v. \tag{2.15}$$

Constraints (2.9), (2.10) and (2.11) guarantee the connectedness of each defined region. These conditions represent an extension of the ordered area assignment conditions proposed by Cova and Church (2000) (see also Martin and Sunley, 1996). More precisely, equation (2.9) forces the model to assign only one area per region with order 0, equation (2.10) implies that each area must be assigned with some order to some region, and equation (2.11) preserves the order of the assignment of the areas to the regions, i.e., guarantees that an area is assigned to a region with order o only if there is another neighboring area belonging to the same region assigned with order o-1.

Restrictions (2.12) and (2.13) provide the interaction between the regionalization (x_i^{ko}) and the interaction in the product space ($t_{ij}^{lm,k}$). Equation (2.13) bounds the decision variables $t_{ij}^{lm,k}$ so that the model is allowed to activate them (and increase the value of the OF) when the interaction between the goods l and m produced in the areas i and j (respectively) occurs within the same region $k \in \{1, \cdots, p\}$, i.e., this restriction generates the trade-off in the model that guarantees that only the most important links of the product space are preserved. On the other hand, we use equation (2.13) to force the model to activate all the variables $t_{ij}^{lm,k}$ whenever areas i and j are within the same region and their interaction is penalized by the distance decay function f so that the OF has to effectively pay for the generation of large regions. This restriction prevents the model from producing degenerate solutions, i.e., one massive region and $p - 1$ small regions. Controlling the size of the regions through the distance decay allows us to formulate the model for any geography without the risk of having empty feasible sets (infeasibilities). Finally, equations (2.14) and (2.15) are integer conditions on the decision variables.

2.3.4 Examples

We present two examples to illustrate how our model works. In the first example we show the behavior of the first term of the OF, whereas the second shows the influence of the second term on the solution, i.e., explicitly, how the second term of the OF breaks ties in a solution reached by the first term, seeking to maximize complexity. In both cases, we will take the Euclidean distance between the centers of the regions as distances between areas.

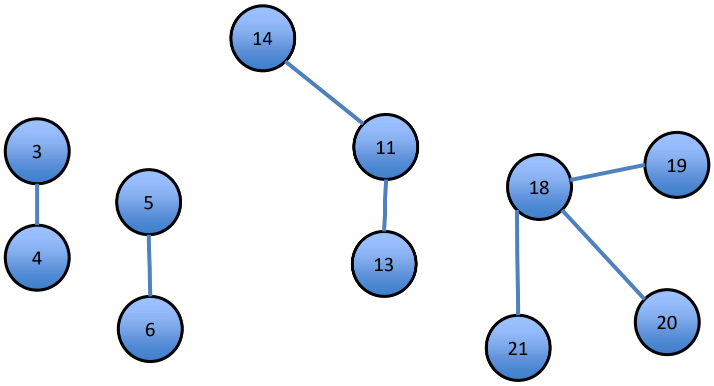
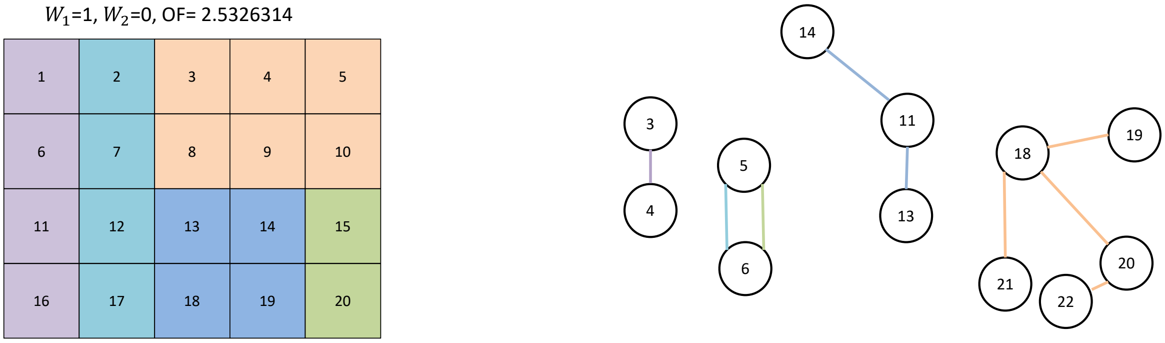


Figure 2.5: PS for example 1. This is the PS of the country described in Figure 2.2. It contains 4 clusters of industries perfectly separated and omits industries 9, 10 and 17 which are isolated.

We will take α, i.e., the decaying parameter of the function f in equation (2.3), such that it gives a value of zero to the radius of the region in the scenario in which all of them have the same number of areas $\frac{n}{p}$. Thus, since the geography in these toy examples is given by square areas in rectangular grids (see Figure 2.6b and Figure 2.7c) we will take $R_{av} = \sqrt{\frac{n}{p}}$ as the radius of an average region. Therefore, we have that $\alpha = \frac{ln(2)}{ln\left(\frac{M}{R_{av}}\right)}$.

For the first example we suppose a country with geography and production described in Figure 2.2 and, therefore, we have the decaying parameter $\alpha = 0.7564$. Consider also the PS depicted in Figure 2.5. Notice that the PS given in Figure 2.5 is a subset of the PS presented in the Figure 2.1, although some links are removed or added so that we obtained 4 disconnected clusters on the graph. Additionally, notice that in this case the graph is entirely exhausted by the set IG, equivalently, there are no target goods. Hence, the weights of the vertices are irrelevant for the objective function. From the conveniently designed distribution of industries on the array, there are five spatial clusters of industries where each one of corresponds to one of the four connected components of the PS in Figure 2.5. Thus, for p=5 the p-IE model finds an optimal solution of the form presented in Figure 2.6a.

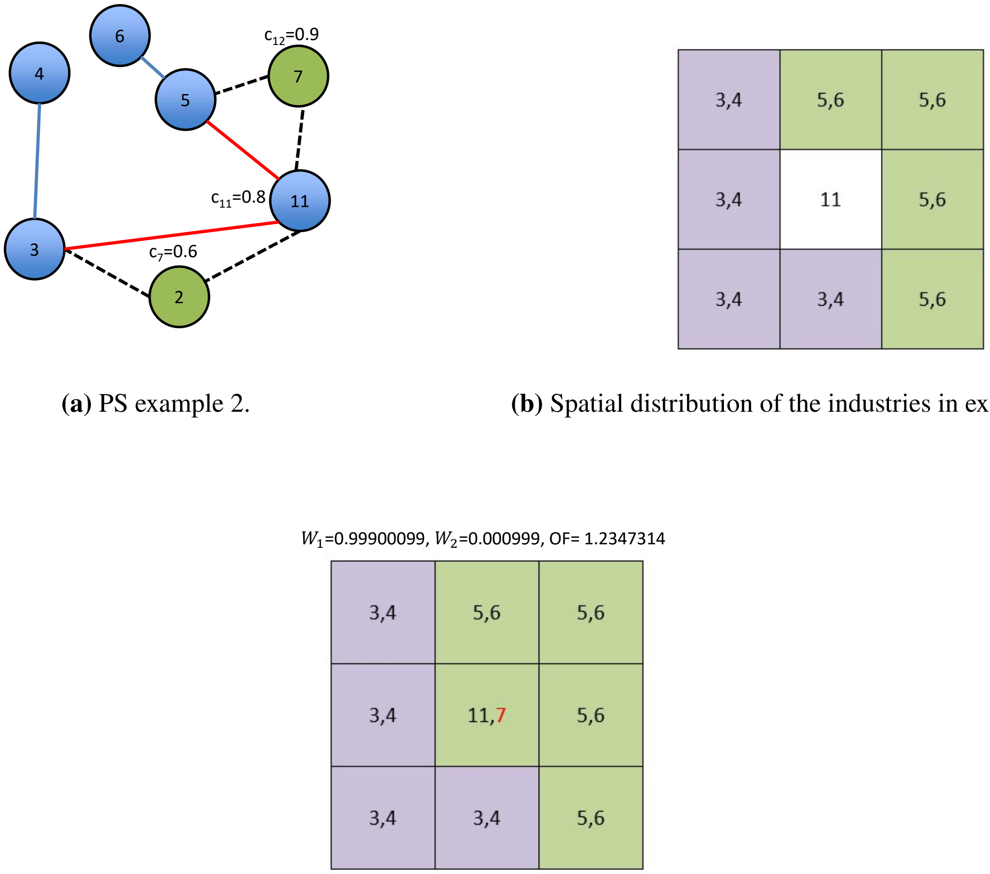


(a) Optimal region assignation.

(b) Industrial interaction per region.

Figure 2.6: This figure summarizes the output of the optimization for this example. The five shaded areas in the first part correspond to the optimal regions, while the color of the links in the second part indicates the regions in which such interactions take place.

In the second example we illustrate the effect of the second term. We assume a simpler scenario where the country has nine areas with a corresponding decaying factor given by $\alpha = 2.4094$ and faces a PS as depicted by Figure 2.7b. In this case we assume that the country produces four goods as labeled in Figure 2.7a so that in the purple areas goods 3 and 4 are produced with the same intensity, goods 5 and 6 are produced with the same intensity in the green areas and good 11 is produced in the white area. We assume that the links in the PS, particularly the one between goods 5 and 11 and the one between goods 3 and 11 are equal. Under this scenario, as we observed in the previous example, the purple and green regions should be preserved by the first term of the objective function. Our attention is focused now on how the white area in the middle is assigned to one of the two regions. Notice that the interaction between goods 5 and 11 causes the arising of good 7, which is more complex than good 2. Thus, we expect that a solution of the p-IE model with $p = 2$ should assign good 11 to the green region, which effectively happens as we can observe in the Figure 2.7c.



(a) PS example 2.

3,4	5,6	5,6
3,4	11	5,6
3,4	3,4	5,6

(b) Spatial distribution of the industries in example 2.

W_1=0.99900099, W_2=0.000999, OF= 1.2347314

3,4	5,6	5,6
3,4	11,7	5,6
3,4	3,4	5,6

(c) Optimal region assignation.

Figure 2.7: This figure summarizes example 2. In Figure 2.7a we have the PS for this economy with 5 available nodes and two targets with their respective complexity. Figure 2.7b shows the spatial distribution of the available goods (3,4,5,6,11) which we assume are equally produced and divided into the following two spatial clusters: the one in which 3 and 4 interact (purple) and the one in which 5 and 6 interact (green). Figure 2.7c presents the optimal solution for example 2, and depicts how the area containing 11 is assigned to the cluster generated by 5 and 6 , thus causing the arising of 7 (in red).

2.4 The heuristic

In this section, we address the computational complexity of the *p*-IE problem by designing a heuristic solution. The use of heuristic methods has been a recurrent alternative within the family of the *p*-regions models (Rosenthal and Strange, 2004; De Groot et al., 2016; Laura et al., 2015; She et al., 2017). Regardless whether it is a *p*-regions or a Max-*p*-regions problem the structure of a heuristic for region design consists of two main blocks, as follows: (1) generate a set of initial feasible solutions and (2) take the best initial feasible solution and improve it by using a local search process based on a tabu search, simulated annealing, greedy algorithm, among others (Rosenthal and Strange, 2004;Duque et al., 2007). Computational experiments performed by Rosenthal and Strange (2004) show that the tabu search performs significantly better within the family of the p-regions problems. The use of a tabu search on an initial feasible solution is a practice that dates back to the nineties with the seminal contribution by Opeshaw and Rao (Openshaw and Rao, 1995).

Algorithm 2.4.1 presents the heuristic for the p-IE problem. The construction phase is inspired by the "growing regions" strategy first proposed by Vickrey (1961), which generates an initial feasible solution in two steps, as follows: (1) select a set of p areas, which are called "seed" areas, and (2) starts assigning the neighboring unassigned areas until all areas are assigned to a region (i.e., a region "grows" around each initial seed). There exist many variants of both steps; in our case, we select the initial seeds using the k-means++ algorithm proposed by Arthur and Vassilvitskii (2006), which seeks a good, spread out, location of the initial seeds. The k-means++ has a probabilistic component that allows exploring different sets of initial seeds. Each seed area becomes a growing region around which the rest of the areas are assigned. For the second step, the algorithm identifies the bordering areas (unassigned areas that share a border with a growing region) and selects the one with the minimum distance to its neighboring growing region. The distance between a bordering area i and a growing region GR is calculated as the sum of the squared differences of the vector of the exports r_{li} and the average vector of exports of the areas already assigned to the growing region GR (see equation (2.16)),

$$d_{iGR} = \sum_{l} (r_{li} - \text{avg}(r_{lGR}))^2. \tag{2.16}$$

Each time an area is assigned to a region it is necessary to update the set of neighboring unassigned areas and the process repeats until all areas are assigned to a region. Since the construction of an initial feasible solution is a fast process, we generate multiple initial feasible solutions (*maxitr*) and choose the one that maximizes the objective function $Z(\cdot)$. The best solution, P_p^{best}, is then passed to the next block,i.e. the local search.

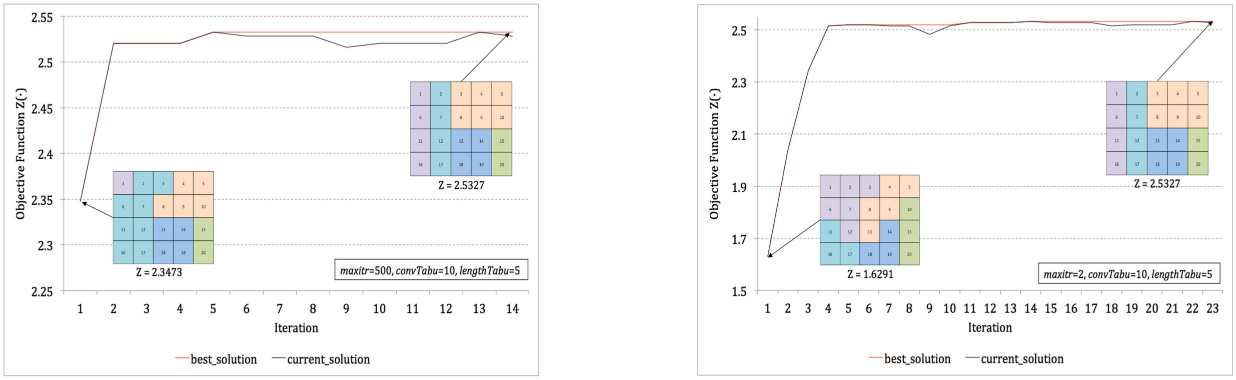
The second block, i.e., the local search, uses a tabu search algorithm (Glover, 1977; Glover, 1989 and Glover, 1990), which has shown good performance in spatial aggregation models (Rosenthal and Strange, 2004; Openshaw and Rao, 1995). Given a feasible solution, the neighboring solutions, N^s, are obtained by moving the bordering areas (i.e., areas that share a border with a neighboring region) to the neighboring regions, one at a time, while preserving feasibility. The tabu search explores these neighboring solutions to find improvements in the objective function $Z(\cdot)$. One key aspect of the tabu search is that it allows for a possible worsening of the objective function as a strategy for escaping from local optima with the hope that it will lead to an even better solution. The search stops after a predefined number of nonimproving moves (*convTabu*). Although multiple variants of this algorithm exist, we decided to use the simplest version because the local search is computationally intensive and the previous literature in the area of spatial clustering did not find clear advantages of using more sophisticated version of the tabu search algorithm (Rosenthal and

Strange, 2004; Openshaw and Rao, 1995).

Algorithm 2.4.1: THE-P-INNOVATION-ECOSYSTEMS (
A : Set of areas,
p : Number of regions,
$maxitr$: Number of initial feasible solutions to generate,
r_{li} : Production of good l in area i,
$lengthTabu$: Length of the tabu list,
$convTabu$: Number of non-improving moves before stop.)

```
Comment: Aggregate $n$ areas into $p$ spatially contiguous regions such that
$Z(S_p)$ is maximized.

$P_p^{best} = \emptyset$, best partition.
Construction Phase: Grow regions from seed areas
for $i = 1, 2, \cdots, maxitr$
  do { $seeds$ = k-means++$(A, p, r_{li}, W)$
       $S_p$ = GrowRegions$(seeds, A, r_{li}, W)$
       if $Z(S_p) > Z(P_p^{best})$
         then { $P_p^{best} = S_p$
Local Search Phase: Tabu search
$P_p^* = P_p^{best}; P_p^{current} = P_p^{best}; tabuList = \{\}; c = 1$
while $c \leq convTabu$
  do { $N^s$ = Set of feasible neighboring solutions of $P_p^{current}$
       if $N^s = \emptyset$
         then { $c = convTabu$
         else { for $P_p^{new}$ in $N^s$
                  do { if $P_p^{new} \notin tabuList$
                         then { if $Z(P_p^{new}) < Z(P_p^*)$
                                  then { $P_p^* = P_p^{new}$
                                         $P_p^{current} = P_p^{new}$
                                         $c = 1$
                                         $N^s = \{\}$
                                         $tabuList.add(P_p^{new})$
                                  else { $P_p^{current} = P_p^{new}$
                                         $convTabu = convTabu + 1$
                                         $N^s = \{\}$
                         else { if $Z(P_p^{new}) < Z(P_p^*)$
                                  then { $P_p^* = P_p^{new}$
                                         $P_p^{current} = P_p^{new}$
                                         $c = 1$
                                         $N^s = \{\}$
                                         $tabuList.add(P_p^{new})$
                                  else { $N^s = N^s - P_p^{new}$
                                         $tabuList.pop()$
return $(P)_p^*$
```

To illustrate the way in which each block of the tabu search contributes to finding a good solution we present the result obtained for the example in Figure 2.6 which aggregates twenty areas into five regions with an optimal objective function $Z = 2.5327$. To test each component (construct initial feasible solutions and local search), we ran the tabu search twice with different configurations: the first test was based upon generating 500 initial feasible solutions with the hope of finding a very good initial solution before moving to the local search. In the second test only two initial feasible solutions were generated. This second test relies on a local search to improve the initial solution as much as possible (hopefully to reach the optimum). In both cases, we stopped the tabu search after ten nonimproving moves (*convTabu*=10). Figure 2.8 shows the results of both configurations. When we explore 500 initial feasible solutions (Figure 2.8a) the algorithm finds a solution that takes only four iterations of the local search to find the optimal solution (plus ten nonimproving moves for convergence). When we generate only 2 initial feasible solutions (Figure 2.8b) the solution that enters the local search is considerably further from the global optima compared to the previous configuration. In this case, the local search process requires thirteen iterations to find the optimal solution (plus ten nonimproving moves for convergence). Note also how the local search process escapes a local optimal between iterations 8 and 13 in Figure 2.8b.



(a) 500 initial feasible solutions.

(b) 2 initial feasible solutions.

Figure 2.8: Performance of Tabu Search under two different configurations, in Figure 2.8a with a high number of initial feasible solutions, and in Figure 2.8a with a low number of initial feasible solutions.

2.5 Case of study

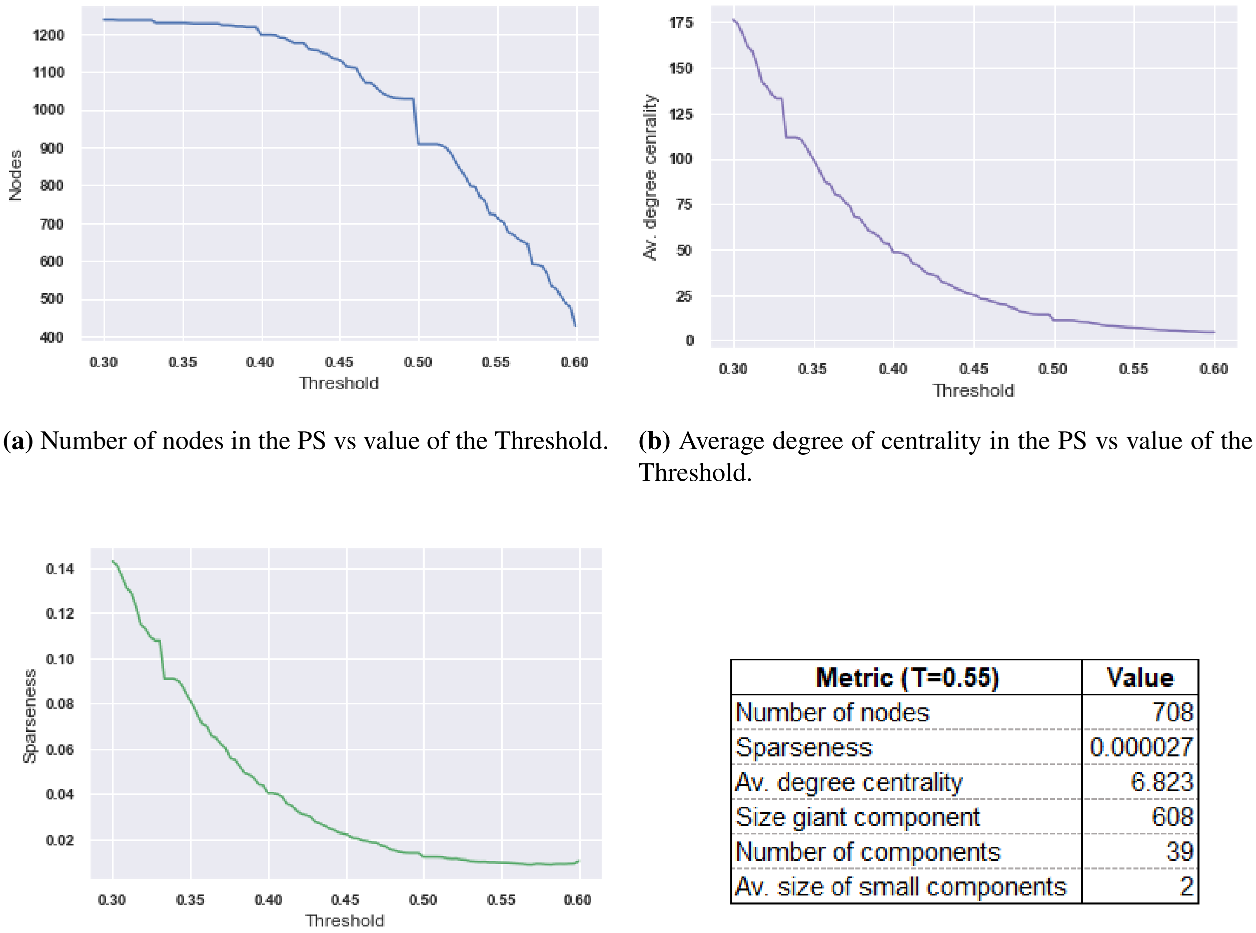
As an application of the p-IE model we address the problem of finding optimal industrial ecosystems for Colombia.

2.5.1 Data

PS data

We developed a version of the PS using trade data from the Centre dÉtudes Prospectives et d'Informatio Internationales (CEPII), which contains data for 128 countries, for the time period of 1995 to 2010 and includes observations of 1,240 products classified under the nomenclature of the harmonized system at the 4 digit level (HS4). We computed the weights of the network using equation (2.1). Following the lines of Hidalgo et al. (2007a) we opt for a threshold of 0.55 to define the PS, although there is a range of possible thresholds that would work for our purposes. Additionall, since our dataset differs from the one used for the computations in Hidalgo et al. (2007a), we provide in Figure 3.7 the respective analysis and diagnostics to justify why this election is still reasonable in our case. In our case, we select a PS with 708 nodes that includes the basket of exports of Colombia. Another important difference is that we have higher average degree centrality (6.823 vs 4 of the PS in Krugman (1991b)). In our case, there are 100 nodes outside of the giant component. Nevertheless, all of them are in small components with negligible size. Thus, the connectivity and sparseness properties are basically preserved by the giant component and we can restrict our analysis to the latter one.

For a better visualization of our version of the PS and for the subsequent descriptive analysis of our results we identified 7 main clusters on the graph using the standard Louvain Modularity algorithm (Blondel et al., 2008). These clusters contain 92.92% of the nodes of the giant component. We can also identify the position of Colombia on the PS by detecting the set of nodes in which Colombia has an RCA greater than 1. We summarize this information in Figure 3.8. Colombia reveals comparative advantage in 217 goods and 187 of them belong to the giant component of the PS.



(a) Number of nodes in the PS vs value of the Threshold.

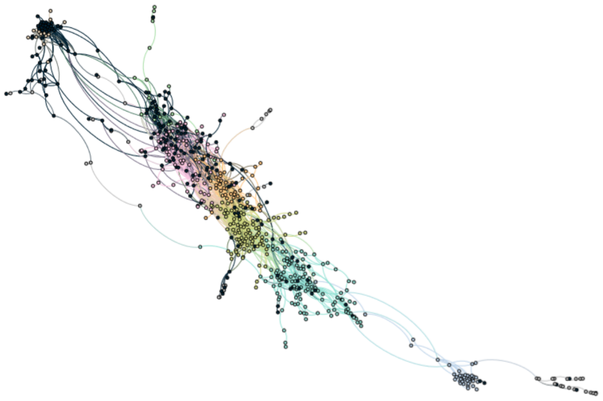
(b) Average degree of centrality in the PS vs value of the Threshold.

(c) Sparseness of the PS vs value of the Threshold.

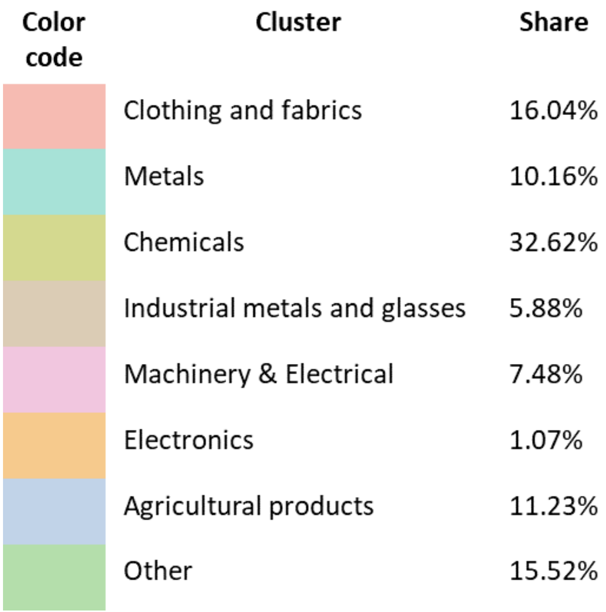
Metric (T=0.55)	Value
Number of nodes	708
Sparseness	0.000027
Av. degree centrality	6.823
Size giant component	608
Number of components	39
Av. size of small components	2

(d) Description of the PS when a threshold of 0.55 is used.

Figure 2.9: We present how the network is becoming more informative as we increase the threshold. This is reflected in how the number of nodes in Figure 2.4a decays slowly, while the degree centrality in Figure 2.4b and the sparseness in Figure 2.4c decrease rapidly. With a threshold of 0.55 we have a significant number of nodes (708), most of them in the giant component and with small components of a negligible size (2 nodes in average) as depicted in Figure 2.4d.



(a) Location of Colombia on the PS.



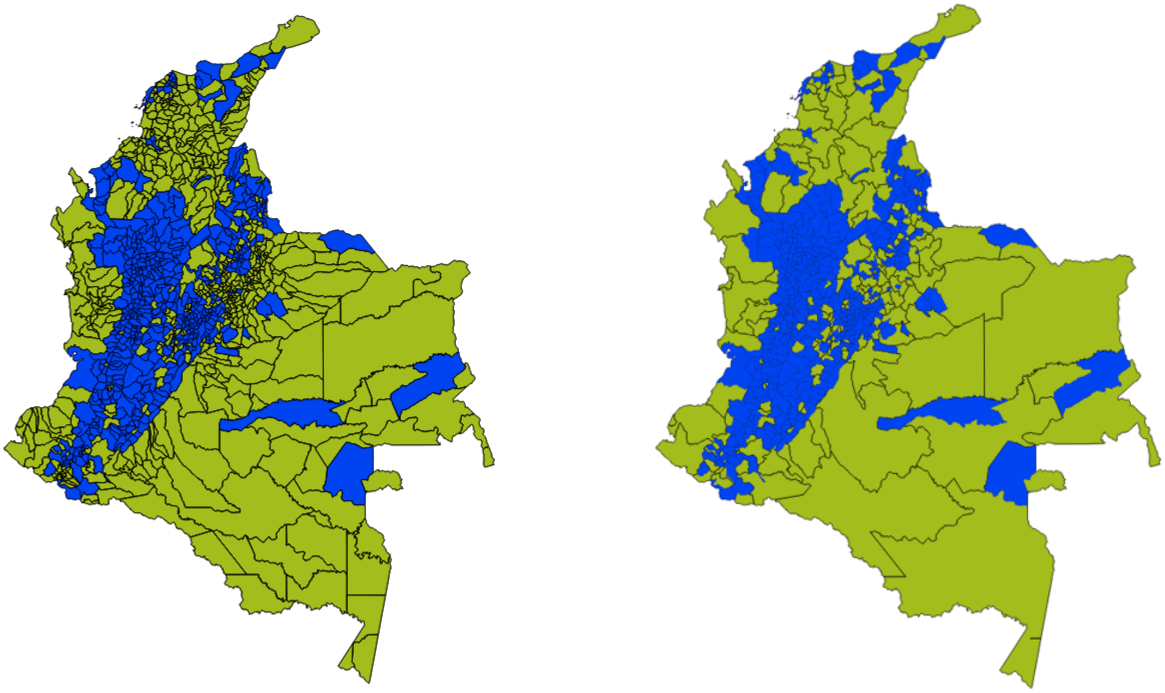
Color code	Cluster	Share
🟥	Clothing and fabrics	16.04%
🟦	Metals	10.16%
🟨	Chemicals	32.62%
🟫	Industrial metals and glasses	5.88%
🟪	Machinery & Electrical	7.48%
🟧	Electronics	1.07%
🟦	Agricultural products	11.23%
🟩	Other	15.52%

(b) Share of Colombian nodes in the PS.

Figure 2.10: In the first part we depict the communities detected with the Louvains method in the giant component of the PS and we mark in black, on top of them, the nodes in which Colombia reveal comparative advantage. The color code of the communities and the share of Colombia in each one of them is summarized in the second part. The order in which the communities are listed corresponds with their relative size compared with the giant component of the PS.

Geography data

The georeferenced data required to compute the interactions between industry areas are obtained from the open source information in the Datlas tool of Bancoldex. This dataset contains the exports per industry in HS4 and per municipality for the years 2008-2017. We use this export data for the year 2014 to compute the distribution matrix (analogous to the one depicted in Figure 2.2b) and the subsequent parameters required for the p-IE model. Since several municipalities in Colombia have a very incipient or nonexistent industry, we found zero exports for 591 out of 1120 municipalities. We overcome this issue by merging the municipalities with zero exports with their neighboring municipalities with positive exports in at least one industry. We defined these new areas by implementing the Max-p Regions algorithm (Duque et al., 2012) and using exports as the decision variable. This algorithm merges the municipalities into an optimal number of spatially continuous areas, following a queen adjacency criterion, so that the distance between the municipalities are minimized. The distance function is defined as the Euclidean difference of the coordinates of the centroids of the municipalities plus the difference of the department ID of each municipality (the departments in Colombia are regional divisions similar to the states in the USA). This distance is chosen such that the resulting regions are compact and prefer to remain within the same department (see Figure 2.11), preserving geographic cohesion. Finally note that the Max-p Regions algorithm creates new areas such that only one of the member municipalities has positive exports.



(a) Municipalities with positive exports.

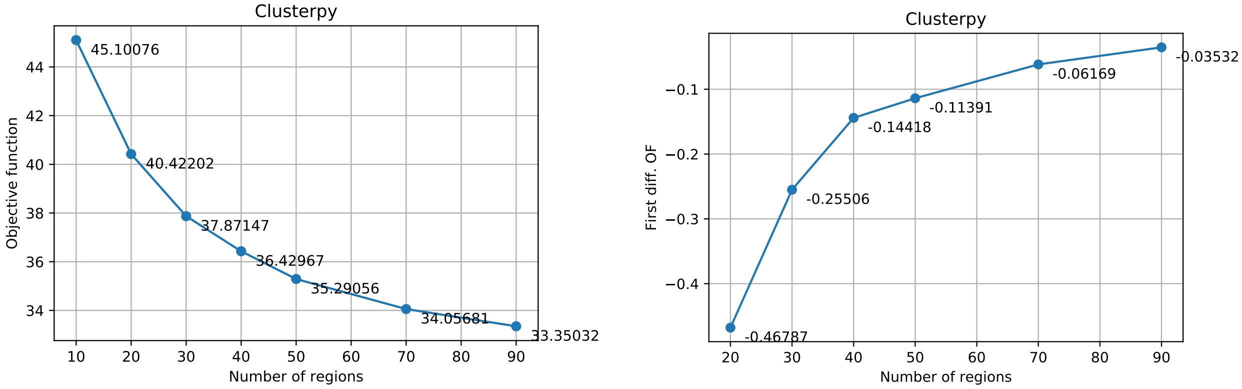
(b) Areas obtained after applying the Max-p regions model.

Figure 2.11: The first part shows the map of Colombia divided in 1,120 municipalities, where 529 of them (marked in blue) have positive exports in some industry. Second part depicts the regions obtained after applying the Max-p Regions algorithm to the first part.

The distance matrix (d_{ij}) is computed as the Euclidean distance between the centroids of the resulting 529 areas. We determined the decay parameter α similarly to that of of the examples in Section 2.3.4. In the same manner as that case we will take the parameter such that the distance decaying function f vanishes at the radius of an average region of size $\frac{n}{p}$. However, since the geography in this case is less rigid, we assume that the optimal regions could resemble a circle more than a square. Hence, we will take $R_{av} = \sqrt{\frac{n}{\pi p}}$ as the radius of an average region and $M = 1,382.2$ km. Therefore, we have that $\alpha = \frac{\ln(2)}{\ln\left(\frac{M}{R_{av}}\right)}$.

2.5.2 Results

We solved the *p*-IE model using our heuristic for a range of values of *p* between 10 and 90 as depicted in Figure 2.12. Since the heuristic algorithm (at least in the first stage) behaves such as a clustering algorithm we opted for a simple application of the Elbow method to select a first candidate for the optimal number of regions. From this analysis we determined 40 as our first candidate.



(a) OF vs Number of regions.

(b) First difference OF vs Number of regions.

Figure 2.12: In the first part we show how the optimal value of the objective function behaves in terms of the number of regions and depict how the slope starts to flatten significantly for large values of *p*. In part two we depict the behavior of the objective function, detecting that the largest variation in the slope occurs at *p*=40.

However, the solution corresponding to *p*=40 has 15 singletons (regions constituted by one area). Different implementations of our model for smaller scenarios suggest that when the number of regions is high, the model has the tendency to isolate areas with less valuable interactions turning them into singletons so that the most valuable interactions are preserved within the larger regions. This and several discussions with policy makers willing to implement this model in the Colombian case made us opt for a regional division without singletons; thus, we proposed to reduce the number of regions such that each singleton is incorporated into a larger region. For this reason, we propose the model for Colombia with *p*=25. For this value of *p* we have a decaying power $\alpha = 0.20041$ and a configuration as depicted by Figure 2.13. We classify these regions according to their share in the total amount of exports of Colombia during 2014. We chose this classification to be consistent with the metrics of localization and of the PS that were based upon the export data. Table 2.1 provides an initial description of the geographic characteristics of the regions and the corresponding export share for each one of them.

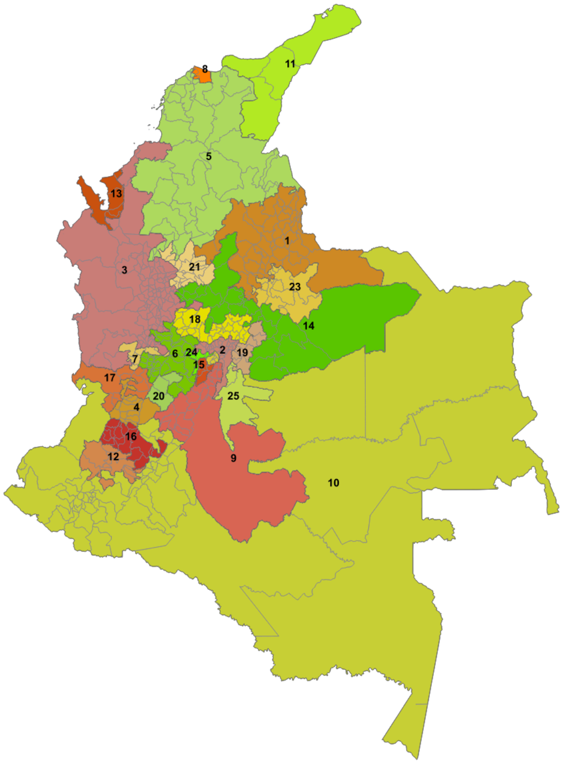


Figure 2.13: This figure depicts the solution for p=25 classifying the resultant regions with colors and labels. We label them in decreasing order with respect to their share in the total exports of Colombia in 2014 (see Table 2.1).

In Table 2.2 we summarize some relevant properties of the regions from the perspective of the Hausmann and Hidalgo's theory. The first two metrics correspond to a measure of the size of the subgraph spanned by each region within the PS. The column of industries shows how many industries of the giant component of the PS are produced in the region. The second column corresponds to the sparseness of each subgraph (see Barabási et al., 2016), i.e., for a region with N industries and L links among these industries its sparseness is given by $\frac{2L}{N(N-1)}$.

This also provides a sort of measure of the extent of the interactions among the key industries inside the region. The third metric indicates the relevance of the goods produced within the region as stepping-stones in the PS. We compute the fraction of shortest paths between any pair of nodes passing through the nodes of the region, i.e., the group betweenness centrality for these nodes (see Brandes, 2001). This measure gauges the extent in which the goods produced in the region are determinants in the diffusion processes in the PS. The fourth column is a weighted average of the normalized complexities of the goods produced within the region. More explicitly, for a region R producing the set of goods P (a subset of the PS) we consider the following complexity measure:

$$C_P = \sum_{i \in R} \sum_{l \in P} r_{il} c_l \tag{2.17}$$

Region ID	**Export share**	**Number of areas**	**Area (Km^2)**
0	19.123%	48	54,760
1	19.049%	28	4,958
2	18.565%	102	85,578
3	10.650%	16	7,095
4	6.619%	40	102,746
5	6.173%	36	13,755
6	3.895%	8	2,774
7	3.894%	3	1,298
8	2.309%	21	70,797
9	2.136%	58	369,591
10	1.787%	4	33,649
11	1.679%	13	9,673
12	0.896%	5	7,349
13	0.828%	23	74,130
14	0.714%	4	1,627
15	0.430%	18	9,394
16	0.417%	11	10,406
17	0.306%	29	10,762
18	0.208%	10	5,051
19	0.207%	3	4,199
20	0.066%	20	8,430
21	0.026%	6	687
22	0.013%	14	14,621
23	0.008%	5	1,166
24	0.002%	4	11,152

Table 2.1: Summary of some initial characteristics of the resultant regions.

Region ID	Number of industries	Sparseness	Betweenness centrality	Complexity
1	73	7.116%	0.40138	1.322
2	88	6.792%	0.44730	15.509
3	89	6.639%	0.44676	13.757
4	80	7.184%	0.44494	9.441
5	78	7.426%	0.43160	2.699
6	57	8.521%	0.28186	0.580
7	57	10.338%	0.24048	0.386
8	78	6.993%	0.43553	3.465
9	43	6.755%	0.29255	0.161
10	38	10.242%	0.33938	0.189
11	25	8.333%	0.24083	0.073
12	21	5.714%	0.09762	0.170
13	41	13.537%	0.19502	0.046
14	21	16.190%	0.14960	0.020
15	74	6.479%	0.37626	0.974
16	30	7.586%	0.21120	0.151
17	40	7.692%	0.24234	0.239
18	47	7.031%	0.31164	0.076
19	20	8.421%	0.26320	0.095
20	1	100.000%	1.00000	0.002
21	55	8.350%	0.34788	0.186
22	11	5.455%	0.09763	0.008
23	5	10.000%	0.06123	0.001
24	31	7.957%	0.29851	0.068
25	5	0.000%	0.04698	0.006

Table 2.2: Summary of some attributes computed for each region for the solution corresponding to p=25.

We want to point out that some strange features of this solution are a direct consequence of Colombian export structure. For example, the large region 10 is formed by 58 areas where 10 of the municipalities make up 70% of its size in the PS; this reflects the absence of exports (or significant exports) in this part of the country and precludes from defining finer divisions. Another example is given by region 20. This region only exports one good in which Colombia has a comparative advantage (coffee beans), though this good has small betweenness centrality and complexity values. These two radical cases show the main reasons that could force the model to deviate from the expected solutions, i.e., gathering areas into macro-regions or defining regions that specialize in goods with low complexity.

Analysis of regions 1 and 3

We conclude our analysis of the output with a comparison between region 1 and region 3. We choose these two for the following two reasons: (1) both have relatively similar size in km^2; (2) the export structures of both regions are distinct enough so that we can exemplify with them two types of regions in Colombia, i.e., regions following a diversification strategy (region 3) and regions specialized in agriculture or extractive economic activities (region 1).

In Figure 2.14 and Figure 2.15 we provide a thorough description of region 1 and region 3 that may be useful for policy making if we regard them as innovation ecosystems. We provide a detailed account of the exports of each region in the form of tree diagrams in part (c) of each figure. These diagrams are classified according to the communities identified and described in Figure 3.8b. It is worth noting that these communities do not follow any standard industrial classification, but are completely determined with clustering methods (Louvain algorithm). Thus, it is possible to have industries that do not seem to belong to the corresponding productive sector. The size of the boxes within each category are weighted by the betweenness centrality (BC) of the industry as a node in the PS. As discussed before, the BC tells us how central a node is in the diffusion of a country on the PS. Actually, we may think of them as stepping-stones in the process of reaching more valuable or complex nodes. Thus, since region 1 and region 3 are intended to foster innovation processes over the short and long run, policy makers should focus not only on the complex goods, but on those that are essential to move towards complex goods.

Region 1 specializes in the refinement of petroleum and contributes 18% of the 44% of the share of the petroleum in Colombian exports. Even more dramatically, the origin of this export is concentrated in one city in region 1 (Barrancabermeja) which has one of the most important refineries of the country. Additionally, after taking the subset of goods that belong to the PS with the prescribed threshold (which excludes petroleum) we see that region 1 is still specialized in the production of commodities such as melons and tomatoes (see Figure 2.14c). In contrast, region 3 has relatively strong industries that contribute significantly to the production of complex goods (reaction & catalytic products, acrylic polymers, other rubber products, etc.) combined with the production of commodities (in the rural parts of the region) such as fermented milk, cheese and so forth (see Figure 2.15c).

The graphic information provided in parts (b) and (c) of Figure 2.14 and Figure 2.15 enables us to give a more detailed answer to the issue related to determining the current situation of the region. In fact, from Table 2.2 we already know the average complexity, the number of industries and how central the goods of each region are, on average. From Figure 2.14 and Figure 2.15 we can explain in what sense region 3 is more complex and more central (on average) than region 1. Since the boxes in the tree diagrams in Figure 2.14c and Figure 2.15c also contain information about the complexity of goods normalized between 0 and 1 (Comp) and the share of the country in total exports of each good of the country during that year (Sh), we can easily observe that the products in region 3 account for both a large share of exports and a high complexity (in contrast with region 1). Additionally, the network of available and target goods rendered in part (b) of Figure 2.14 and Figure 2.15 allow us to observe how existing industries will probably thrive. For example, we point out that in region 3 the clusters of Clothing and fabrics and Chemicals are larger and better connected than in region 1. This is of particular interest in terms of complexity since the cluster of Chemicals in each region contains (on average) more complex goods than the other clusters.

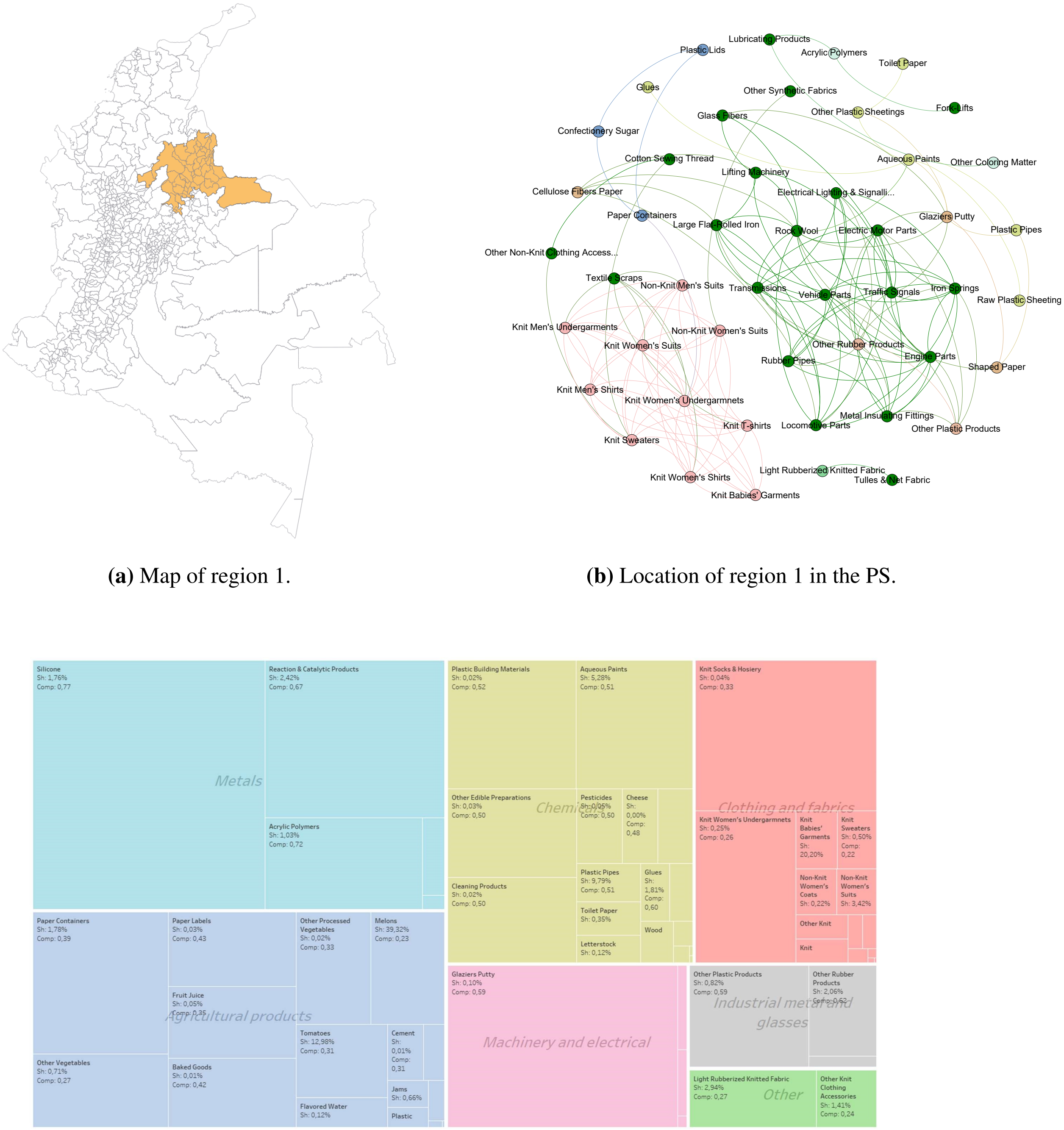
Finally, we present prospective strategies of innovation for each region. Our implementation of the p-IE model finds 25 target nodes for Colombia. All 25 targets are described in Table 2.3 together with their respective complexity values and the probability of each of the two regions explored to introduce these new goods in their economy. These probabilities are computed as a weighted sum of the following type,

$$\text{Prob Reg}_j(v) = \frac{\sum_{i \in R_j} \sum_{l \in P} r_{il} y_{lv}}{\sum_{l \in P} y_{lv}}. \tag{2.18}$$

where v is the target good, and R_j is the set of areas forming the region j (for j=1,2). This computation is a linearized and continuous version of the formula proposed by Alshamsi et al. (2018). Since our computations use as inputs the shares of the regions with respect to the total production of the country, these probabilities must be understood in a relative way, i.e., how likely a region is to develop a comparative advantage in one good with respect to another. We notice that region 3 has a positive probability to branch into all 25 products, whereas region 1 only exhibits 21 products with positive probabilities. We can illustrate how to use this information with two simple examples. Combining parts (b) and (c) of Figure 2.14 and Figure 2.15 with Table 2.3 we notice that region 3 has a considerably higher probability of developing comparative advantage in the last 4 goods listed in Table 2.3 because all of them are highly related to goods that are intensively produced by region 3 and that belong to the cluster of Clothing and fabrics which is highly connected in this region. On the other hand, the most complex target good (fork-Lifts) has a considerably higher probability to be produced in region 3 compared with region 1. This difference underscores the fact that region 3 produces a significantly higher share of the products in the cluster of Metals related to forklifts in the PS as we can observe in Figure 2.15b and Figure 2.15c.

Name	Complexity	Prob Reg1	Prob Reg3
Fork-Lifts	0.76808	0.095%	14.164%
Lubricating Products	0.74126	0.070%	7.620%
Transmissions	0.72552	0.101%	4.427%
Electrical Lighting & Signalling Equipment	0.70289	0.129%	6.088%
Vehicle Parts	0.69214	0.130%	4.803%
Large Flat-Rolled Iron	0.68408	2.518%	4.728%
Rock Wool	0.68167	1.619%	3.608%
Lifting Machinery	0.67676	2.320%	0.310%
Traffic Signals	0.67154	0.104%	2.662%
Iron Springs	0.67059	0.149%	3.452%
Locomotive Parts	0.66780	1.946%	4.356%
Rubber Pipes	0.65993	0.129%	5.466%
Engine Parts	0.64322	0.153%	3.529%
Metal Insulating Fittings	0.63526	0.245%	9.380%
Electric Motor Parts	0.62512	0.119%	3.983%
Glass Fibers	0.60791	0.001%	0.017%
Whey	0.58710	0.000%	14.066%
Electrical Insulators	0.58070	0.000%	4.167%
Letterstock	0.53960	0.000%	7.697%
Bovine Meat	0.50841	0.000%	17.160%
Tulles & Net Fabric	0.34832	0.719%	8.273%
Other Non-Knit Clothing Accessories	0.32376	0.054%	15.447%
Other Synthetic Fabrics	0.32241	0.376%	27.458%
Cotton Sewing Thread	0.31512	0.102%	26.012%
Textile Scraps	0.29533	0.602%	60.167%

Table 2.3: Target industries for the solution ranked by Complexity. Here, Complexity stands for the standardized complexity index of the goods (17), Prob Reg 1 and Prob Reg3 are the probabilities of region 1 and region 3 to produce such goods, respectively, computed using (2.18). All 25 industries have a positive probability to appear in region 3, however, there are four industries that do not have any related activities in region 1 and correspond to the zero values in the third column.

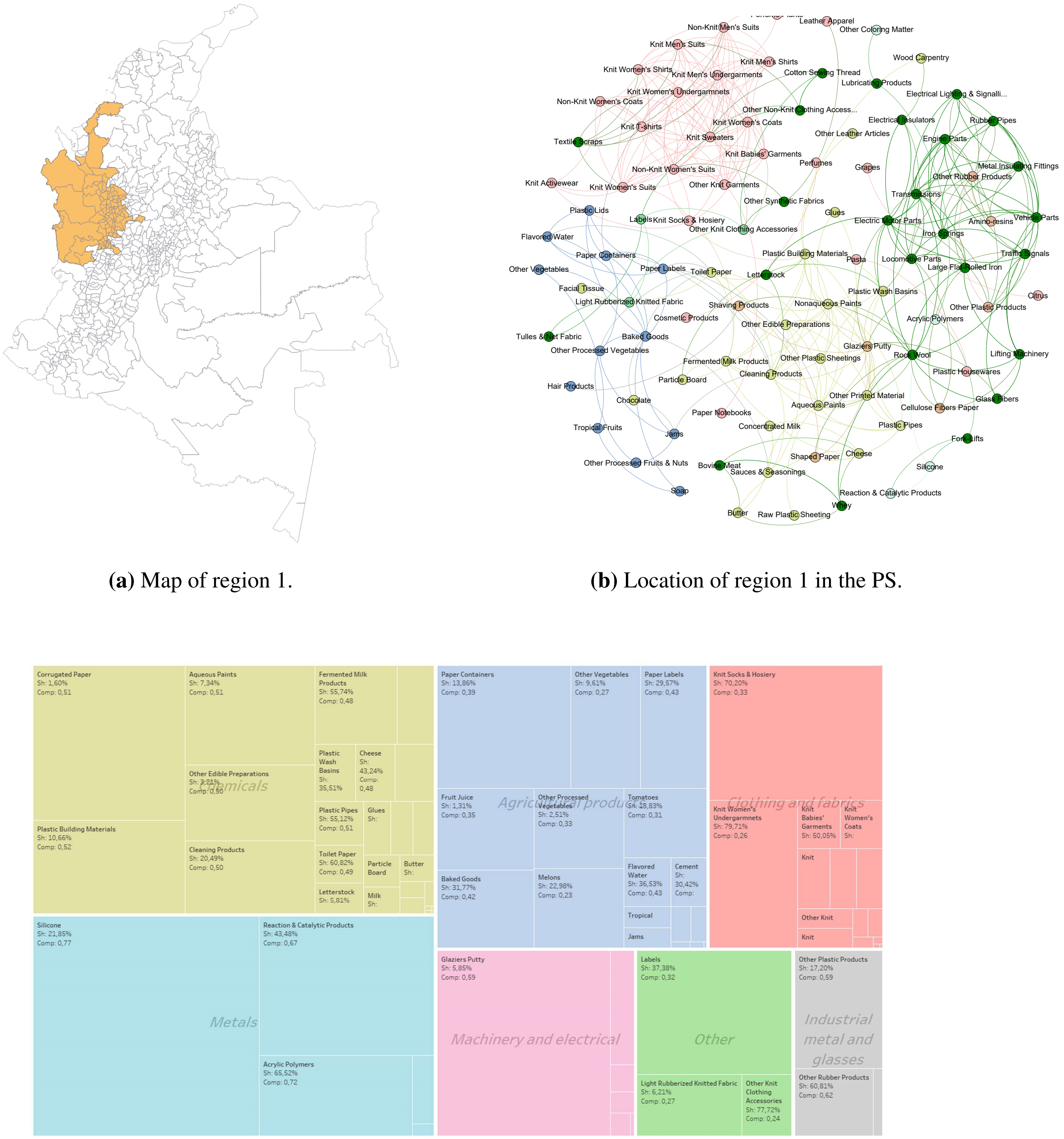


(a) Map of region 1.

(b) Location of region 1 in the PS.

(c) Exports of region 1 per cluster.

Figure 2.14: Figure 2.14a depicts the geographic location and the areas that form region 1, Figure 2.14b depicts the target goods of region 1 in green. All the other nodes correspond to available goods in the region colored according to the nomenclature established in Figure 2.14b. The box diagram depicted in Figure 2.14cto the betweenness centrality of the nodes in the PS and each box also contains the share of the region in that activity with respect to the national value (Sh) and the respective normalized complexity index (Comp).



(a) Map of region 1.

(b) Location of region 1 in the PS.

(c) Exports of region 1 per cluster.

Figure 2.15: This figure is analogous to Figure 2.14. We want to stress the fact that the PS of region 1 depicted by Figure 2.15b is a strict subset of the one depicted in Figure 2.14b. In particular, region 3 has a larger set of available and target nodes. Additionally, it is important to notice the change in the composition of the tree map in Figure 2.15c with respect to the one in Figure 2.15c.

2.6 Conclusions

In this paper we introduced a new member of the p-regions clustering methods called the p-innovation ecossytems (p-IE) model. This model adds a new layer of complexity into the set of p-regions models by incorporating an underlying network structure that captures possible interactions between agents within the geographic areas. The main application (and motivation) of the model is the study of the evolution of the economic output based on the recent developments of economic complexity and evolutionary economic theory. More precisely, this model consists of the aggregation of n areas into p spatially contiguous regions that (1) maximizes the number of relevant interactions among the industries within the same region; and (2) identify strategic relationships between industries within regions that maximize the probability of activating innovation processes that allow the region to jump to more complex relevant goods that are yet to appear in the economy. Formally, the p-IE model was formulated as an MIP problem with a corresponding heuristic solution. From the economic point of view, we introduced this model as an application and extension of the theory of Hausmann and Hidalgo (done at the level of countries) to the study of the industrial evolution at regional levels.

From the perspective of policy development, we introduced several metrics as well as suggesting how they might be applied in conjunction with the p-IE model. This approach will help policy makers diagnose and assess the actual performance of the economic regions and the potential performance of each regionalization proposed by the model. We introduced these ideas in a case of study for the municipalities of Colombia and were able to identify 25 innovation ecosystems, which we ranked according to their contribution to the total exports of Colombia. We found a positive association between the complexity of these regions, their level of exports and the growth of their level of exports. Additionally, a further analysis applied to the first and third regions (according to our classification) showed how to obtain inputs that would help the policy makers design policies to exploit the current situation of the regions (amount of production) and to enhance the production of key industries in the innovation process (with high betweenness centrality) so that the target industries (with a high complexity) arise within the regions.

As it is formulated, the p-IE model can be implemented to other clustering problems as long as the situation can be described in terms of the following two inputs: (1) a distribution of areas inside a region that contains agents or attributes of some sort, (2) a network structure that relates the agents or attributes and identifies the most important attributes or agents and the key interactions among them. We suggest two alternative applications of this model. First, given an input-output matrix (of money transfers among industries) we could use the model to produce optimal regions so that the policy maker can identify the supply chains, preserve them within the resulting regions and design strategies to complete the chains inside the regions. Second, we can think in a scenario involving a network of amenities in a city (such as hospitals, schools, museums, gas stations, etc.) and the commuting flows or connectivity flows weighting the links among them. In this case, it would be possible to apply the model to divide the city into optimal zones in which a citizen could find well-connected and complementary sets of amenities. The model could also be used to identify the lacking amenities within each zone and measure the extent to which a given zone is not yet self-contained or self-sufficient. This type of exercise would be a relevant contribution to solve mobility issues and to promote the use of nonmotorized transport methods.

Our approach in this paper is static by nature; therefore, its usefulness may vary when the underlying economic and network structure change. Furthermore, since the main application of this model is intended to enhance the economic growth and the innovation through the implementation of local policies, it is necessary to better understand the behavior of the model in dynamic scenarios; namely, when many target goods are reached and a new regionalization is required to guarantee the optimality of the solution, or when the aims of the policy makers are directed to a specific set of industries rather than the ones dictated by the economic complexity. This extension is the main topic of the third part of this thesis.

Chapter 3

Diversification dynamics and optimal policies for innovation ecosystems.

Abstract

The innovation dynamics in industrial clusters are gradual and adaptive processes that usually take place in specific spatial units (cities, metropolitan areas or regions). The effectiveness of these innovations as boosters of economic growth in a region is strongly determined by the relatedness structure of the economic activities already present in the region and by their level of complexity. In this paper [1], we introduce two different long-run diversification strategies that capitalize the initial endowment of activities in a country, to either maximize the level of economic complexity of the country or to pursue specific targeted innovations. We formulate our models as two-step optimization problems, where the first step determines optimal diversification routes for the country, while the second step applies an iterative spatially constrained algorithm to aggregate the areas of the country into (a prescribed number of) innovation ecosystems. We illustrate both strategies analyzing diversification opportunities for Colombia at its municipalities.

3.1 Introduction

The evolution of the productive structures of the economies is a never-ending, path-dependent and organic process. Important empirical studies have found that the dynamics of diversification and specialization followed by the countries are strongly correlated with their income or economic growth level (Imbs and Wacziarg, 2003; Hausmann and Hidalgo, 2011; Hausmann et al., 2014). This phenomena was addressed in Hausmann, Hidalgo and others (Hausmann and Hidalgo, 2011; Hidalgo et al., 2007b) arguing (in the same lines of the Hecksher-Ohlin reformulation of the Ricardian model) that the types of capabilities available in a country determines whether an industry is feasible (likely to appear) in the economy. Thus, according to Hausmann and Hidalgo's approach, the diversification process of a country can be thought of as the progressive and selective

[1] Joint work with Nancy Lozano of the World Bank.

acquisition of capabilities. These capabilities, as put forward by Hausmann and Hidalgo (2011), are the requirements to the emergence of an industry in the broadest sense: institutional requirements, resources, knowledge, etc. Also, at a microeconomic level, this theory is reinforced by ideas from cognitive theory and evolutionary economics which claim that an industry can diversify its productive portfolio as long as the new goods are not quite distinct in terms of cognitive distance (required know-how) from the ones already being produced (Boschma et al., 2015; Neffke et al., 2011).

Measuring the relatedness among products as the overlapping of required capabilities, although rather indirect, has shed light upon the patterns and restrictions faced by the diversification of economic output. For example, Bahar et al. (2014) considered the geographic nature of the capabilities, showing that the spatial proximity among countries explains significantly their current basket of exports and, moreover, evidencing that there is a high and significant spatial correlation between the diffusion of Revealed Comparative Advantage (RCA) of neighboring countries. This geographic effect is justified partially in Bahar et al. (2014) by the strong effect of the distance in the, arguably, more important capability: knowledge (see Thompson and Fox-Kean, 2005, Bottazzi and Peri, 2003; Bahar et al., 2014). Moreover, works like Neffke (2009) have pointed out that knowledge or know-how and other capabilities face strong mobility restrictions even within countries and then the economic regions should be regarded as the spatial unit in which the emergence of capabilities (and then innovation) is more likely to occur. Empirical studies like Boschma et al. (2015), Neffke et al. (2011), and Boschma et al. (2013) have provided evidence that at the level of regions or cities the evolution of the economic landscape is shaped by the underlying structure of relatedness among goods, encouraging the entrance of industries related with the existing ones and discouraging the survival of the poorly related with the local economic activities.

The p-innovation ecosystem model (p-IE model) introduced in Restrepo et al. (2020) was proposed as a quantitative answer to the problem of determining what local interactions among industries must be enhanced to guarantee the preservation of cohesive economic regions and ensure an optimal diversification that improves the economic complexity (where complexity is understood in terms of Hausmann et al., 2014). Explicitly, the p-IE model is devised to find optimal regional divisions of a country that seeks (1) to capitalize within each region the strategic interactions among industries that already operate in the country and (2) to encourage within each region the interaction of industries whose capabilities (know-how) combined are likely to foster the emergence of new valuable or more complex goods. This optimization model builds upon the theory of relatedness and diffusion of Hausmann and Hidalgo and it is motivated by the empirical findings on the patterns of regional diversification (Boschma et al., 2013; Bahar et al., 2014; Neffke et al., 2011; Boschma et al., 2015).

The p-IE model only addresses the static part of the question so that it enables policy makers to find optimal solutions, but only in the short-term. In the medium and long term, economic activity in a country is expected to diffuse into new industries changing the parameters of the p-IE, limiting the validity of its results for longer periods of time. On the other hand, the p-IE model only considers a one-step diffusion process, i.e. the optimization is carried out such that the country can branch into the closest products of existing industries (see Restrepo et al., 2020 or Section 3.3 for more detail). A dynamic approach to these problem, would instead require an iterative implementation of the p-IE model together with a model that describes the evolution of the productive landscape of the economy, i.e. the acquisition of new goods and the relative change in the distribution of the production of the available goods in the economy. The consideration of these dynamic strategy leads us to several relevant questions: at which rate the economy will evolve into the production of target goods? How would be the change in the shapes of the regions proposed as innovation ecosystems by the p-IE? Is there any clear diffusion patterns in the evolution of the production among the regions proposed by the model?

Governments like Saudi Arabia (Arabia, 2016), Indonesia, Peru and Chile (see Alshamsi et al., 2018) have established diversification as explicit goals in their economic policy agenda. Nonetheless, if a country tries to develop directly complex activities the current availability of capabilities may force it to follow alternative approaches, like diversifying first into more proximate economic activities before investing in the target activity (Imbs and Wacziarg, 2003; Hidalgo et al., 2007b). Thus, diversification policies require long-run schemes adapting the evolution of the economic incentives and, in this particular case, the never-ending process of industry branching. Governments can therefore set out to reach a diversification or complexity goal, or set out to reach specific products that are desired due to their higher complexity and subsequent contribution to value added.

Based on this dichotomy we consider two schemes to address both long run policy objectives. First, the case in which a country defines recursively innovation ecosystems (following the p-IE model) until it reaches a desired level of diversification or complexity; we call this case the non-restricted evolution scenario. Second, given a set of target goods proposed by the policy maker we redefine the notion of target nodes in the p-IE model so that the economy is encouraged to reach the stepping-stones towards the target goods; the iterative implementation of this modification of the p-IE model corresponds to the targeted evolution scenario.

In this work, we combine the ideas of the models of complex contagion, mostly inspired in Alshamsi et al. (2018), with the p-IE model. Our goal is twofold. First, we incorporate a dynamic component into the p-IE model to assess and recommend long run diversification strategies for countries and regions. Second, we offer a new alternative to understand the dynamics of innovation within a country incorporating regional effects, i.e. distance and boundary effects. Thus, we regard this model as a concrete and schematic proposal to implement efficient diversification strategies in an harmonious way coherent with the current industrial state of the countries at a regional level.

We look at alternative strategies of targeted products vs diffusion, using a dynamic version of the p-IE: The evolutionary p-IE models. The evolutionary p-IE models are formulated as two step optimization problems, in the first step, given a set of target nodes, we identify optimal routes (stepping-stones) that minimize the expected time to reach the target good through a process of complex contagion (following Alshamsi et al., 2018). The second step consists of the iterative application of a mixed-integer programming (MIP) which essentially corresponds to the p-IE model Restrepo et al. (2020) and that belongs to the family of the p-regions models devised by Duque et al. (2011). Since the p-IE model by itself is a non-deterministic polynomial-time hard (NP-hard) (Duque et al., 2011; Duque et al., 2012), we propose a heuristic solution to effectively compute a near-optimized (if not optimal) solution at each iterative step.

The evolutionary p-IE models provide detailed and systematic information about the regional and national strategies that should be followed by an economy pursuing long-run policy goals such as reduction of inequality or economic growth. These types of policies usually require structural transformation via diversification into more complex and riskier economic activities which are more distant to the current productive system and therefore harder to develop (see Imbs and Wacziarg, 2003; Hausmann and Hidalgo, 2011; Hausmann et al., 2014; O'Clery et al., 2018).

The rest of the paper is presented as follows: Section 3.2 contains the Literature Review, Section 3.3 presents the the main conceptual elements upon which our model is built up, in Section 3.4 we formulate and discuss the model and in Section 3.5 we implement it to depict the reach of our model in a case of study for Colombia. Finally, Section 3.6 contains the conclusions.

3.2 Literature review

The models proposed in this paper have two main components: (1) the p-IE model, and (2) diffusion on the PS. In this section we provide a brief account of the theories that encompasses the features of these two elements relevant for our work.

The p-IE is part of the family of p-region models introduced in Duque et al. (2011). The p-regions models solve the problem of aggregating n areas into p spatially contiguous regions so that a pre-determined policy goal is optimized. As the general quantitative regionalization methods, the p-regions model seeks to decompose a global policy objective into a prescribed number of regions (p) in which the efforts are decomposed in local tasks so that the policy goal is achieved more easily. Formally, the p-region models are formulated as mixed integer programming (MIP) optimization problems. The original model in Duque et al. (2011) was formulated with three different strategies to ensure spatial connectedness of the regions. This flexibility allows the p-region model to incorporate various types of objective functions without imposing additional constraints in the shape of the regions, and avoiding incurring in infeasible configurations. Thus, these models capture endogenously a broad range of spatial patterns (compact, elongated, among others) generated by the nature of the problem rather than any a priori geometric criteria. Other examples of p-regions models includes the determination of the optimal level of aggregation (Duque et al., 2012), the maximization of the compactness of the regions (Li et al., 2014), the definition of functional regions according certain criterion (Kim et al., 2015), among others.

The main difference between the p-IE and its predecessors is the underlying economic foundation provided by Hausmann and Hidalgo's theory. Two main elements of this theory are incorporated into the model: the Product Space (PS) and the complexity theory. The PS is a network that describes the relatedness structure of tradable goods. Two goods are linked in the PS if they are likely to be co-exported by a country (Hausmann and Hidalgo, 2011). The topology of this network has proven to be a strong predictor of the evolution of the patterns of diversification and specialization of the countries. Moreover, since the PS has a core-periphery structure, earlier work has show that baskets of exports specialized in goods located in the periphery of the PS are highly correlated with low levels of income per capita, whereas a well connected export basket in the core of the PS is highly correlated with high levels of income per capita as well as higher GDP growth (Hausmann and Hidalgo, 2011). On the other hand, the economic complexity of a good is a measure of the amount of capabilities (mostly knowledge) embodied in the good. The main properties and computations related with this measure are explained in (Hausmann et al., 2014). Analogously to the PS, the economic complexity also captures important information of the countries' potential to boost their economic growth through selective innovation (see Hausmann et al., 2014).

In the framework of the *p*-IE, the role of the PS is twofold: (1) capturing the key interactions between goods that are already produced in the economy, and (2) gauging the distances between the existent goods in the economy and the ones that are produced in the rest of the world but not in the country. In other words, the PS describes the current situation of the economic landscape of the country, as well as its short-run and long-run possibilities of diffusing into new products. The other element, the complexity, is used as a criterion to assess whether a good can be regarded as a target for the economy. A product can be consider a target as long as it is more complex than the present basket and more related to the existing economic output of the economy. In that case, the *p*-IE assigns a greater weight to the industrial interactions that can lead to such target product. Considering this, we can rephrase the description of the *p*-IE model given in the introduction to say that this model seeks to identify regions that enhance the industrial interactions so that (1) the most related goods in the PS can be clustered together, and (2) existing goods that are common neighbors of complex goods yet to be produced tend to collaborate to ease the diversification of the country towards this new good. From a policy perspective, this is useful because it points at regions where strengthening internal interactions will be conducive to both facilitating specialization and clustering of related activities while also favoring a move toward more complex, and hence higher value added products in the long term.

The PS has been used extensively to gain a better understanding about the strong path-dependence of the economies in their specialization and diversification processes (see e.g. Hausmann and Hidalgo, 2011; Bustos et al., 2012; Bahar et al., 2014). Since such dynamics are highly idiosyncratic and path-dependent, the topological and probabilistic components of the PS offers a reasonable framework to understand why it is difficult for countries specialized in commodities to innovate into more complex activities which embody more know-how. On the other hand, a series of works in evolutionary economic geography (EEG) have analyzed in a further extent the Hausmann and Hidalgo's findings at regional levels. Studies like Bustos et al. (2012), Frenken and Boschma (2007), Boschma et al. (2012), Boschma et al. (2015), Neffke et al. (2011), Neffke (2009) have argued that the structural changes required to the innovation processes (and, therefore, the economic growth) take place within regions. Regions are identified as the ecosystems in which the firms thrive, fail, disappear and branch due to various externalities (agglomeration, spillovers, etc) broadly discussed in the economic literature for decades. The main novelty in the works mentioned above lies in the identification of the nature of the diversification patterns that can arise within the regions (innovation ecosystems).

Remarkably, Neffke et al. (2011) showed for the regions in Sweden that the innovation process not only follows the dynamics dictated by the PS (more precisely, for an analogous network constructed in their work), but that the evolution of the economic landscape within the regions seeks to preserve the technological cohesiveness, i.e. industries related to the ones already existing in the region are more likely to enter whereas existing industries with a weak link with the other industries in the same region have a higher probability of exiting the region. These findings have been validated in several other studies (see Neffke et al., 2014).

In parallel, studies bringing together the literature on economic complexity and EEG to understand the nature of the path-dependence in the innovation processes, have explored how the information provided by the PS can be used to foster the blooming of emergent economies (either countries or regions). We summarize two approaches that are relevant for this work: first, to spread within the PS with "jumps" further along the space to non-contiguous nodes; and second, defining diffusion strategies on the PS. The first type of strategy is more aggressive and seeks to break the path-dependence predicted by the PS through government investments that promote the rise of industries that are distant from the current product basket or to redirect the diffusion of the country towards zones in the core of the PS that were originally less likely to be reached. This strategy is exemplified in Zhu et al. (2017) which shows the impact on China's region of the investment in new activities and in extra-regional linkages among industries, i.e. investment in the interaction between industries belonging to different regions. Further, work presented in Martin (2010) suggests that path-dependency is not the sole driver of the evolution of economic output, but instead, it is necessary to break path-dependency to induce industrial restructurings and, therefore, foster the structural transformation of the economy.

The second strategy seeks to exploit the topology and probabilistic structure of the PS to determine which are the best routes in the PS to reach desirable goods in efficient times. These models are based on the classical theory of complex contagion in which the diffusion is highly dependent in the interaction of many existing goods (infected nodes) to reach a new one (to infect a healthy node) Alshamsi et al. (2018).

3.3 Conceptual framework

This section contains two main blocks, the first one reviews the p-IE model and its components. The second part introduces some generalizations and adaptations of models of complex contagion on networks. More precisely, in the second part of this section we present the models that describe the dynamics of the inputs in the p-IE. These models will allow us to determine, among other things, the change of the existing production in the country, the arising of new products within the regions and the expected number periods between implementations of the p-IE. In the second part, we also address the problem of identifying optimal diversification strategies to reach long run target goods. These elements will serve as inputs to Section 3.4 where we introduce the free and guided evolutionary p-IE models.

3.3.1 Complexity theory and the PS in the p-IE model

The PS introduced in Hausmann and Hidalgo (2011) is defined as a network that measures the relatedness among all the tradable goods in the world in terms of their probability of being co-exported. More explicitly, the PS (for the purposes of our paper) is a network whose nodes are tradable goods and whose links have as weights the a measure related with the probability of a country of developing a comparative advantage jointly on both goods (see Section 3.5 or Hausmann and Hidalgo, 2011 and Hausmann et al., 2019 for further details and alternative constructions of the PS). The PS has been treated conventionally as an undirected network selecting the links whose weights exceed certain threshold (see Section 3.5 or Hausmann and Hidalgo, 2011). This conception of the PS captures mostly the topology of the relatedness structure of the tradable goods and it is useful to understand in a qualitative way the current state of an economy in terms of its production, mostly because the PS is a scale free network exhibiting a core-periphery distribution of the nodes. As an example of this, Hausmann and Hidalgo (2011) provides evidence of the stylized fact that developed countries are prone to produce well connected goods that are located in the core areas and along the main clusters of the PS, whereas developing countries usually produce poorly connected and peripheral goods. We provide a toy example of this in Figure 3.1 where we depict a general PS and highlight in blue the set of initial goods (IG) in which the (hypothetical) country has a comparative advantage.

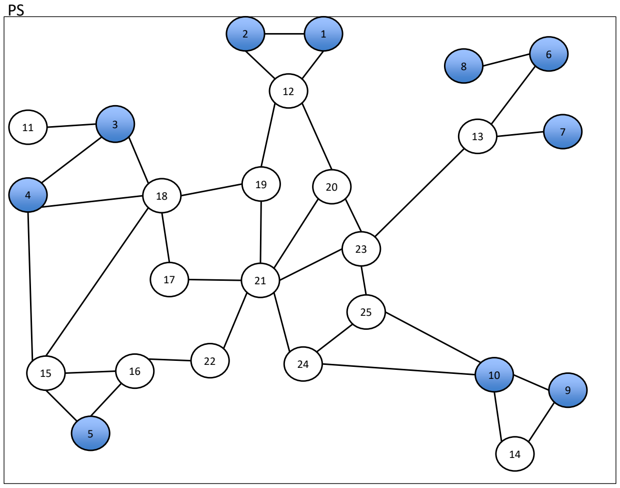


Figure 3.1: Layout of the PS of an abstract economy with 25 goods. Nodes 1 to 10 (in blue) represent the set of initial goods of the country IG. The white nodes are the rest of the goods in the PS (in the world) in which the country has not developed comparative advantage yet.

The set IG has its corresponding projection on the productive areas of the country according to the spatial distribution of the industries producing each good. So, if the country has a set of n productive areas $\{1, \cdots, n\}$, each $l \in IG$ has an associated distribution vector $r_l = (r_{l1}, \cdots, r_{ln}\}$ which indicates that the fraction r_{ll} of the good l is produced in the area i. So, according to this definition $\sum_{i=1}^{n} r_{li} = 1$ for every $l \in IG$. One of the main goals of the p-IE consists in capitalizing the industrial interactions indicated by the PS at the level of regions, this interactions are defined by an interaction term I_{ij}^{lm} defined as $r_{li} r_{mj}$ whenever the areas i and j belong to the same region and 0 otherwise. In Figure 3.2 we illustrate this interplay between the PS and the spatial/geographical distribution of the production of the goods.

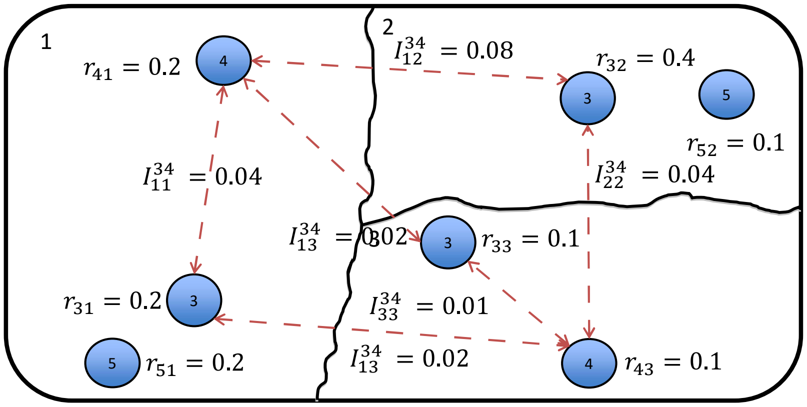


Figure 3.2: Representation of a generic region of our hypothetical country. This region has 3 areas (1, 2 and 3 and produces three out of the ten available goods in this country. This region produces $r_{31} + r_{32} + r_{33} = 0.7$ of the total production of good 3, (analogously) 0.3 of the total production of good 4 and 0.3 of the total production of good 5. The dashed red lines indicates the interactions between the areas arising from the interaction of industries linked in the PS and located in the corresponding areas.

The I_{ij}^{lm} factor is the way in which the p-IE model introduces a region or border effect because it only captures the interactions as long as they take place inside a region. This interaction term is not introduced explicitly in Restrepo et al. (2020), but a binary counterpart of it denoted by $t_{ij}^{lm,k}$ which is 1 if $I_{ij}^{lm} > 0$ when i and j belong to the region k and 0 otherwise. The other geographic component introduced in the p-IE model is a power-law distance decay penalization of the form:

$$f(x) = 1 - 2\left(\frac{x}{M}\right)^{\alpha}. \tag{3.1}$$

where M is the maximal distance between two areas in the country and $\alpha > 0$ is a decaying parameter. Notice that f is strictly decreasing, its range is given by $[-1, 1]$, $f(0) = 1$ and $f(M) = 0$.

So far we have introduced the elements incorporated in the p-IE model used to assess the current state of a country in terms of the industrial interactions at local level (incorporating distance decay and border effects). Before recalling the general form of the model, it is necessary to recall the notion of a target set. Given the static or short run nature of the p-IE model, it focuses on a single first stage of diffusion only, considering candidates belonging to the first order neighborhood of the set IG in the PS. We define this set as the set of candidate goods CG. The p-IE makes a further reduction of the set of candidates in two ways. First, since the model seeks to foster the innovation through industrial interaction, candidate nodes linked to only one node in IG are not regarded as targets simply because no interaction in IG can be established to encourage its introduction into the economy.

Second, the model is intended to diffuse into goods that increase the current complexity of the economy. In this case we refer to the notion of Economic Complexity Index (ECI) introduced in [bottazi]. In these terms, the model regard as target goods the ones whose ECI exceeds all the ECI's of its neighbors in IG, i.e., only preserves the candidates that contribute to an improvement in the complexity of the country's production basket. This set of nodes is defined as the set of target goods TG.[2]

The process of definition of the network that is used as an input for the p-IE model is explained thoroughly in Figure 3.3. The process has three steps, (1) it considers the set of first order neighbors (CS) as depicted in Figure 3.3a, (2) among the nodes in CS the p-IE only consider as target goods (TG) the ones which require the interaction of 2 or more nodes in IG as showed in Figure 3.3b and, finally, (3) it simplifies the network keeping all the elements of TG and keeping the elements of IG that interact either with another set of IG (generating an initial link IL) or with a target node. In this latter case, the p-IE introduces a new type of link (L_v) which associates all the common neighbors of a target node $v \in TG$ in IG. This final step is illustrated in Figure 3.3c. All the notation and nomenclature of this construction is explained in Figure 3.3d.

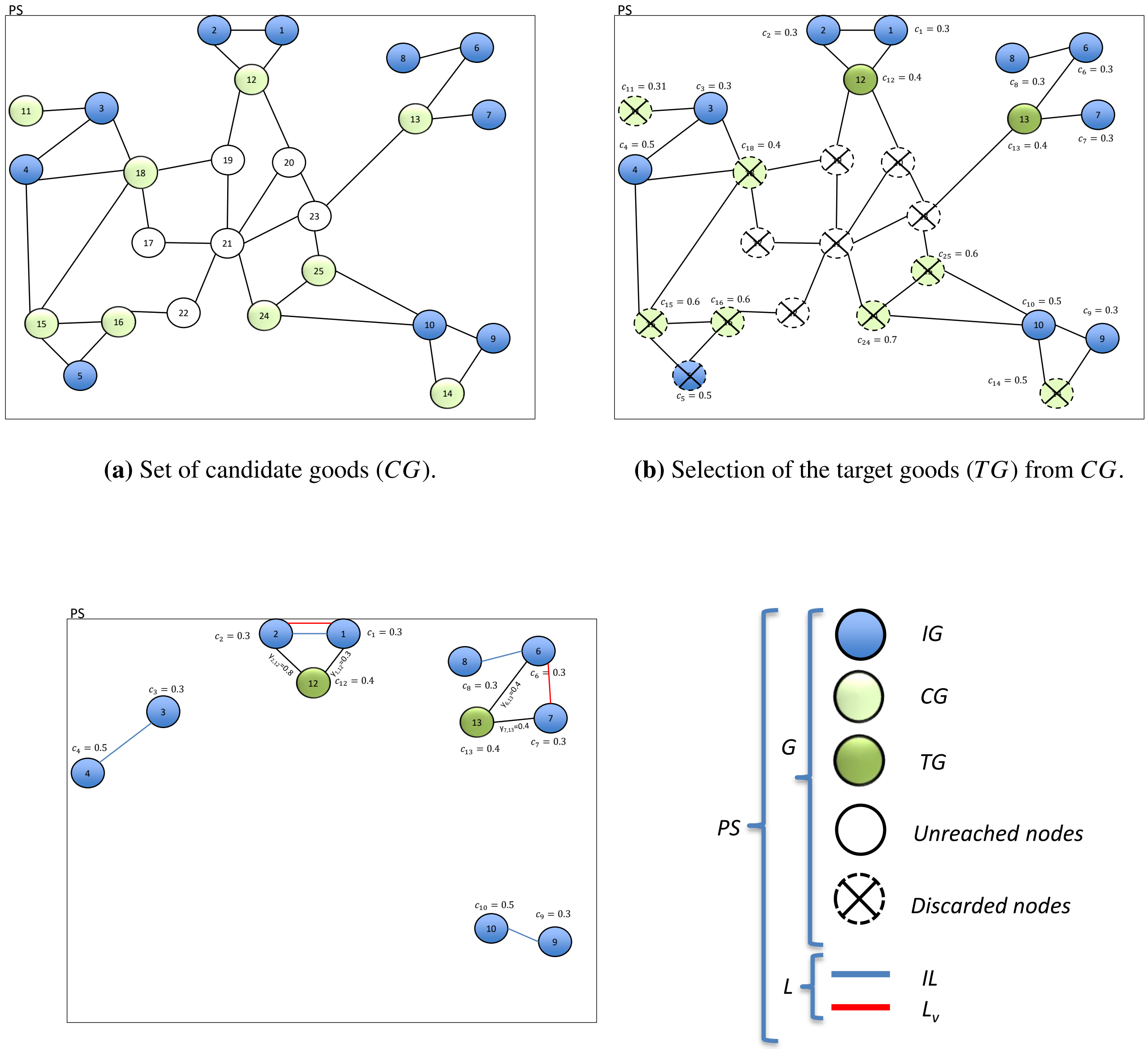
For future reference we will compress this process in the following two functions

$$LI = LI(r_{il}, c_v) \tag{3.2}$$

$$TG = TG(r_{il}, c_v) \tag{3.3}$$

[2] The notion of target set relies on two main principles which are supported by the general work of Hausmann and Hidalgo, (1) complex goods are desirable in the economy, and (2) the interaction among existent industries is the main driver to jump into new economic activities.

We are now in position to recall the p-IE problem.



(a) Set of candidate goods (CG).

(b) Selection of the target goods (TG) from CG.

(c) Final input network used by the p-IE.

(d) Nomenclature.

Figure 3.3: The first part corresponds to the selection of the first order neighbors CS in light green of the initial goods IG in blue. In the second part we select the nodes in CS that that has two or more neighbors in IG and whose complexity (c_v) is strictly greater than any of its neighbors in IG. Part 3 describes the final network after discarding all the nodes that do not have any interactions. Also, in this part we introduce the red links (L_v) connecting common neighbors of a given target good $v \in TG$ distinguishing them from the preexisting links IL between nodes in IG in blue. This part also shows explicitly the weights of the PS between nodes in IG with nodes in TG (y_{lv}). The fourth part summarizes the nomenclature used in for this selection process.

3.3.2 Problem statement of the p-IE

Let $I = \{1, \cdots, n\}$ be the set of areas of a country with n areas and let $p \in \{2, \cdots, n-1\}$ be the prescribed number of regions. Let Π be the set of feasible p-regionalizations given by contiguous partitions of I, i.e. collections of areas the form $P_p = \{R_1, \cdots, R_p\}$ such that

- $R \neq \emptyset$ for $R \in P_p$
- $R \cap R' = \emptyset$ for any pair of distinct elements $R, R' \in P_p$.
- $\bigcup_{R \in P_p} R = I$.
- Each region $R \in P_p$ is connected.

For the regionalization $P_p = \{R_1, \cdots, R_p\}$ the first term of the objective function of the p-IE is given by

$$\mathrm{FT}(P_p) = \sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{(l,m) \in LI} r_{li} r_{jm} t_{ij}^{lm,k} f(d_{ij}), \tag{3.4}$$

where the d_{ij} is the distance between area i and j. This term takes care of measuring in which extent the links IL generates intraregional collaboration between the existing industries and, therefore, guaranteeing the preservation of the main industrial interactions within each region.

On the other hand, the second term is given by

$$\mathrm{ST}(P_p) = \sum_{k=1}^{p} \sum_{v \in TG} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{(l,m) \in L_v} t_{ij}^{lm,k} (r_{il} y_{lv} + r_{jm} y_{mv}) f(d_{ij}) c_v, \tag{3.5}$$

where c_v is the ECI of the target node v normalized on the interval $[0, 1]$ and y_{lv} is the weight between l and v in the PS. We obtain the objective function from a convex combination of (3.4) and (3.5) so that the first term dominates the second (see Restrepo et al., 2020 for further details). Thus, we recall the objective function of the p-IE model

$$Z(P_p) = W_1 FT(P_p) + W_2 ST(P_p). \tag{3.6}$$

At last, we have that the p-IE problem amounts to determine the maximum of Z over the set of feasible regionalizations Π.

3.3.3 Complex contagion.

The innovation processes described by the PS has been thought as of the dynamic capabilities acquisition (know-how) followed by countries and regions (Hausmann and Hidalgo, 2011; Hidalgo et al., 2007b; Boschma et al., 2015; Neffke et al., 2011; O'Clery et al., 2018). These processes are highly path-dependent since the arising of a new capability requires the presence of several related economic activities requiring similar capabilities. Acquiring capabilities, in turn, is a systematic and nested process that takes place in industrial ecosystems where the R&D activities of the firms are constantly reinforced (Bustos et al., 2012; O'Clery et al., 2018). A type of phenomena that adjusts this characteristic is the so-called complex contagion diffusion (Centola, 2018; Centola and Macy, 2007). To put it in contrast, simple contagion models are those in which one node in a network infects easily all its neighbours. This kind of diffusion processes are common in the theory of disease spreading. On the other hand, the complex contagion models require the interaction of several nodes to reinforce the power of contagion so that they can diffuse into a new node. The nodes which have not been reached yet by the process are called inactivated (or healthy) nodes, whereas the nodes invaded by the process are called activated (or infected) nodes. In simple contagion, one node can infect others easily. In contrast, in complex contagion, the combined influence of many nodes are required to infect a new one.

In what follows, we will introduce the mathematical elements to define the free and guided evolutionary p-IE models. Intuitively, these two approaches are devised as alternatives to continue the contagion dynamic depicted in Figure 3.3 into further nodes. These alternatives will be thoroughly explained in Section 3.4. We start devising two new pieces of the model upon the ideas of the complex contagion models (Centola, 2018; Centola and Macy, 2007, Alshamsi et al., 2018) that are common to both evolutionary p-IE models. In the first piece of the model, we compute how likely it is (relatively) for the country to step into the nodes in the set TG (nodes 12 and 13 in Figure 3.3c). The second piece of the model, allows us to explain how these new nodes are incorporated spatially into the productive system of the country, i.e., how its production is distributed among the areas of the country. Besides these common elements, we present some new concepts exclusive of the guided evolutionary p-IE model. In contrast, with the target nodes (TG) of the p-IE model, we introduce the notion of long run target nodes that are not necessarily first order neighbors of the current state of the economy (IG). Thus, reaching long run target nodes in the PS require a dynamic strategy which guides the economy to make decisions foreseeing several steps into the future. This definition and the related mathematical issues are the main topic of the last part of this section.

Contagion dynamics

This first piece of the dynamic model is formulated as an extension of the model proposed in Alshamsi et al. (2018) to compute probabilities of activating nodes. In Alshamsi et al. (2018), they assume a non-directed network with adjacency matrix (a_{lm}), i.e. $a_{lm} = 1$ if nodes l and m are linked in the network and zero otherwise. Under this setting the probability of activating a note k is given by

$$p_k = B \left(\frac{\sum_l A_l a_{lk}}{\sum_l a_{lk}} \right)^{\alpha}, \tag{3.7}$$

where $A_l = 1$ if the node l is activated and zero otherwise and $B, \alpha > 0$ are parameters. In here B is merely a scaling parameter, whereas α plays a central role in the modeling. Notice that the fraction in (3.7) represents the ratio of neighbors of k already activated; thus, for large values of α compared to one (the model is superlinear) the probability of activating k is higher the larger the number of activated neighbors.

In contrast, if α is close to zero the activation probability is less sensitive to the number of activated neighbors of k and approaches the case of simple contagion. Therefore, we can regard α as the parameter that controls how complex the contagion process is. For the case of the PS, in Alshamsi et al. (2018) it is found a slightly superlinear behaviour with $\alpha = 1.03$ and with $B = 0.16$.

We start generalizing this idea in a straightforward way to the case in which we assume a directed network with weights. We recall that the PS used for the p-IE is the same introduced in Hidalgo et al. (2007b), which comes from taking the minimum between the weights of the links joining two nodes in both directions. In this case, we opt to keep the directed version of the PS to stress the fact that the innovation processes follow a directed dynamics. We will call this network the directed PS and we will denote its (weighted) adjacency matrix by $W = (w_{lm})$. Notice that the links of the undirected and directed versions of the PS are related by the expression $y_{lm} = \min\{w_{lm}, w_{ml}\}$. With these ingredients in mind, we define the *intensity of contagion* of a node k as follows:

$$I_k = \left(\frac{\sum_l^k A_l w_{lk}}{\sum_l w_{lk}} \right)^{\alpha} . \tag{3.8}$$

Notice that, in this case, we dropped the term B because we are not longer interpreting I_k as probability, but as an intensity. We can regard w_{lk} as the transmission intensity from node i to node k. Hence, the intensity of contagion of k is a superlinear power of the fraction of the transmission intensities provided by the activated neighbors of k. We will use the numbers I_k in two ways, (1) to rank the target nodes for an economy according to their intensity of contagion (this one does not require the power α either), and (2) to compute the probability of a good to enter into the economy, relative to the probability of the other goods that are target to enter during the same period. The ranking proposed is deceasing in I_k, regarding target nodes k with high I_k as the most likely to be develop in the economy. Additionally, since all the computations are carried in a relative way, we can dispense with the incorporation of the factor B. In Figure 3.4 we show how to use equation (3.8) for the computation of the intensities of contagion for the target goods obtained in Figure (3.3c).

Updating of distribution vectors

The introduction of a new activity into the p-IE requires information of the relative spatial distribution of the good inside the country. Parallel to this, the entering of new activities into the country and the interaction among the existing industries also induces changes in the distribution of the nodes in IG. These considerations motivate the second application of the complex diffusion problem: the evolution of the distribution of the goods in the country. In terms of the notation introduced at the beginning of this section, the aim of this part consists in defining the evolution of the quantities r_{li} which corresponds to the share of area i in the production (exports) of good l for both, entering and existing activities. In this model, we are not admitting the possibility of deactivation of nodes (equivalently, industry disappearances). Henceforth, an industry can only become relatively more or less relevant with respect to other areas inside the same country. Under this framework we propose a scheme to update the distribution vectors r_i in two steps, in the first step, we define the vector of raw distributions in the period $t + 1$ as

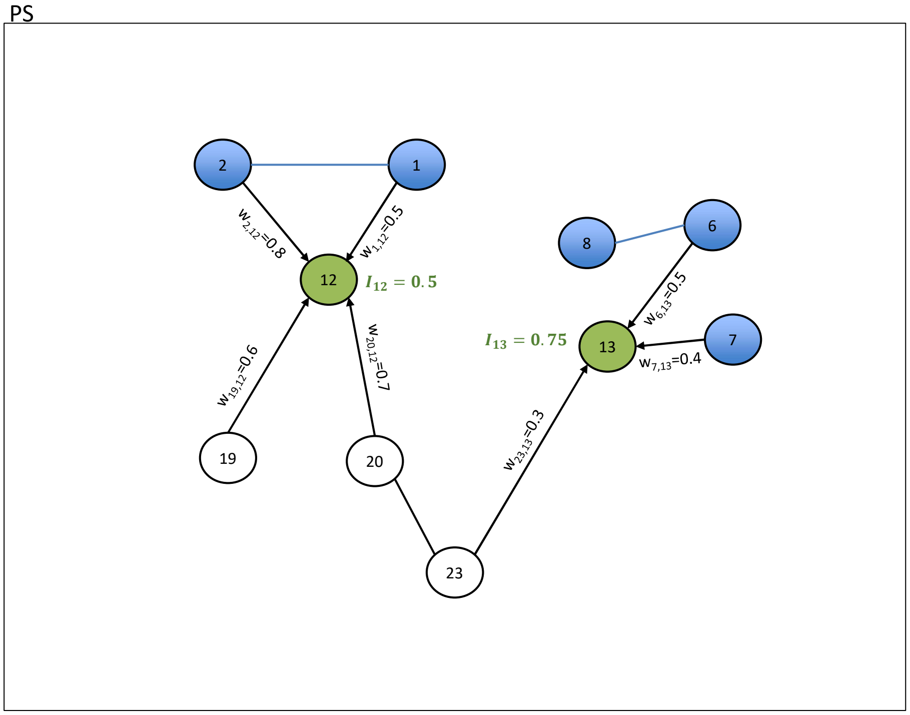


Figure 3.4: This figure zooms into the upper right part of the PS rendered in Figure (3.3). We change some links by arrows to stress the directed nature of the computation of the intensities. Notice that both, the active (in blue) and the inactive set (in white) play a role when computing the intensities. In this case, target 13 is ranked above target 12, which intensities of contagion are 0.75 and 0.5, respectively.

$$f_{li,t+1} = r_{li,t} + q_{li,t},$$

where $q_{il,t}$ is the new production of the good l in the area i induced by the influence of the related industries within the same region. We propose this autoregressive scheme to guarantee a coherent evolution of the geographic distribution of the activities. For the determination of $q_{li,t}$ we combine the spatial features of the p-IE model and the philosophy behind the complex contagion model, namely (3.7). Since the diffusion process is supposed to take place after the definition of the regions obtained from the p-IE model, we may assume that each one of the areas i belong to a region R_j of a regionalization with p regions P_p. We also recall that the distance decay function f, introduced in equation (3.1), defines the level of impedance between areas i and j as $f(d_{ij})$. We expect industries belonging to areas with negative impedance to not interact with one another. Thus, given areas i and j, we define the geographic interaction between them $G_{ij} = f(d_{ij})$ if $f(d_{ij}) > 0$ and if i and j belong to the same region in P_p, and zero otherwise. Notice that the term G_{ij} captures both, distance and border effects. Finally, we define

$$q_{ki,t} = \left(\frac{\sum_{l,j}^{k} G_{ij} r_{lj,t} w_{lk}}{\sum_l w_{lk}} \right)^{\alpha} . \tag{3.9}$$

We notice first that the quotient inside the power α in the previous expression is always between $(0, 1)$. This quantity encompasses three concepts: geographic effects G_{ij}, spatial distribution of the nodes $r_{li,t}$ and transmission intensity w_{lk}. So, the new production of a node k in the area i is closer to one as long as the region produces with high intensity $(r_{jl,t})$ all the neighbors of l in the directed PS in areas j nearby to i according to the geographic effect G_{ij}.

For future references we compress the updating processes of the production using the notation

$$r_{li,t+1} = \text{UpdateProduction}\left(r_{li,t}, TG, P_p\right) . \tag{3.10}$$

Finally, we normalize the vectors $f_{l,t+1} = (f_{li})_{\{i=1,\cdots,n\}}$ so that its components add up one. Thus, we define

$$r_{l,t+1} = \frac{1}{\sum_{i=1}^{n} f_{li,t+1}} f_{l,t+1} .$$

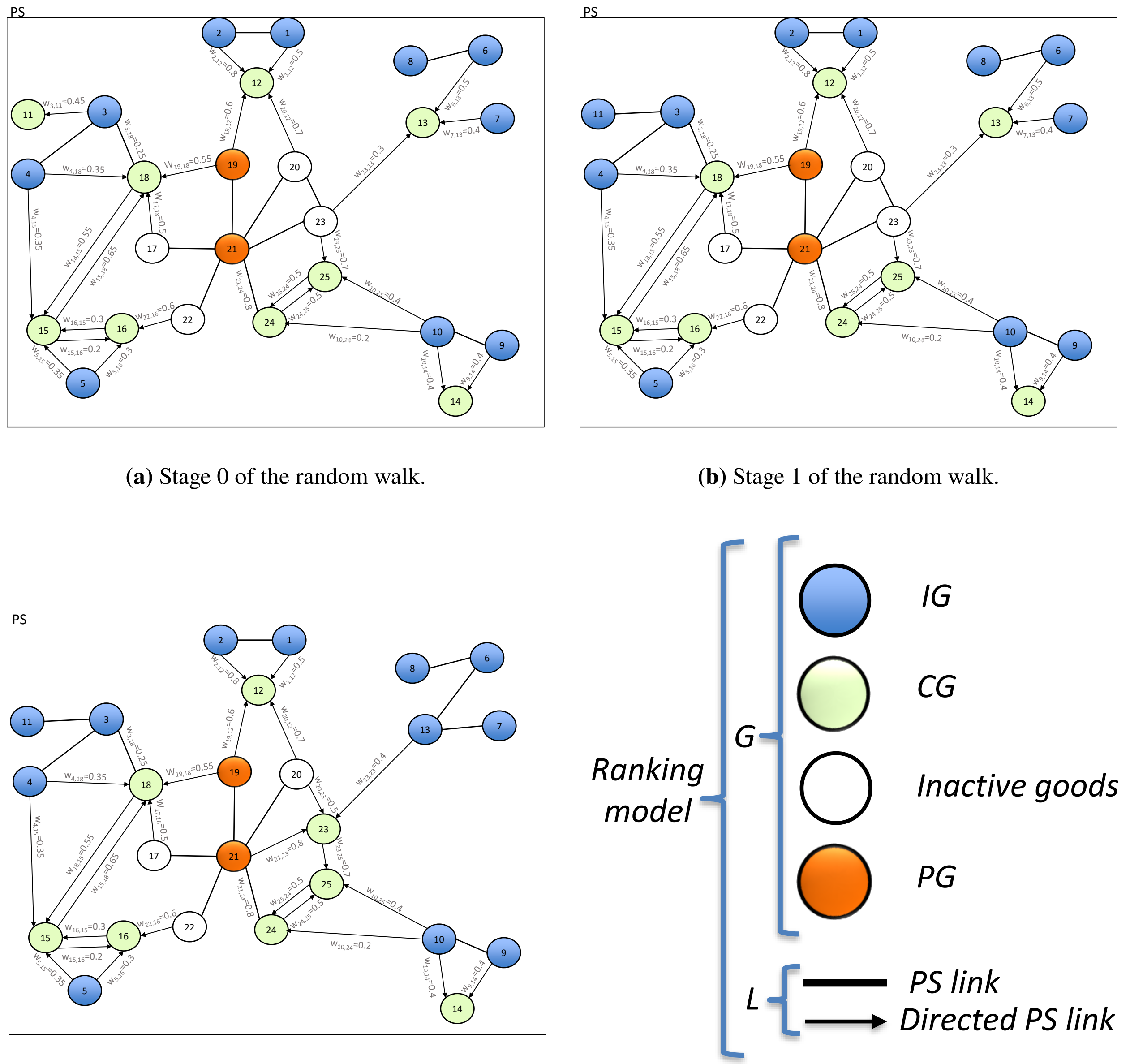
We conclude this part pointing out that the evolution of the output described by equation (3.9) does not pretend to forecast the factual production patterns, since it leaves aside demand and external shocks to the economy. In fact, this equation is intended to describe the dynamics of innovation and relative specialization of regions inside a given country. Thus, its results capture the nuances of the innovation theory of Hausmann and Hidalgo in a regional setting and must be interpreted accordingly.

3.3.4 Long run target nodes and definition of stepping-stones

In this part, we introduce theoretical elements of the guided evolutionary p-IE model which will be introduced in Section 3.4. The guided evolutionary p-IE model also starts with a set of initial goods IG just like the p-IE and tries to reach an objective or long run target set defined by a policymaker, we call this set of nodes Policy Goods (PG). The guided version of the model identifies the optimal diversification route in terms of regionalizations and innovation patterns a country has to follow to reach the set PG in the shortest time, given a set of initial goods IG and a set of long run targets PG.

This problem, as we mentioned in Section 3.2, is central to define diversification strategies that foster economic growth among other interests of the policy makers. In particular, in Alshamsi et al. (2018) several diversification strategies are explored in the context of complex contagion such as targeting nodes with high probability of being activated, nodes with large degree centrality, nodes with low degree centrality, etc. Our approach combines elements of complex contagion and the spatial components of the p-IE model to determine optimal activation routes of nodes to reach the long run targets; these routes are the so called stepping-stones.

More explicitly, our diversification strategy consists in defining a ranking of the stepping-stones to reach the nodes PG using equation (3.8) in a probabilistic way. Let us illustrate this idea with Figure 3.5. Given a node $v \in PG$ (either 19 or 21) we consider a random walk starting at the set of initial goods IG trying to diffuse to v through the intensities defined in (3.8). This process, at each step, selects one of the neighbors of the active set (which is IG at stage 0) to be activated randomly. This set corresponds to CG which are in Figure 3.5a. For each $k \in CG$ we compute its intensity I_k. Notice that $11, 14 \in CG$ have the highest probability of activation (which is one) because they are only connected to activated nodes. In the second stage, the process randomly infects only one of the elements in CG with a probability proportional to I_k. In the case of Figure 3.5b 11 is activated, which is plausible given that $I_{11} = 1$. From here we see that the random walk may follow inconsequential paths, like invading 11 which is a leaf of the network. Figure 3.5c shows the second stage of the random walk in which node 13, with intensity $I_{12} = 0.75$, is activated. The random walk continues this process until it reaches the node v.



(a) Stage 0 of the random walk.

(b) Stage 1 of the random walk.

(c) Stage 2 of the ranking walk.

(d) Nomenclature.

Figure 3.5: In the first part we provide a full description of the initial stage of the random walk showing the initial set IG colored in blue, the set of the first order neighbors of IG, CG in yellow and the Policy goods PG. This part also shows the directed links of the PS with its respective weights (w_{lm}) from which the probabilities to jump into the elements of CG are computed. The second part corresponds to the second stage of the random walk in which the node 11 is activated. In the first stage the node 13 is activated implying that node 23 becomes a candidate good. Part four summarizes the nomenclature used in the example.

Although single paths may be inefficient like the one rendered in Figure 3.5, the repetition of this process tends to create paths that are, in average, more likely to occur (and therefore, more efficient). We repeat the process for each $v \in PG$ several times to find the paths which connects more likely the set IG with the nodes in PG (see Section 3.4 and 3.5 for more details about the number of simulations). Using these simulated paths as inputs we define the stepping-stones for the guided model. The ranking process have the following two steps

- We rank the nodes in PG. Since the main goal of the country consists in developing as many activities in PG as possible, the nodes of this set with more appearances in the paths followed by the random walk are more relevant. Thus, we assign a value to each node $v \in PG$ given by its relative frequency in all the simulated paths.

- For each step of the diffusion we only rank the first order neighbors of the available set, i.e,. CG. We rank these nodes looking at the relative frequency of all the nodes in which the random walk started its travel from IG to v PG. Thus, each node $v \in PG$ has an associated list of first stepping-stones ranked by its relative frequency in the number of appearances as a first step of the random walk. Then we compute for each $g \in CG$ the sum of the values assigned to it by each node $v \in PG$ weighted by the value of the corresponding v, which we denote as IP_g^0. Finally, we define the *innovation potential index* (IP_g) of g as follows

$$IP_g = \frac{IP_g^0}{\sum_{l \in CG} IP_l^0}. \tag{3.11}$$

  For future reference we compress the computation of IP as a function of the number of simulations of the complex contagion random walks ($numsimul$), PG and IL in the following way

$$IP = IP(IL, PG, numsimul). \tag{3.12}$$

The main advantages of this approach are, on the one hand, it allows to assess the value of a node as stepping-stone to reach the policy goals and, on the other hand, it is more flexible, incorporating alternatives between different diversification strategies so that countries can find suitable diversification paths that fit their production or market constraints and guide them, at the same time, through an efficient route to reach their innovation goals.

We finish this section stressing that the paths followed by the random walks are not the conventional paths joining pairs of nodes defined in graph theory, but the list of nodes sequentially activated up to the activation of any of the elements in PG.

3.4 The model

In this section, we explain how the p-IE model is introduced in a dynamic framework; this is the free and directed evolutionary p-IE models. Since both models uses intensively certain components of the heuristic solution of the p-IE, we start rewriting its pseudo-code in a compact way in Algorithm 3.4.1. More explicitly, both, the free and the directed model we present in this section use intensively the local search of the p-IE heuristic. The local search uses a tabu search algorithm explained in Restrepo et al., 2020 (see also Glover, 1977, Glover, 1989, Glover, 1990 and Rosenthal and Strange, 2004). Given a feasible solution, the neighboring solutions, N^s, are obtained by moving bordering areas (i.e., areas that share a border with a neighboring region) to neighboring regions, one at a time, while preserving feasibility. The tabu search explores these neighboring solutions seeking improvements in the objective function. One key aspect of the tabu search is that it allows for temporal worsening of the objective function as a strategy for escaping from local optima. The search stops after a predefined number of non-improving moves (convTabu).

```
Algorithm 3.4.1: pIE (
$A$ : Set of areas,
$p$ : Number of regions,
$maxitr$ : Number of initial feasible solutions to generate,
$r_{li}$ : Production of good $l$ in area $i$,
$weight_v$ : Weight of the target good $v$,
$lengthTabu$ : Length of the tabu list,
$convTabu$ : Number of non-improving moves before stop,)

Comment: Aggregate $n$ areas into $p$ spatially contiguous regions such that
$Z(S_p)$ is maximized.

$LI = LI(r_{il}, c_v)$
$TG = TG(r_{il}, c_v)$
$P_p^{best} = \emptyset$, best partition.
Construction Phase: Grow regions from seed areas
for $i = 1, 2, \cdots, maxitr$
  do { $seeds$ = k-means++($A, p, r_{li}, W$)
       $S_p$ = GrowRegions($seeds, A, r_{li}, W,$)
       if $Z(S_p) > Z(P_p^{best})$
         then { $P_p^{best} = S_p$
Local Search Phase: Tabu search

  do { $P_p^*$ = Tabusearch $\left(P_p^{best}, r_{il}, d_{ij}, IL, TG, weight_v,\ lengthTabu,\ convTabu\right)$
return $(P)_p^*$
```

3.4.1 The free p-Innovation Ecosystems Evolutionary model

After finding the optimal solution for a given value of p, the parameters of the model are updated recursively depending on the initial solution and the number of targets invaded. The dynamic of the production parameters ($r_{il,t}$) is ruled by equation (3.10), whereas the amount of goods introduced into the economy before running Algorithm 3.4.1 is determined by a recursive process described in Algorithm 3.4.2. Briefly, the target goods are allowed to enter in order, one by one, depending on their probability to appear in the economy according to equation (3.8); then, we run the local search part of Algorithm 3.4.1 with convTabu=1 until we find any change in the current solution.

```
Algorithm 3.4.2: $p$-IIE (
$A$ : Set of areas,
$p$ : Number of regions,
$maxitr$ : Number of initial feasible solutions to generate,
$r_{li,0}$ : Production of good $l$ in area $i$ at the initial period,
$c_v$ : Complexity of the target good $v$,
$lengthTabu$ : Length of the tabu list,
$convTabu$ : Number of non-improving moves before stop,
$numper$ : Number of periods,)

  Comment: Aggregate $n$ areas into $p$ recursively during $numper$ periods such
  that $Z(S_p)$ is maximized.

  Initial solution
  $LI_0 = LI(r_{il,0}, c_v)$
  $TG_0 = TG(r_{il,0}, c_v)$
  $P_p^{best,0} = pIE(A, p, maxitr, r_{il,0}, IL_0, TG_0, c_v, lengthTabu, convTabu).$
```

```
**Evolution of the solution**
for $t = 1, 2, \cdots, numper$
    do
        $P_p^{best,t} = P_p^{best,t-1}$
        $TG_t^{aux} = \emptyset$
        while $P_p^{best,t} == P_p^{best,t-1}$ or $TG_t^{aux} \neq TG_{t-1}$
            do
                $v = \text{argmax}_{v \in TG_t \setminus TG_t^{aux}} \{I_v\}$
                $TG_t^{aux} = TG_t^{aux} \cup \{v\}$
                $r_{il,t}^{aux} = \text{UpdateProduction}\left(r_{il,t-1}, TG_t^{aux}, P_p^{best,t-1}\right)$
                $LI_t = LI\left(r_{il,t}^{aux}, c_v\right)$
                $TG_t = TG\left(r_{il,t}^{aux}, c_v\right)$
                $P_p^{best,t} = \textbf{Tabusearch}\left(P_p^{best,t-1}, r_{il,t}^{aux}, IL_t, TG_t, c_v, 1, 1\right)$
        $r_{il,t} = \text{UpdateProduction}\left(r_{il,t-1}, TG_t^{aux}, P_p^{best,t-1}\right)$
        $P_p^{best,t} = \textbf{Tabusearch}\left(P_p^{best,t-1}, r_{il,t}^{aux}, IL_t, TG_t, c_v,\ leghtTabu,\ convTabu\right)$
return $(P)_p^{best,t}$
return $(r)_{il,t}$
```

3.4.2 The guided p-Innovation Ecosystems Evolutionary model

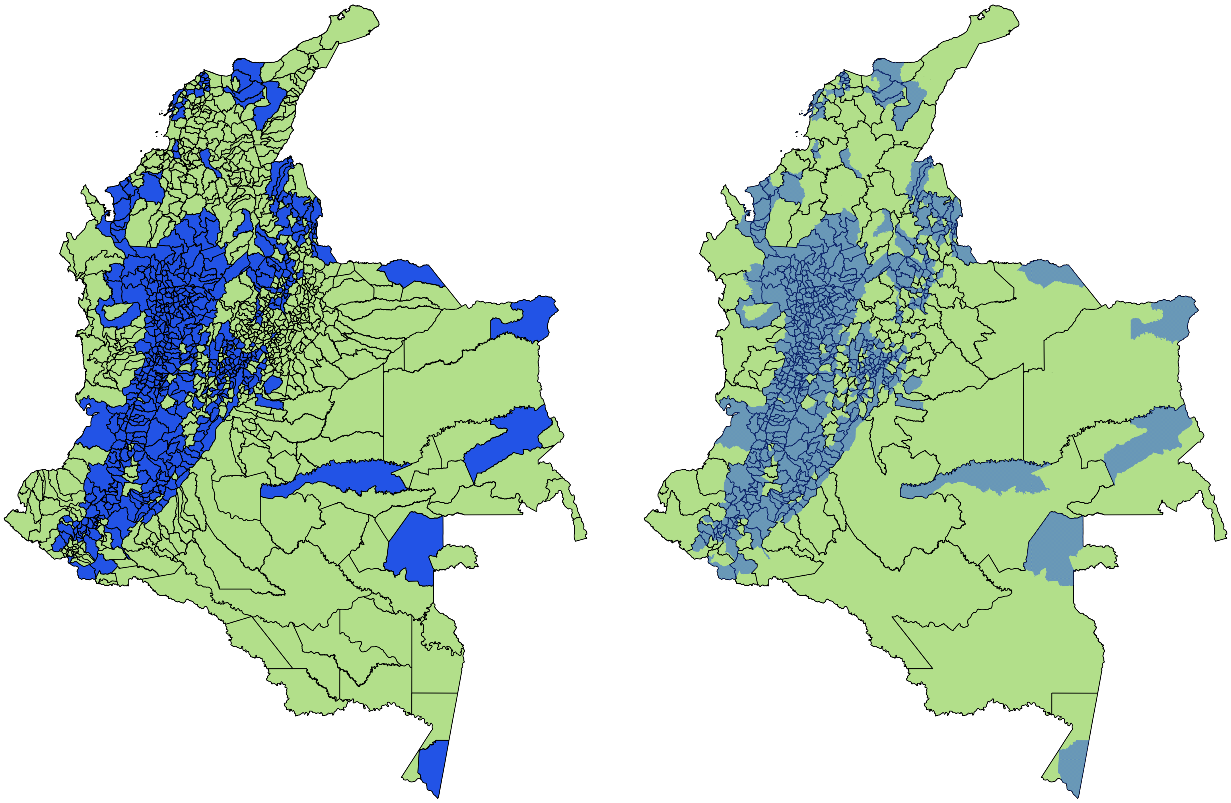
The main differences between the guided and the free evolutionary p-IE models are the introduction of the long run target goods or policy goods (PG) and the replacement of the complexity (c_v) as the weight of the candidate goods per period by the innovation potential (IP). A flexibility of this alternative model that is not explored in this paper are the alternatives to determine ideal candidate goods. Nonetheless, for the sake of the comparability with the free evolutionary p-IE model we opt to regard as candidate goods only the ones that still ensure the complexity improvement with respect to the current state of the economy. An implication of this is that this model also guarantees the complexity improvement of the global economy. In contrast with the complexity, the innovation capacity is a dynamic measure, since its definition depends on the current state of the economy (IL). This implies that the innovation capacity must be updated per simulated period. Analogously to Algorithm 3.4.2, Algorithm 3.4.4 also determines endogenously the length of each period using the local search of Algorithm 3.4.1 and the ranking index introduced in equation (3.8).

```
Algorithm 3.4.4: $p$-IIE-GUIDED (
$A$ : Set of areas,
$p$ : Number of regions,
$A$ : Set of areas,
$p$ : Number of regions,
$maxitr$ : Number of initial feasible solutions to generate,
$r_{li,0}$ : Production of good $l$ in area $i$ at the initial period,
$c_v$ : Complexity of the target good $v$,
$lengthTabu$ : Length of the tabu list,
$convTabu$ : Number of non-improving moves before stop,
$numper$ : Number of periods,
$numsimul$ : Number of simulations of the random walkers,
$PG$ : Long run target nodes or policy goods,)

Comment: Aggregate $n$ areas into $p$ recursively during $numper$ periods such
that $Z(S_p)$ is maximized.

Initial solution
$LI_0 = LI(r_{il,0}, c_v)$
$TG_0 = TG(r_{il,0}, c_v)$
$IP_v^0 = IP(IL_0, PG, numsimul)$
$P_p^{best,0} = pIE(A, p, maxitr, r_{il,0}, IL_0, TG_0, IP_v^0, lengthTabu, convTabu)$.

  Evolution of the solution
  for $t = 1, 2, \cdots, numper$
    do
      $P_p^{best,t} = P_p^{best,t-1}$
      $TG_t^{aux} = \emptyset$
      while $P_p^{best,t} == P_p^{best,t-1}$ or $TG_t^{aux} \neq TG_{t-1}$
        do
          $v = \text{argmax}_{v \in TG_t \setminus TG_t^{aux}} \{I_v\}$
          $TG_t^{aux} = TG_t^{aux} \cup \{v\}$
          $r_{il,t}^{aux} = \text{UpdateProduction}\left(r_{il,t-1}, TG_t^{aux}, P_p^{best,t-1}\right)$
          $LI_t = LI\left(r_{il,t}^{aux}, c_v\right)$
          $TG_t = TG\left(r_{il,t}^{aux}, c_v\right)$
          $P_p^{best,t} = \textbf{Tabusearch}\left(P_p^{best,t-1}, r_{il,t}^{aux}, IL_t, TG_t, IP_v^{t-1}, 1, 1\right)$
      $r_{il,t} = \text{UpdateProduction}\left(r_{il,t-1}, TG_t^{aux}, P_p^{best,t-1}\right)$
      $P_p^{best,t} = \textbf{Tabusearch}\left(P_p^{best,t-1}, r_{il,t}^{aux}, IL_t, TG_t, IP_v^{t-1}, leghtTabu, convTabu\right)$
      $IP_v^t = IP(LI_t, TG_t, numsimul)$
  return $(P)_p^{best,t}$
  return $(r)_{il,t}$
```

3.5 Study Case: Colombia

3.5.1 Geographic data

As an application of the ideas introduced in this paper, we study the dynamics of complexity and innovation in Colombia. The georeferenced data required to find the distribution vectors of the nodes in the country was obtained from the open source information in Datlas tool of Bancoldex. This data set contains the exports per industry in HS4 and per municipality for the years 2008-2017. We use information of 2010 for the computation of the distribution vectors r_l for each good l in which Colombia has relative comparative advantage, i.e., Relative Comparative Advantage strictly greater than 1 (see Hidalgo et al., 2007b, Balassa, 1964). In this work, we opt for a more stringent criteria than in Restrepo et al. (2020) to define the effective production of Colombia. In this case, we only consider goods belonging to the giant component of the PS in which Colombia has RCA strictly greater than 1. This condition leaves us with 494 out of 1120 municipalities with effective production. As in Restrepo et al. (2020), we use the Max-p algorithm (Duque et al., 2012) to aggregate the Colombian areas such that the municipalities with zero exports are merged in regions with one leading municipality with positive exports. We depict the areas with effective production highlighted in blue (Figure 3.6a) and the resultant Max-p areas in (Figure 3.6b). The latter areas are going to be used as inputs into the model.



(a) Municipalities with effective production. **(b)** Areas obtained after applying Max-p regions.

Figure 3.6: Figure 3.6a shows the map of Colombia divided in 1120 municipalities, where 494 of them (marked in blue) export goods in which Colombia has relative advantage and belong to the giant component of the PS. Figure 3.6b depicts the regions obtained after applying the Max-p Regions algorithm to Figure 3.6b.

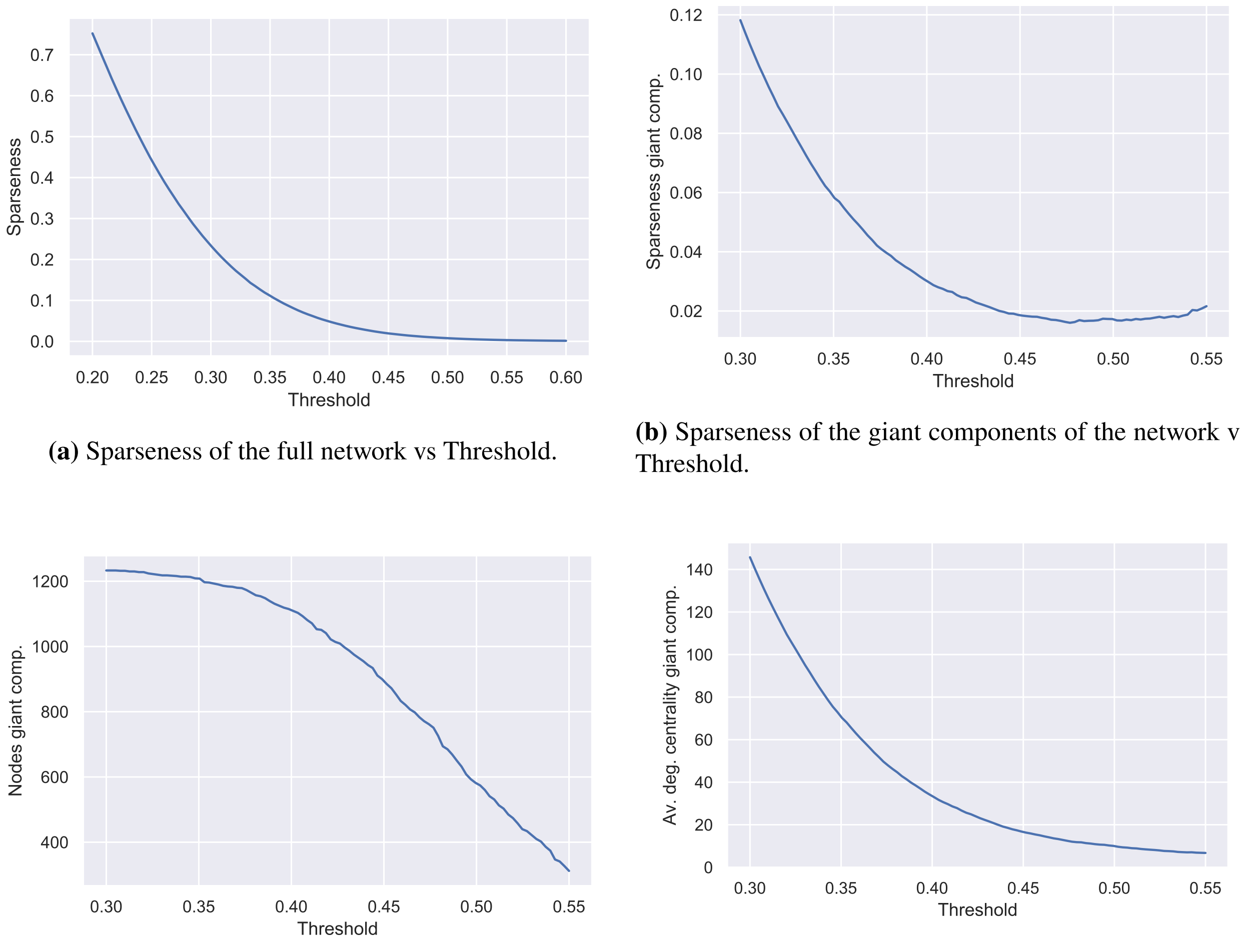
The distances d_{ij} are computed as the Euclidean distance between the centroids of the resulting 494 areas. We determine the decaying parameter α in the decaying function (3.1) analogously as in Restrepo et al. (2020). We recall that the parameter α is chosen such that the distance decaying function f vanishes at the radius of an average region of size $\frac{n}{p}$. Therefore, if we assume optimal regions circular shaped

$$\alpha = \frac{ln(2)}{ln\left(\frac{M}{R_{av}}\right)},$$

with $R_{av} = \sqrt{\frac{n}{\pi p}}$. Thus, since the maximal distance among regions is $M = 1,382.2$ km, we have that α only depends on p.

3.5.2 PS data

We compute a version of the PS using trade data from the Centre dÉtudes Prospectives et dInformations Internationales (CEPII), which contains data for 128 countries, for the time range 1995-2010 and includes observations of 1,240 products classified under the nomenclature of the Harmonized System at 4 digit level (HS4). The directed and undirected weights are computed as in Hidalgo et al. (2007b). Given the dynamic nature of our study, we compute the weights of the PS using data for the full window time of our data set. Figure 3.7 summarizes the diagnostics for the determination of the threshold of the PS. Notice that we focus our analysis in the giant component of the PS to guarantee the good behavior of the diffusion processes on it. Aside of this, we remark that the choosing of a lower threshold (0.45) than in Hidalgo et al. (2007b) and Restrepo et al. (2020) leads to a less sparse network with higher degree centrality. We require a bigger network so that the diffusion range of Colombia in the PS enable us to study the innovation dynamics for several number of periods.



(a) Sparseness of the full network vs Threshold.

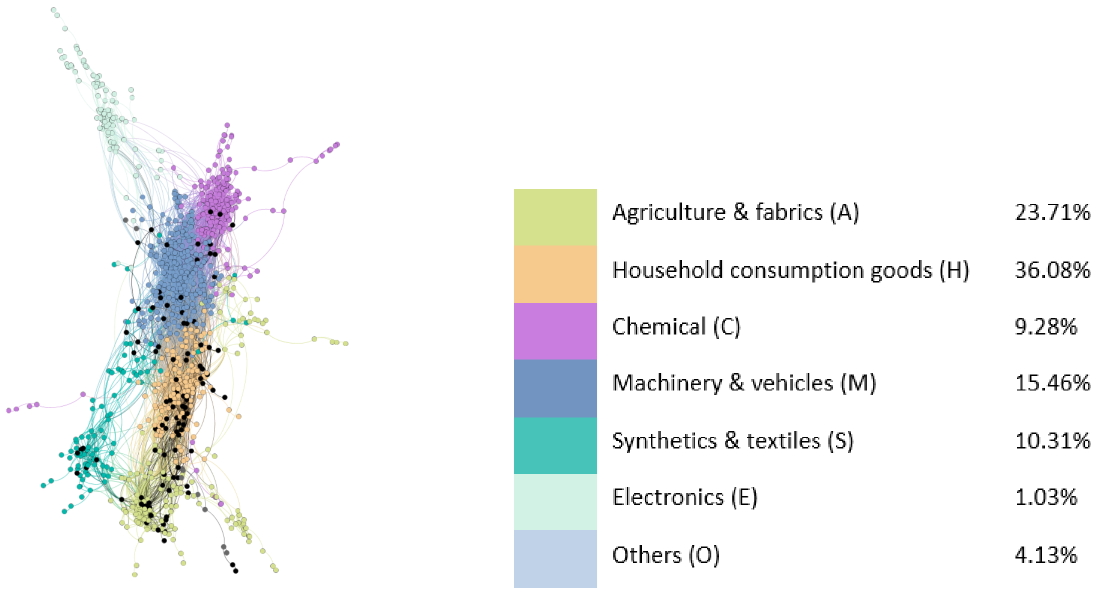
(b) Sparseness of the giant components of the network vs Threshold.

(c) Number of nodes of the giant component vs Threshold.

(d) Average degree centrality of the giant component vs Threshold.

Figure 3.7: The first two diagrams show how the sparseness of the network and the sparseness of the giant component of the network, respectively, decay rapidly as we increase the threshold. In contrast, the third chart depicts a slower decay of the number of nodes (namely, in a concave way). Lastly, the fourth chart shows the decaying of the average degree centrality of the network. We select a threshold of 0.45. This network has an sparseness of 0.0194, the sparseness of it connected component is 0.0185, the number of nodes is the giant component is 896 and the average degree of the giant component is 16.58.

With the threshold of 0.45 the number of nodes of our PS is 896. Colombia in 2010 had comparative advantage in 144 goods, from which 97 belong to the connected component of the PS. We divide the PS in 7 categories: Agriculture & fabrics (A), Household consumption goods (H), Chemical (C), Machinery & vehicles (M), Electronics (E), Synthetics & textiles (S) and Others (O). The first 6 categories amount to the 98.5% of the network and are obtained from merging some communities computed through the classical Louvain Modularity algorithm (Blondel et al., 2008). The distribution of Colombia in the PS in 2010 and the nomenclature for the categories are summarized in Figure 3.8b



🟨	Agriculture & fabrics (A)	23.71%
🟧	Household consumption goods (H)	36.08%
🟪	Chemical (C)	9.28%
🟦	Machinery & vehicles (M)	15.46%
🟩	Synthetics & textiles (S)	10.31%
⬜	Electronics (E)	1.03%
🟦	Others (O)	4.13%

(a) Initial location of Colombia on the PS.

(b) Share of Colombian nodes in the PS.

Figure 3.8: In Figure 3.8a we depict the communities detected with the Louvains method in the giant component of the PS and we mark in black, on top of them, the 97 nodes in which Colombia reveal comparative advantage. The color code of the categories and the share of Colombia in each one of them is summarized in Figure **??**.

3.5.3 Application of the evolutionary p-IE models for the Colombian case

As described in Section 3.4 we start implementing the p-IE for the initial period, in our case 2010. We find the optimal value of p implementing the same strategy as Restrepo et al. (2020). This is, we run the model for several values of p as it is shown in Figure 3.9. Thus, from the Elbow method we select p=14 as our first candidate for the optimal value of p.

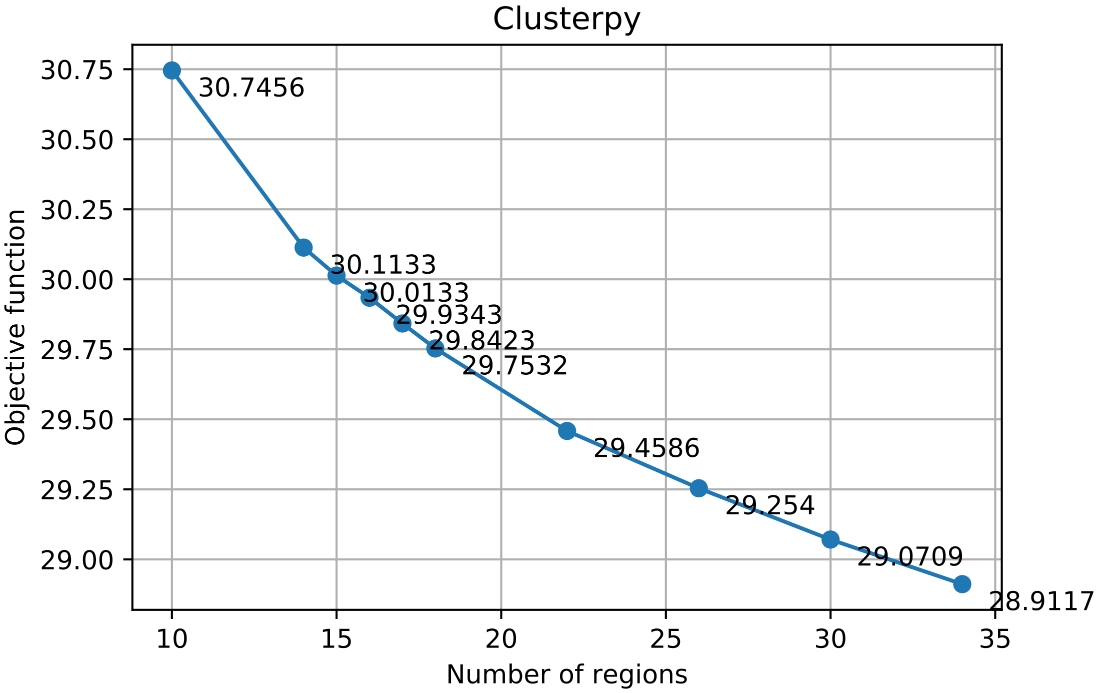
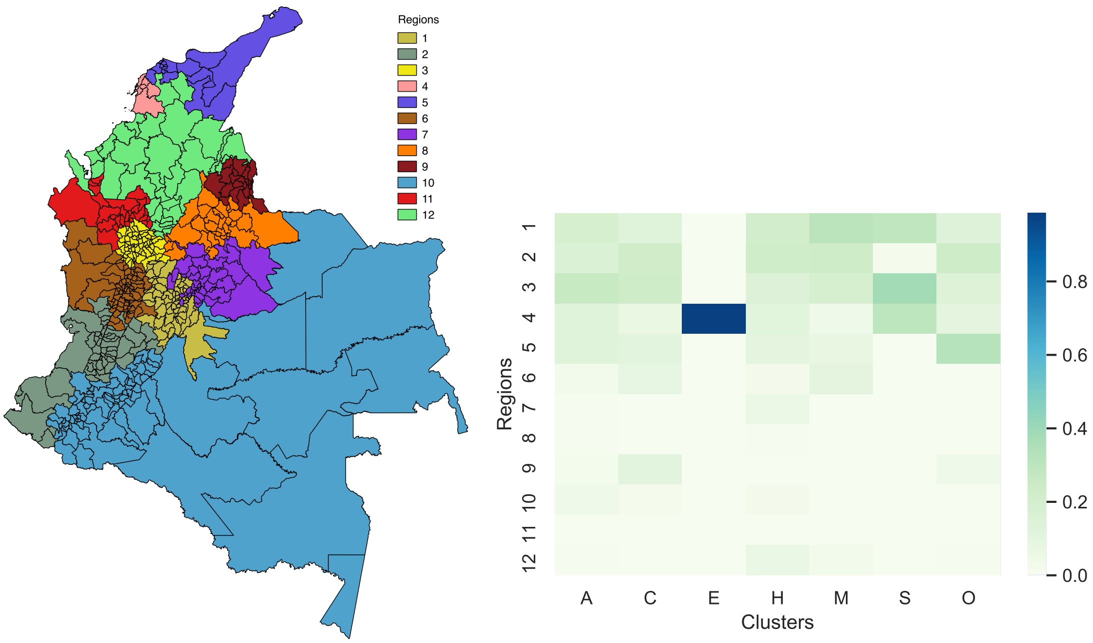


Figure 3.9: This figure shows how the optimal value of the objective function is convex in the number of regions and depicts how the slope starts flattening significantly faster for big values of p. We deduce from the change in the slopes that the Elbow occurs at $p = 14$.

However, as in the case of the p-IE model, our model can produce optimal regionalizations in which some regions have only one area (singletons). This outcome is not desirable because the emergence of singletons corresponds to solutions that do not take advantage of the possible interaction between regions (for a further discussion see Restrepo et al., 2020). In our case, the solution for $p = 14$ has 2 singletons. We overcome this issue running the algorithm for $p = 12$, i.e., the optimal p according to the Elbow method minus the number of singletons in the optimal solution. The solution for this value of p is rendered in Figure 3.10 together with the distribution of the production of the clusters per region in the heat map presented in Figure 3.10b.



(a) Solution of the p-IE model for $p = 12$.

(b) Distribution of the production per regions.

Figure 3.10: Figure 3.10a shows the initial regional configuration, corresponding to the solution for $p = 12$ of the static model. Figure 3.10b represents the distribution of the production of the clusters per region in the country. This heat-map is normalized per column such that each one of the adds up 1. For example, the 98.3% of the production of the only good belonging to the Electronics cluster is produced in region 4 (Styrene Polymers); or the 4 goods of the category in which Colombia has RCA are intensively produced in region 1 (Aluminium and copper scraps and prepared equine and bovine hides) with 43% of the national production of such goods.

The order chosen for the regions obeys a notion of complexity already introduced in Restrepo et al. (2020). The relative regional complexity is defined as the weighted sum of the complexities of the complex goods produced within a region P. Explicitly, if μ is the mean of the complexities c_l of all goods traded in the world

$$C_P = \sum_{\{l:\, c_l > \mu\}} \sum_{i \in P} r_{il} c_l. \tag{3.13}$$

Thus, a region has a high complexity if it is relatively intensive (with respect to the other regions of the country) in the production of goods whose complexity is above the average. This measure has a natural interpretation in terms of the PCI (Prodcut Complexity Index) introduced in Hausmann et al. (2014). Since the complexity index of a good l (c_l) is a linear transformation of the PCI, our measure is proportional to a weighted average of goods whose PCI is positive. This and other descriptive measures of the regions are summarized in Table 3.1. This table also contains information about the share in the national exports per region, the number of areas (after applying the Max-p regions algorithm) of each region, the extension in Km^2 per region, the number of regions in which Colombia has relative comparative advantage operating within the region and the relative regional complexity.

Region Id	Share	Number of areas	Area (Km^2)	Number of industries	Complexity
1	10.82%	62	35,514,730	95	6.6717
2	10.41%	48	67,737,886	80	6.1108
3	9.84%	64	14,333,749	83	5.8031
4	30.49%	4	7,159,731	53	4.4138
5	5.51%	15	46,504,913	78	3.0283
6	10.86%	71	47,605,505	68	2.1869
7	2.83%	37	51,107,579	54	1.5878
8	7.75%	30	43,957,739	32	0.1394
9	0.46%	20	16,136,694	39	0.1276
10	10.66%	82	558,369,739	38	0.0743
11	0.07%	29	31,411,544	33	0.0037
12	0.28%	32	106,888,437	34	0.0001

Table 3.1: Description of the initial solution of the p-IE for 12 regions.

The strong influence of commodities in the Colombian economy is reflected in the exports share of some regions. For instance, region 4 exports refined petroleum corresponding to 21.34% of the exports of the country. Nonetheless, its complexity and industrial diversity is substantially less than the corresponding values of 3 other regions that have a standard contribution to the Colombian exports. Another example that is worth pointing out (which is also consistent with the findings in Restrepo et al., 2020) is the existence of the macro-region 10. This region has a significant contribution to the national exports (10.66%), although its industries have a low complexity index and few interactions among them. Moreover, the extension of this region could be a little misleading since the largest areas have low production and only the eastern part of the region has a significant contribution to national exports. Thus, this region provides an example of how the p-IE model could gather several areas to create some macro-regions in optimal configurations to guarantee some industrial interaction, even if this implies incurring in negligible negative spatial interactions given the extent of the region (because of the spatial decay penalization).

The next step in the free evolutionary p-IE model according to the pseudo-code in Section 3.4 consists in introducing one by one the goods detected as target goods in stage 0 (TG), starting with the good most likely to be produced according to equation (3.8). We continue with this process until the local search (Tabu) finds an improvement in its first iteration (see Algorithm 3.4.2). The recursive implementation of this procedure yields the output presented in Table 3.2. The first column shows how the morphology of the regions changes from iteration to iteration, in the second column we mark in red the new nodes that are incorporated into the economy between iterations. Lastly, the third column depicts the evolution of the participation of the output of each region, per cluster (modeled by equation 3.10).

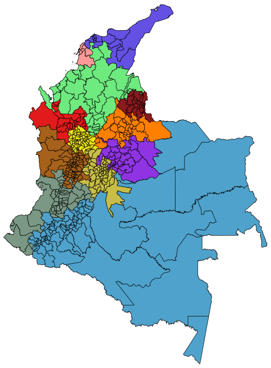
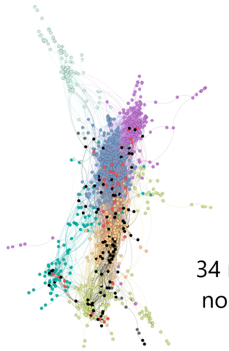
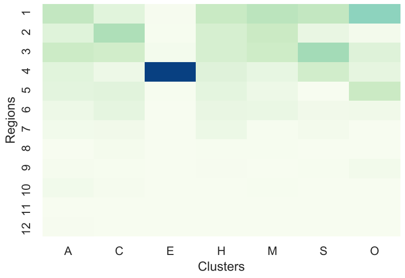
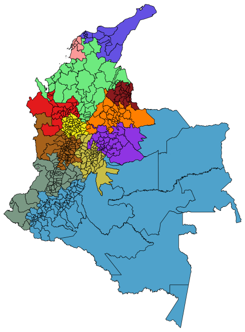
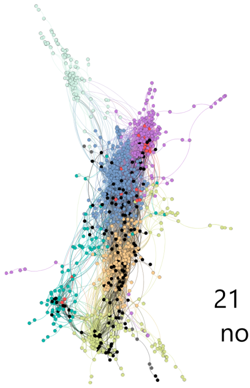
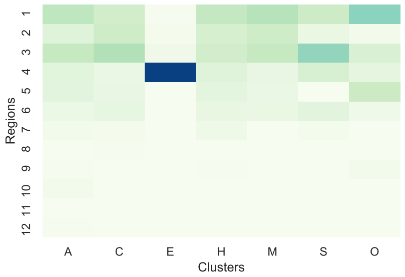
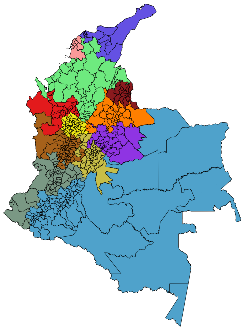
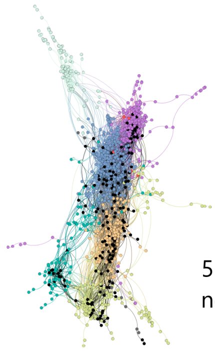
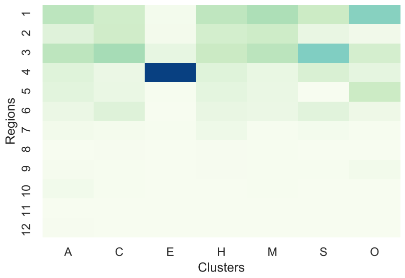
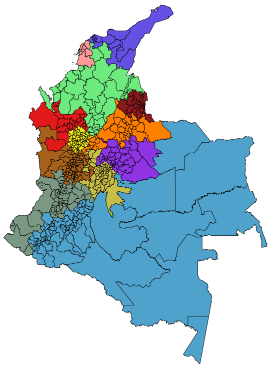
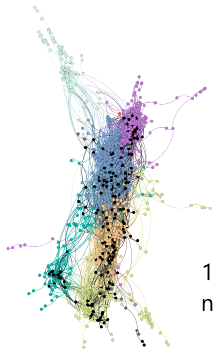
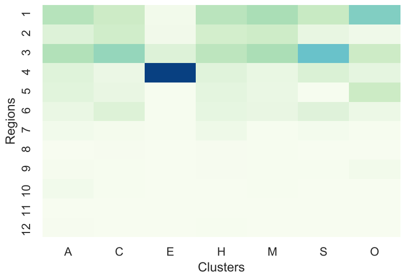
Period	Map	PS	Product distribution
1		 34 new nodes	
2		 21 new nodes	
3		 5 new nodes	
4		 1 new node	

Table 3.2: Description of the evolution of the output of the free model.

We remark the continuity in the change in both, the morphology of the regions and the distribution of the products (columns 1 and 3 in Table 3.2, respectively) so that we can keep track of the changes throughout the whole process. We describe quantitatively the evolution of the shape of the regions using a metric d that measures the distance between two clusterings. Given two p-regionalizations $R_1 = (n_1, \cdots, n_p)$ and $R_2 = (m_1, \cdots, m_p)$ with the same number of clusters we define its distance as

$$d(R_1, R_2) = \sum_{i=1}^{p} \frac{|n_i - m_i|}{\frac{n_i + m_i}{2}}. \tag{3.14}$$

We explore in detail the properties of d in the Appendix 3.7. At this point we stress that in our context it measures the costs incurred by the country due to the variation of the regional configurations between two periods. The total variation of the regional configurations, i.e., the sum of the variations from the starting period to the last period is 1.56 which amounts to an overall variation of 156% of the regions along the time (see Section 3.6 or the Appendix 3.7 for a further discussion about this result). We also notice how the evolution of the exports among regions tends to accentuate the specialization patterns evidenced in Figure (3.10b). The fifth (and final) iteration of the model does not introduce any new target node into the economy, the only variation is a subtle change in the assignation of one area between regions, thus we opt to omit it.

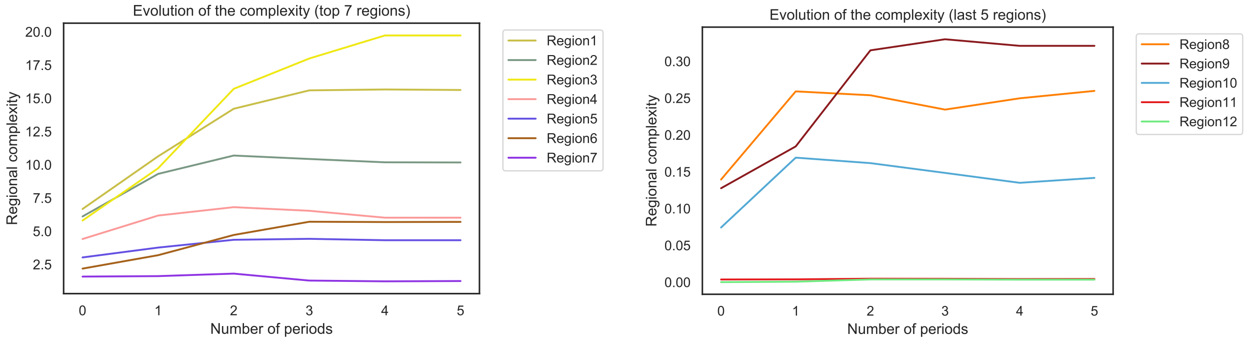
In Table 3.3 we present four measures describing the final situation of the twelve regions from the economic and the geographic perspective. The most remarkable geographic change took place in region 8. This region evolves generating a corridor in the middle of the country. In addition, region 6 expands to include 10 additional areas throughout the whole process; besides these, all other changes were subtle. Before explaining the economic evolution of the regions, it is important to notice that the model proposed to describe the regional evolution of the output introduced in equation (3.9) predicts more production for the regions whose production is more clustered spatially. More specifically, regions in which the industries linked in the product space tend to be spatially clustered are prone to augment their relative production throughout the years. This is the case for region 3 which is highly industrialized and quite compact. This region, in contrast with region 8 tends to reduce its extension along the time, but, at the same time, increases its share in the total exports. It is important to point out that the diffusion process followed by each region by period 5 reaches all the target nodes of the country (157). Thus, the measure of the number of industries per region is not longer informative. Instead of this measure, the combination of the last column of Table 3.3 with the heat-maps describing the evolution of the exports share per region provides a better understanding of the evolution of the production.

We notice three clear patterns in the evolution of the output of the regions, (1) regions whose share remained in the top 4 (approximately 10% or more), (2) regions whose final share dropped 1% and 10% and, (3) regions whose final share remained or dropped below 1%. This behaviour is strongly predicted by the initial complexity of the regions. For instance, all the regions whose complexity is below 1 ended with a complexity strictly less than one. Region 10 provides a striking example of the influence of the initial complexity in the evolution of the output. Notice that its complexity started below 0.01 but its share started above 10%. However, its production by the fifth period dropped below 1% (it belongs to the third group mentioned above). Another example of the influence of the complexity in the output can be found in region 4. As explained below Table 3.1, its high share in the exports obeys its intensive production of refined oil. In Table 3.3 we see how its share decreased dramatically, but remains close to 10%. These and other changes can be further explained from the evolution of the regional complexity depicted in Figure3.11. Given the strong segmentation of the complexity between values above 1 and below (i.e., persistence of the regional

Region Id	Exports share	Number of areas	Area (km^2)	Complexity
1	25.74%	55	29,274,816	15.6256
2	15.12%	42	56,918,469	10.1736
3	30.11%	60	11,383,625	19.7277
4	9.35%	5	8,153,318	6.0184
5	7.23%	14	45,511,327	4.3232
6	7.99%	81	53,946,990	5.7021
7	2.19%	32	50,901,402	1.2529
8	0.48%	37	48,450,947	0.2598
9	0.74%	20	16,136,694	0.3210
10	0.75%	85	561,372,138	0.1417
11	0.07%	31	37,790,084	0.0044
12	0.22%	32	106,888,437	0.0035

Table 3.3: Description of the final regional configuration of the free model.

complexity above or below 1, respectively) we depict the evolution of regional complexity for two groups, region 1 to region 7 in Figure 3.11a and region 8 to region 12 in Figure 3.11b.



(a) Evolution of the regional complexity of regions with complexity above 1.

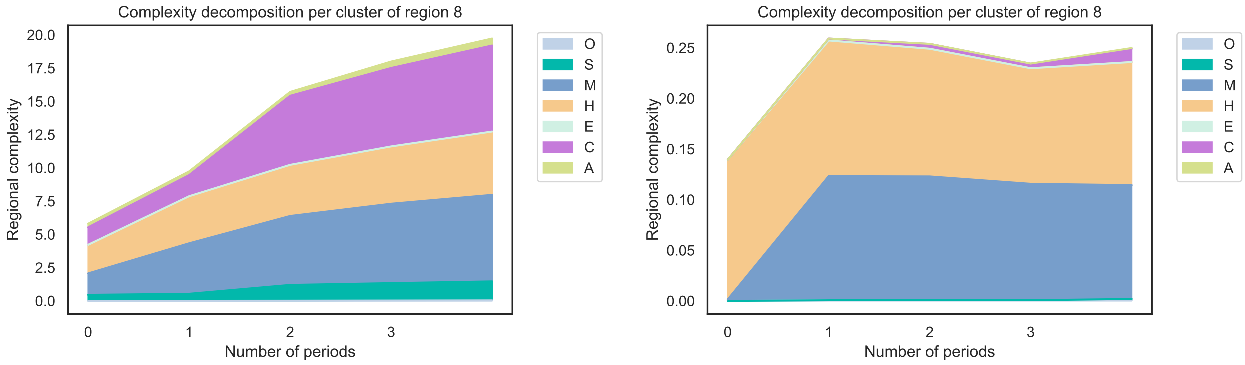
(b) Evolution of the regional complexity of regions with complexity below 1.

Figure 3.11: Figure 3.11a shows the evolution of the complexity of the regions with high complexity along the 6 periods. The most remarkable feature of this graph is how region 3 overtakes regions 1 and 2. Figure 3.11b shows the evolution in the complexity of regions with low complexity.

3.5.4 Analysis of regions 3 and 8

We conclude the analysis of the output of the free model for Colombia comparing regions 3 and 8. We chose these two regions given their differing processes: region 3 evolves to gain compactness dropping several areas in the process, whereas, region 8 augmented its size seeking to gain industrial interactions and becoming more elongated. Besides its substantially different initial conditions in terms of complexity and industries we notice interesting patters in their complexity evolution through the time. For instance, region 3 boosted its complexity building stronger internal links between its key industries. In contrast, region 8 improved its complexity initially thanks to the first wave of innovations entering into the region. Nonetheless, this process stabilizes rapidly leaving it stagnated in a low complexity level.

The difference of scales between the regional complexities is explained by the combination of the relative sizes of the economies (exports shares) and the complexity distribution among the goods produced initially by each region. Beyond the evident difference of exports shares, we can see the impact of the specialization patterns of each region in Figure 3.12. Both regions have a significant contribution coming from Household consumption goods and Machinery & vehicles. However, as we can see in Figure 3.12a , region 3 has a more diverse output with a significant presence in the sector of Chemicals and textiles (either from the cluster of Agriculture & fabrics or from Synthetics & textiles). On the other hand, region 8 was specialized almost exclusively in the production of Household consumption goods in the first period, with an incipient presence of the industry of Machinery & vehicles which strengthens through time as we can see in Figure 3.12b.



(a) Evolution of the regional complexity of region 3 per cluster.

(b) Evolution of the regional complexity of region 3 per cluster.

Figure 3.12: Figure 3.11a and Figure 3.11b shows the evolution of the regional complexity of regions 3 and 8, respectively, and the contribution of each industrial cluster to this evolution. We remark that the process of industrial diversification and the complexity gain seem to be concomitant processes. With region 3 strengthening its complexity as long as it diversifies into more and more complex sectors, whilst region 8 boosts initially complexity when it diversifies into a new sector (Machinery & vehicles), but looses quickly this impulse due to its poor diversification.

We proceed zooming in the initial and final compositions of each region to find exactly which location patterns on the PS and, more specifically, within the clusters in which they are specialized on, could drive their economies to the trajectory predicted by this model. The dashboards contained in Table 3.4 and Table 3.5 provide a detailed account of initial and final stages of the evolution performed by regions 3 and 8, respectively. The first column of each dashboard shows the initial geometry of the region; the initial position of the region per cluster with their respective color code and in gray the target goods that are not yet developed in the economy, but that will be developed throughout the process at some point; lastly, the first column contains a tree map with the goods already produced in the country, where the sizes of the boxes are proportional to the betweenness centrality (BC in short, see e.g. Barabási et al., 2016) of each node in the PS and also each box contains information about the share of the region in the good and its complexity. This column provides a diagnostic of the current situation of the regions according to their capacity to innovate and to incorporate effectively complex goods into their economy. For this purpose, the second image in the first column shows the position of the region in the graph containing the innovation potential of the Colombian economy. From here, we see that region 3 and 8 have a strong presence in goods of the clusters of Household consumption goods and Machinery & vehicles. The main qualitative difference between them is the presence of region 3 in Chemicals and Synthetics & textiles. This information is complemented with the tree map which ranks the available goods per region depending in their innovation potential. The betweenness centrality shows how likely is a good to be relevant in the diffusion routes on the PS. So, a priori, a good with high betweenness centrality appears as a natural target to forest the innovation of the region.

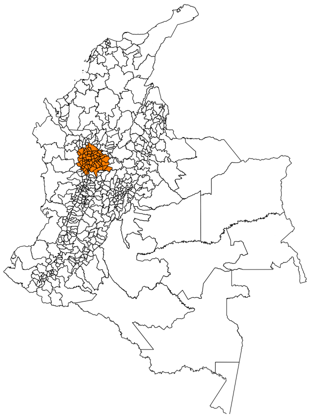
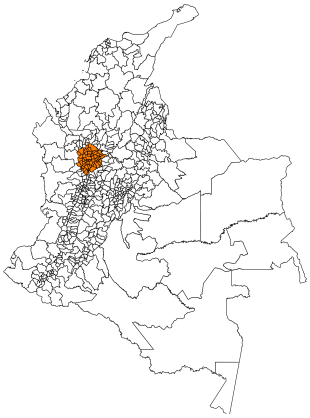
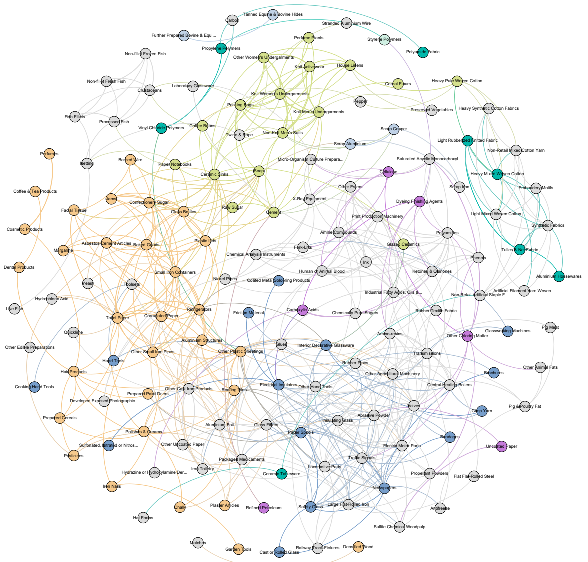
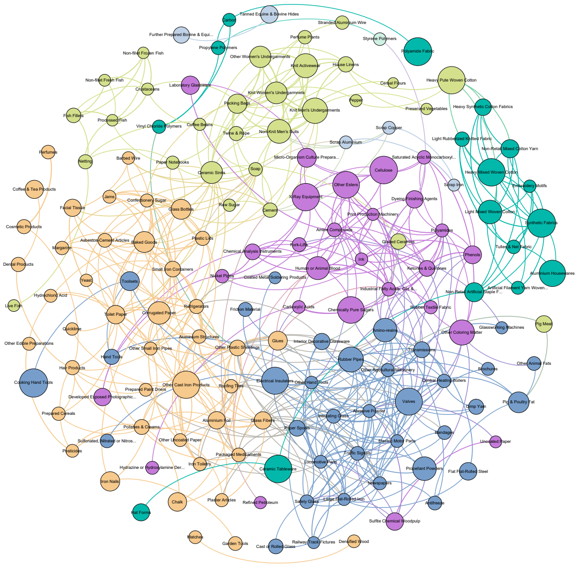
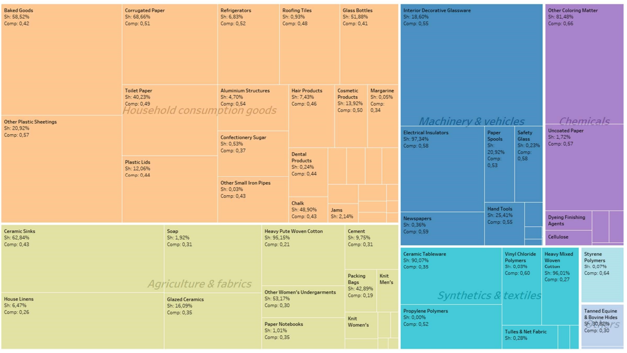
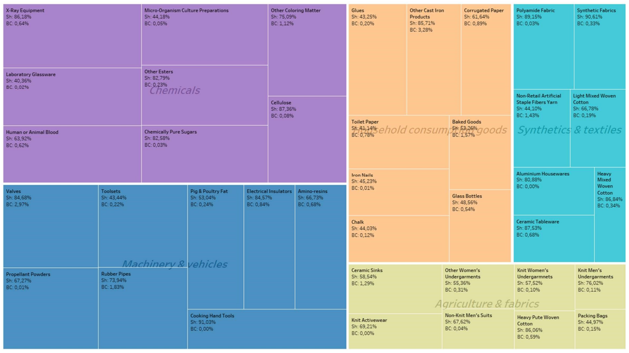
First period	Last period
	
	
	

Table 3.4: Evolution dashboard of region 3.

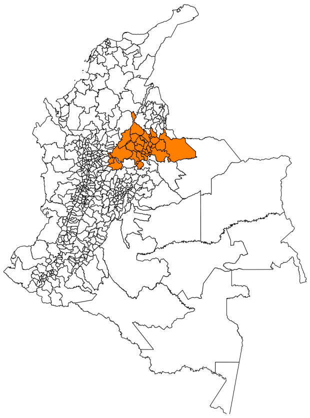
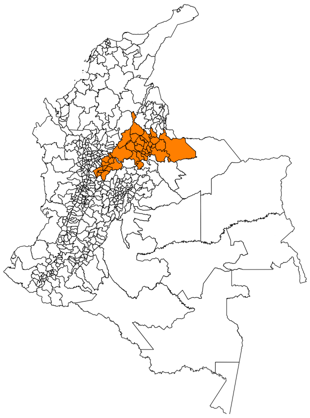
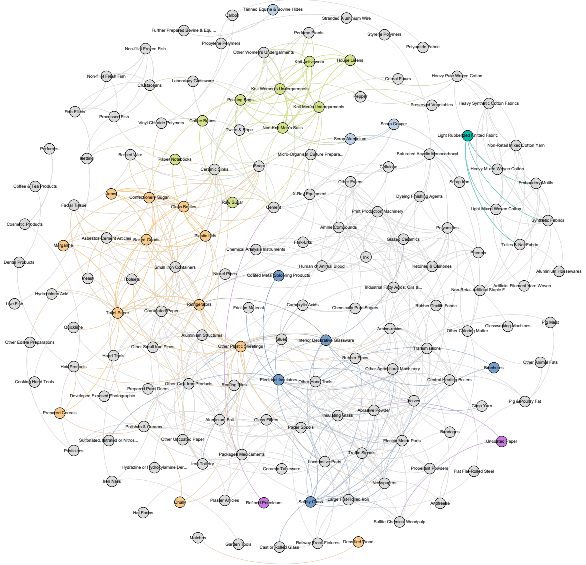
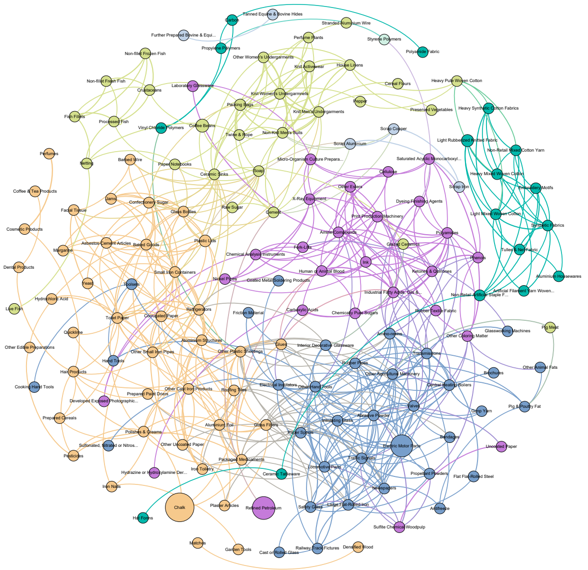
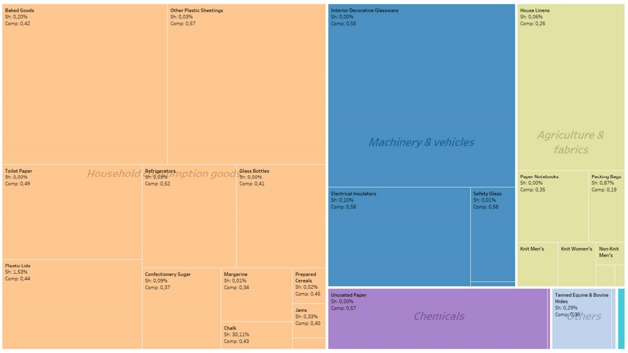
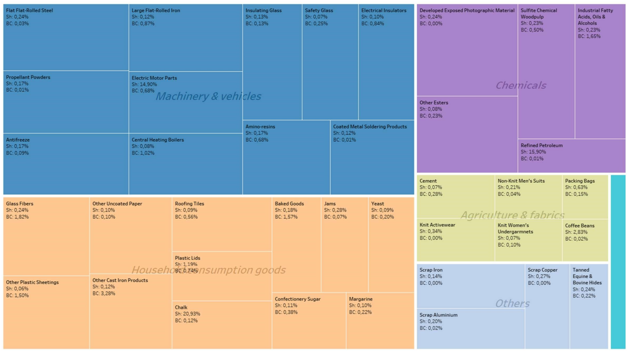
First period	Last period
	
	
	

Table 3.5: Evolution dashboard of region 8.

The second column of each dashboard provides a diagnostics of the final situation of each region after the implementation of the free model. From our approach to update the regional production (equation 3.9) and the connectivity of the PS (we are working in its giant component), all regions invade, in some extent, all the possible 157 nodes. So, we represent the innovations of each region in a nuanced way depicting the final graph with nodes whose size is proportional to the share of the region in the exports of the corresponding good. Hence, this size is interpreted as the capacity of each region to specialize and to innovate in that activity. Lastly, the tree diagram of the second column is weighted by complexity instead of betweenness centrality. Therefore, this tree diagram shows which are the most valuable goods reached by each region. A good example of how to use this dashboard to analyze the evolution of the regions is provided by the diffusion of region 3 into the Chemicals cluster. We notice that the size of the nodes within the cluster of Chemicals are significantly bigger than the corresponding ones for region 8. This combined with the tree diagrams of the second column for each region explain in a more detailed way the divergence of the complexity of both regions. Fact that was pointed out initially in the analysis of Figure 3.12.

3.5.5 Guided model

The guided model has an immediate application to the current economic goals of the Colombian government, since the present Colombian government has explicitly stated as policy goal in its agenda to develop systematically a set of goods related to creative and cultural activities also known as Economía Naranaja (Duque and Buitrago, 2013). Based on this, we selected a subset of goods that qualifies with this criteria as our long run policy goods (PG). We select 25 policy goods following the priority sectors mentioned in analytical underpinnings of the National Development Plan 2018-2022 (DNP, 2019).

We compute the guided model using the same data as in the free model. The first step towards the implementation of the guided model consists in finding the innovation potential (IP), introduced in equation (3.11), of the candidate goods (CG) for the initial state of the Colombian economy in 2010. For the sake of comparability between the free model and the guided model, we restrict the paths taken by the random walkers to take place in the routes that improves the current complexity of the economy at each step (see also Algorithm 3.4.4 in Section 3.4). We determine IP recursively through a probabilistic approach as described in Section 3.3.4. This process requires the calibration of the number of simulations to ensure the stability of the inputs of the model (see Section 3.7). After the computation of the innovation potential we proceed to find the number of regions p using the Elbow method and subtracting the number of singletons as it is explained in the implementation of the free model (see also Restrepo et al., 2020). In this case, after repeating the analysis we obtain $p = 17$. The solution for this value of p is presented in Figure 3.13 together with a heat-map showing the distribution of the production of each region per cluster analogous to the one rendered in Figure 3.10.

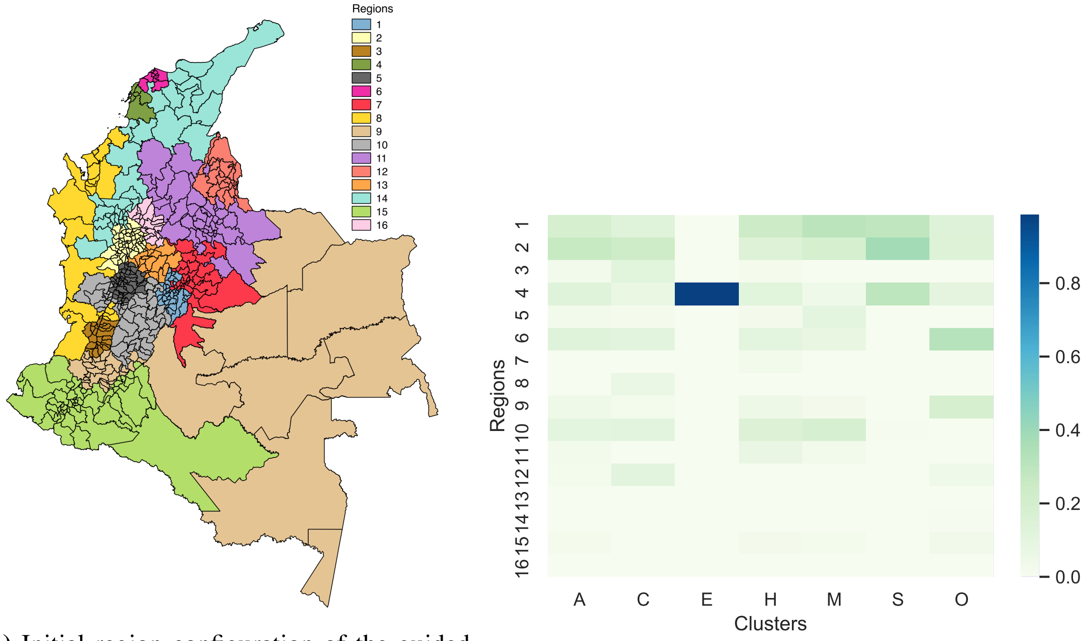
The ordering of the regions in this case describes how close each region is to being able to incorporate policy goods (or orange goods, in this particular case) PG to their economy. For this, we define, analogously to the regional complexity (3.13), the regional innovation potential:

$$IP_p^0 = \sum_{\{v \in TG_0\}} \sum_{i \in P} q_{iv} IP_v, \tag{3.15}$$

where TG_0 are the target goods at the initial period, IP_v is the innovation potential of the target good v and q_{iv} is the probability of area i to introduce good v in the next period (see equation 3.9). Since the national innovation potential is defined inductively, we follow a similar process to define the innovation potential of a region at period t as follows:

$$IP_p^t = IP_p^{t-1} + \sum_{\{v \in TG_t\}} \sum_{i \in P} q_{iv} IP_v. \tag{3.16}$$

Following similar criteria as in the free model we set $p = 16$. We use the ordering introduced by the index in equation (3.15). This ordering is described in both, Figure 3.13 and Table 3.6. Furthermore, Table 3.6 summarizes other important properties of the regions, among them the total exports share for the initial period.



(a) Initial region configuration of the guided model.

(b) Initial distribution of the production per regions.

Figure 3.13: This Figure is analogous to Figure 3.10 and describes in Figure 3.13a the initial regions of the guided model, whilst Figure 3.13b shows the relative distribution of the regions per cluster.

We notice that some regional configurations obtained in the free model are also present in the guided model. For example, the macro region 10 in the free model corresponds to the macro region 9 in this case. Another interesting example is region 4 (which corresponds to region 4 in the free model which still has the largest export share (according to Table 3.6. These and other similarities are explained by the fact that the first term of the OF, which is common in both models, dominates considerably the second term. Nonetheless, several features of the model change dramatically, e.g., the number of regions in this initial stage.

Region ID	Exports share	Number of areas	Area (km^2)	Number of industries	Innovation Potential
1	10.66%	37	9,698,164	95	3.4929
2	9.97%	60	13,708,828	83	2.5097
3	9.37%	23	8,848,594	80	2.1444
4	30.49%	4	7,159,731	53	1.3834
5	9.92%	47	10,216,336	64	1.1435
6	5.27%	10	4,411,321	77	0.4664
7	2.81%	32	45,586,408	52	0.4112
8	0.93%	19	66,859,918	37	0.1023
9	7.25%	32	421,025,955	26	0.0867
10	2.72%	49	40,272,082	29	0.0681
11	7.76%	37	87,101,094	44	0.0473
12	0.46%	23	22,326,401	39	0.0380
13	0.12%	16	12,373,084	12	0.0198
14	0.55%	35	108,429,447	37	0.0056
15	1.67%	51	159,191,499	23	0.0029
16	0.05%	19	9,519,383	16	0.0001

Table 3.6: Description of the initial regional configuration of the guided model.

The iterative algorithm of the guided model (Algorithm 3.4.4) follows the same lines as the free evolutionary p-IE model presented in Section 3.4. The main difference in this case is the preferential regional design that facilitates the diffusion towards target goods with high innovation potential. These goods are introduced in the same order as in the free model, up to the local search (Tabu) finds an improvement in its first iteration (see Algorithm 3.4.4). We depict the evolution of the output of this model in Table 3.7. This table is completely analogous to Table 3.2. The only difference appears in the second and third column in which we mark in orange the number of innovations that correspond to policy goods. Notice that, after some point, the orange entries of the matrices in the third column become constant, reflecting the connectedness of the PS per region. In this case, the convergence of the algorithm is faster compared with the free model, stopping after 4 iterations. We depict 3 of them in Table 3.7 omitting the last period since it only introduces one extra policy good and does not change the regionalization attained in the third step.

Only 18 out of the 25 targets selected are reached as shown in Table 3.7 (adding the policy good of the last iteration). This is a result of the restrictions imposed in the selection of the target goods, (see Section 3.3) which guarantees the complexity improvement in both, the guided model and the free model. In the case of the free model, it behaves like a greedy algorithm always trying to maximize the complexity at each stage. In contrast, the guided model combines a certain degree of "greediness" only allowing goods that improve the total complexity of the country, but, at the same time, accelerates the diversification towards certain subset of goods. In this case, the constraint imposed by the complexity could make some policy goods infeasible (just like in our case). However, this combination of strategies is desirable in a guided diversification model because it achieves a twofold objective, (1) prevents the economy to incorporate unnecessary goods in its way to the long run targets and, (2) it is conservative enough to avoid routes that lead the economy to sets of goods with low complexity.

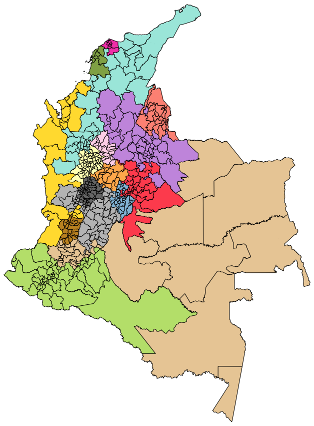
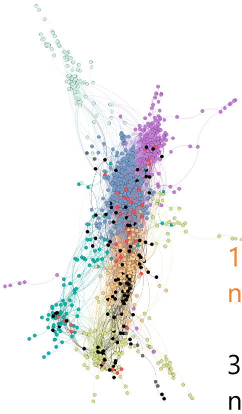
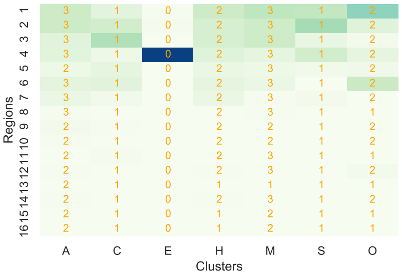
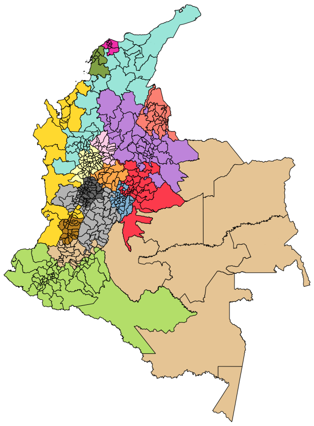
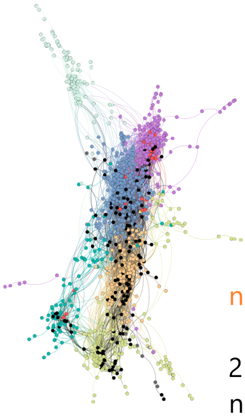
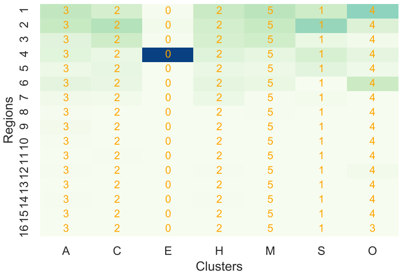
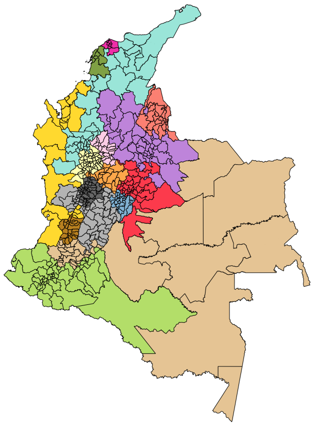
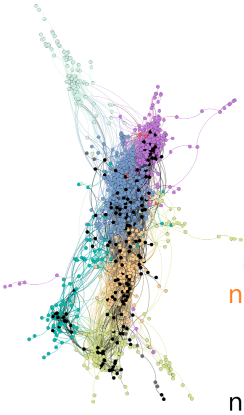
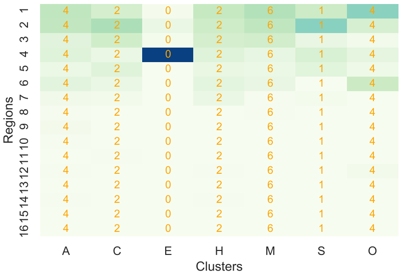
<table>
<tr><th>Period</th><th>Map</th><th>PS</th><th>Product distribution</th></tr>
<tr><td>1</td><td></td><td>
12 policy nodes
35 new nodes</td><td>
Regions / Clusters
<table>
<tr><th></th><th>A</th><th>C</th><th>E</th><th>H</th><th>M</th><th>S</th><th>O</th></tr>
<tr><td>1</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>2</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>3</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>4</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>5</td><td>2</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>6</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>7</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>8</td><td>3</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>1</td></tr>
<tr><td>9</td><td>2</td><td>1</td><td>0</td><td>2</td><td>2</td><td>1</td><td>2</td></tr>
<tr><td>10</td><td>2</td><td>1</td><td>0</td><td>2</td><td>2</td><td>1</td><td>2</td></tr>
<tr><td>11</td><td>2</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>1</td></tr>
<tr><td>12</td><td>2</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>13</td><td>2</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>14</td><td>2</td><td>1</td><td>0</td><td>2</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td>15</td><td>2</td><td>1</td><td>0</td><td>1</td><td>2</td><td>1</td><td>1</td></tr>
<tr><td>16</td><td>2</td><td>1</td><td>0</td><td>1</td><td>2</td><td>1</td><td>1</td></tr>
</table></td></tr>
<tr><td>2</td><td></td><td>
5 policy nodes
21 new nodes</td><td>
Regions / Clusters
<table>
<tr><th></th><th>A</th><th>C</th><th>E</th><th>H</th><th>M</th><th>S</th><th>O</th></tr>
<tr><td>1</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>2</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>3</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>4</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>5</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>6</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>7</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>8</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>9</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>10</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>11</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>12</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>13</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>14</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>15</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>4</td></tr>
<tr><td>16</td><td>3</td><td>2</td><td>0</td><td>2</td><td>5</td><td>1</td><td>3</td></tr>
</table></td></tr>
<tr><td>3</td><td></td><td>
2 policy nodes
5 new nodes</td><td>
Regions / Clusters
<table>
<tr><th></th><th>A</th><th>C</th><th>E</th><th>H</th><th>M</th><th>S</th><th>O</th></tr>
<tr><td>1</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>2</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>3</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>4</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>5</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>6</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>7</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>8</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>9</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>10</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>11</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>12</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>13</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>14</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>15</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
<tr><td>16</td><td>4</td><td>2</td><td>0</td><td>2</td><td>6</td><td>1</td><td>4</td></tr>
</table></td></tr>
</table>

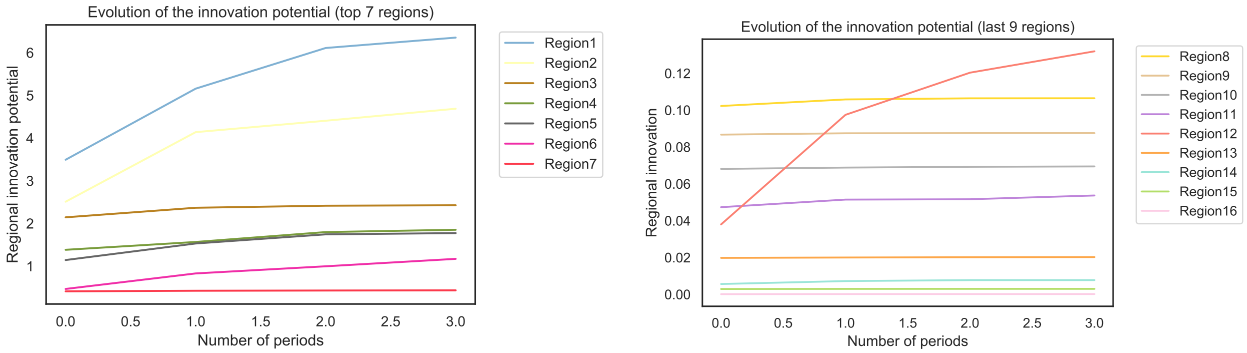
Table 3.7: Description of the evolution of the output of the guided model.

Table 3.8 summarizes the final situation of the 16 regions from the economic and the geographic perspective. We focus again in the most remarkable geographic and economic changes. In terms of the geography, we notice that region 7 was the region with the most conspicuous change dropping and adding areas and ending up its evolution with 4 more areas. We measure the change of the regional configurations using the metric introduced in equation (3.14). In this case obtain a total variation of 0.87, which amounts to an overall variation of 87% in the assignation of areas to regions (see Section 3.6 or the Appendix 3.7) for a further discussion about this).

Similarly to the free model, we see that the model predicted a substantial change of relative regional output. In this case, the output is directly proportional to the innovation potential of the region. The strongly autoregressive component of the innovation potential (i.e., equation 3.16) forces the index to never decrease from period of period. More interestingly, we notice in 3.14a that the innovation potential of the 2 regions with the highest initial innovation potential increases substantially over the time. Also, region 12 innovation potential increased so that it ended in position 8th in Figure 3.14a. All the other regions depicted in Figure 3.14 had an almost flat IP curve.

Region ID	**Exports share**	**Number of areas**	**Area**	**Innovation potential**
1	24.74%	31	8177942	6.3542
2	28.80%	58	13325413	4.6882
3	14.88%	23	8848594	2.4266
4	10.04%	5	8153318	1.8526
5	7.58%	47	10216336	1.7748
6	7.63%	9	3417734	1.1714
7	3.08%	36	46753060	0.4340
8	0.51%	19	66859918	0.1065
9	0.61%	32	421025955	0.0876
10	0.35%	49	40272082	0.0695
11	0.41%	37	87101094	0.0537
12	0.79%	23	22326401	0.1320
13	0.08%	20	13110068	0.0203
14	0.39%	35	108429447	0.0078
15	0.06%	51	159191499	0.0030
16	0.05%	19	9519383	0.0002

Table 3.8: Description of the final regional configuration of the guided model.

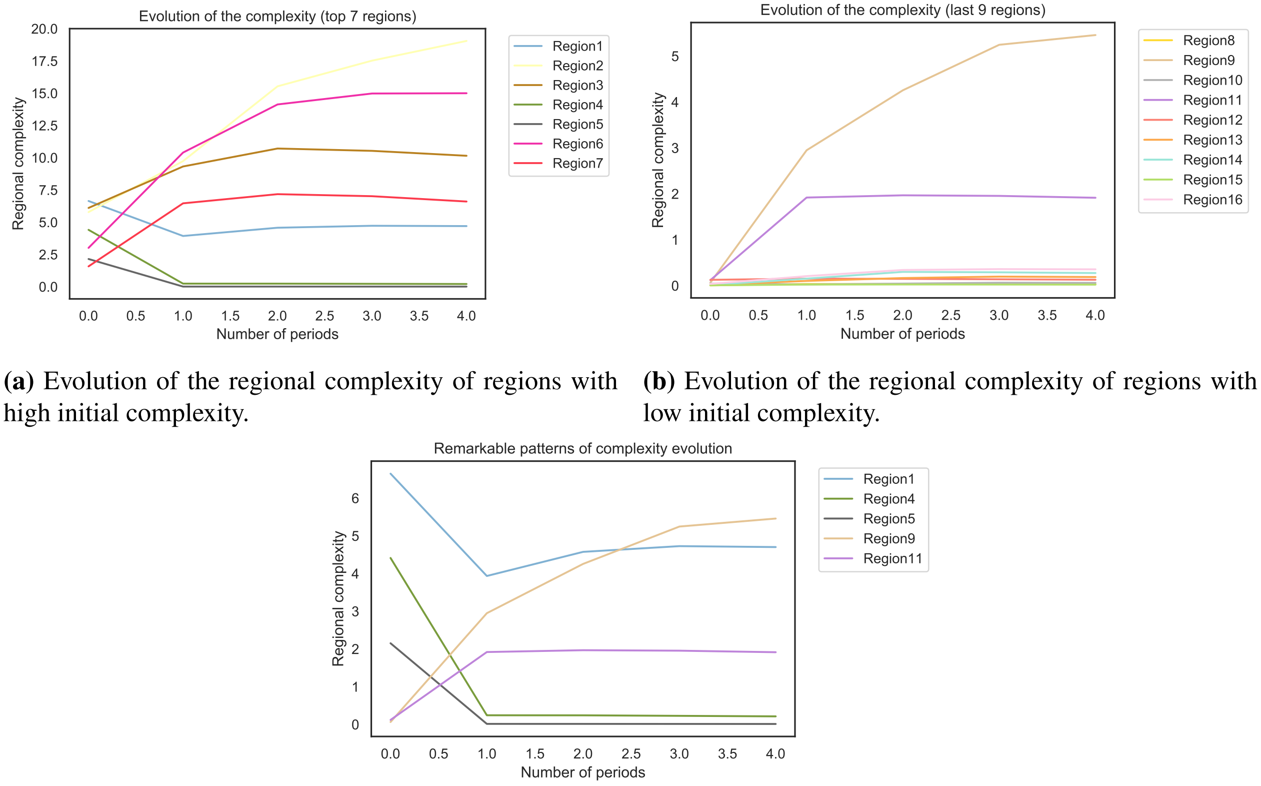


(a) Evolution of the regional innovation potential of the first seven regions.

(b) Evolution of the innovation potential of the last 9 regions.

Figure 3.14: This figure shows the evolution of the IP of the regions with high initial innovation potential for 3 regions. The most remarkable feature of this graph is how region 12 overtakes regions 8,9,10 and 11. Figure 3.14a and Figure 3.14b shows the evolution in the innovation potential of regions with high and low innovation potential, respectively.

We also consider the evolution of the regional complexity (see equation (3.13)) in this case. We recall that both, the guided and the free model, are constrained to only allow innovations that improve the average complexity at each stage (see Section 3.4). Thus, the overall complexity of the country must increase with the implementation of any of these models. However, the evolution of the complexity at regional scale may follow more involved patterns. For the free model we already offer a simplified taxonomy of the evolution patterns of the regions according to the graphs depicted in Figure 3.11. In this case, we saw that the complexity of the regions mostly increased along the time. In contrast, in the guided model there are regions whose complexity significantly fall over the time whereas other evidenced dramatic complexity improvements. An striking example of this is reflected in Figure 3.15c, where many regions with initial low complexity end up having significantly more complexity that other regions which initially had more complexity (this even includes the region whose complexity was the gratest initially). These differences, depicted in Figure 3.15, suggest that the implementation of the guided model could generate economic incentives that can alter significantly the growth path of the regions and, as a consequence of this, it may increase the existing gap between the regional economies and, at last, increase the inequality between them.



(a) Evolution of the regional complexity of regions with high initial complexity.

(b) Evolution of the regional complexity of regions with low initial complexity.

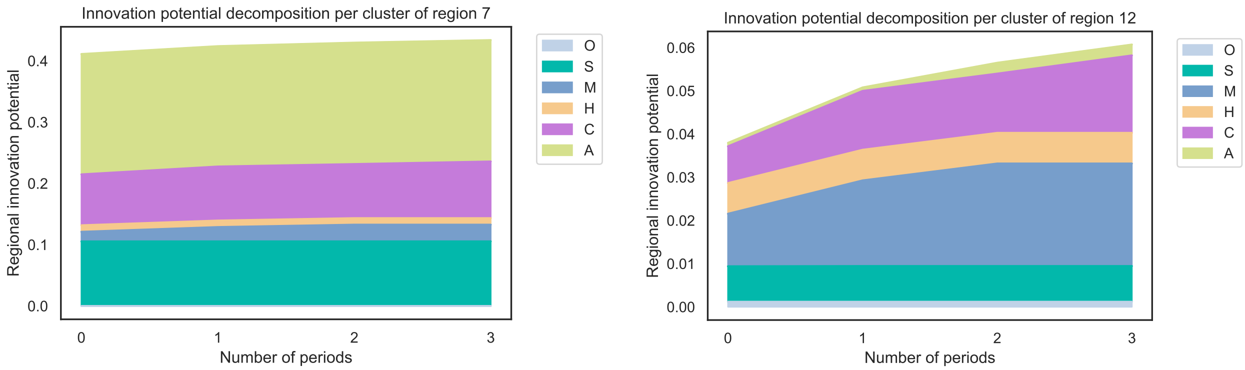
(c) Significant take overs in the evolution of the regional complexity.

Figure 3.15: Figure 3.15a shows the evolution of the complexity of the regions with initial high complexity along the 5 periods simulated by the model. Figure 3.11b shows the evolution in the complexity of regions with initial low complexity. We remark that radical changes in the relative ranking of the regions in terms of their complexity over the time. Figure 3.15c reflects the changes that are not directly captured in any of the other two parts of the figure.

3.5.6 Analysis of regions 7 and 12

We conclude the analysis of the output of the guided model for Colombia comparing regions 7 and 12. As explained before, we chose these two regions since one of them (region 7) had the most evident morphological change dropping and adding areas throughout the implementation of the algorithm, whereas, the other one exhibited the largest relative improvement in its IP along the time. We start pointing out that region 7, despite of exchanging several areas with its neighboring regions, only increased its innovation potential from 0.4112 to 0.4340. In contrast, region 12 did not change at all geographically during the 3 periods and drastically improved (in a relative way) its IP from 0.038 to 0.132.

We give a further look to the goods that served as stepping-stones to reach the policy goods starting with the decomposition of the IP in terms of the industrial clusters. In Figure 3.16 we start analyzing the evolution of the complexity of each region at a cluster level. Figure 3.16a adds information to the already noticed from Figure 3.14a there is not substantial variation either in the IP within any of the clusters. On the other hand, the increment of the IP in region 12 is mostly explained by the innovations introduce in the region from the cluster of Machinery & vehicles as it can be seen in Figure 3.16b.



(a) Evolution of the IP of region 7 per cluster.

(b) Evolution of the IP of region 12 per cluster.

Figure 3.16: Figure 3.11a and Figure 3.11b shows the evolution of the IP of regions 3 and 8, respectively, and the contribution of each industrial cluster to this evolution.

We complement the information illustrated on Figure 3.16 with the more comprehensive dashboards rendered in Table 3.9 and Table 3.10. As in Section 3.5.4 we zoom in on the initial and final stages of each region. In both cases, we provide a description of the initial state of the region in terms of its location on the PS and, depending of the type of goods already produced, we diagnose which are the best candidate industries to invest in to maximize the chances of the region to reach the long run targets in the PS. For instance, in Table 3.9 we present in the first column a subset of the PS containing the goods produced in the region in the initial period colored according to the nomenclature established in Figure 3.8. The other nodes (in gray) in this graph correspond to the goods that will be activated eventually in the contagion process underlying to the guided model. The tree diagram depicted below the graph on the first column provides more detailed information about the goods already produced in the region. The size of the boxes are proportional to the betweenness centrality of the nodes and, therefore, it suggests which goods have better connectivity on the graph. This metric is accompanied with the share of the product in the region exports (sh) and the IP of each good (multiplied by 100). We point out that this IP is not exactly the same introduced in equation (3.11) since it is computed for goods already produced in the country. The exact definition of IP_g for $g \in IG$ is

$$IP_g = \sum_{v \in TG_0} w_{gv} IP_v, \tag{3.17}$$

where IP_v was introduced in equation (3.11) and w_{gv} is the standard weight of the link between the two goods in the PS. The value IP_g basically reflects in which extent the good g is linked to goods with a significant innovation potential. Summing up, the first column of each dashboard provides useful information to determine the current state of the economy in terms of production of goods that are relevant to achieve the diversification goals of the region. In contrast, the second column of each dashboard shows the prospective scenario of the output of a region that followed the recommendations of the guided model.

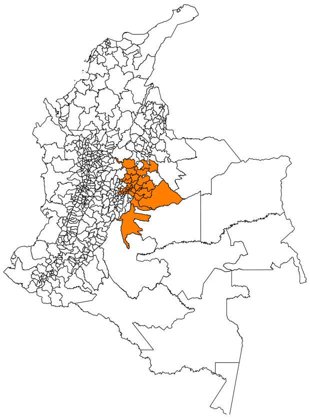
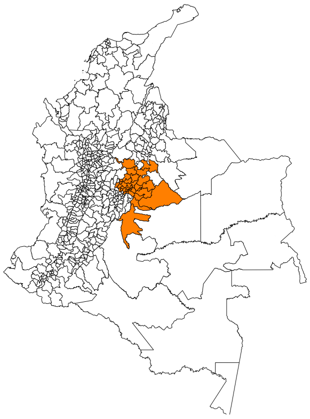
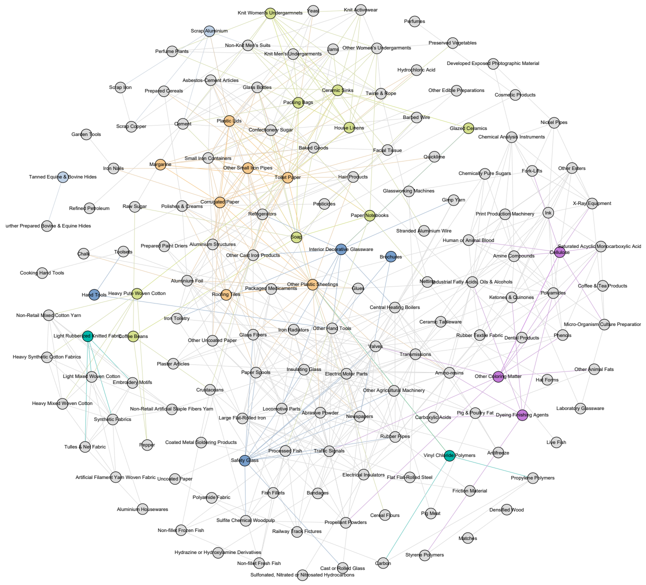
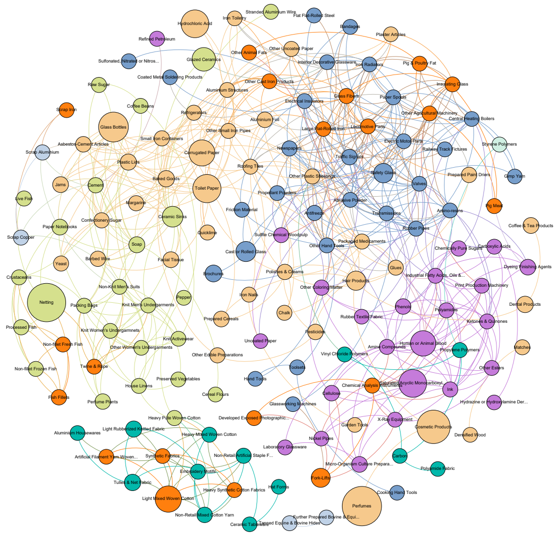
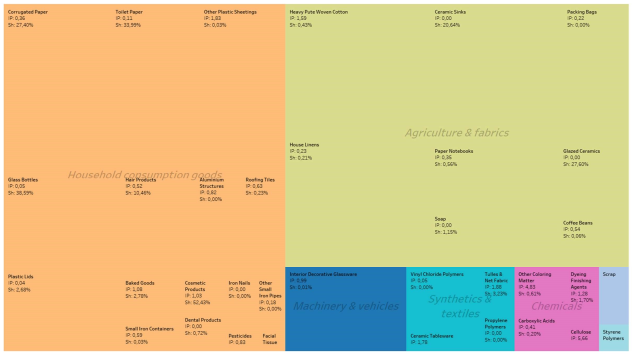
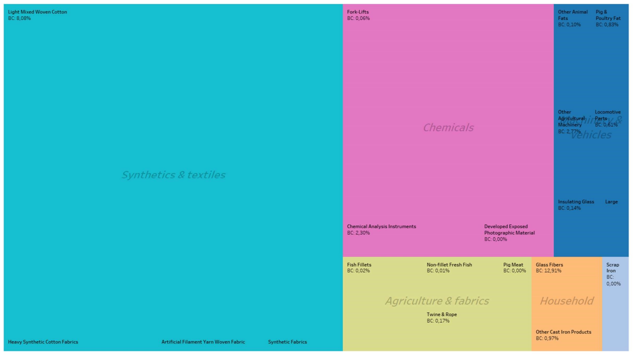
First period	Last period
	
	
	

Table 3.9: Evolution dashboard of region 7.

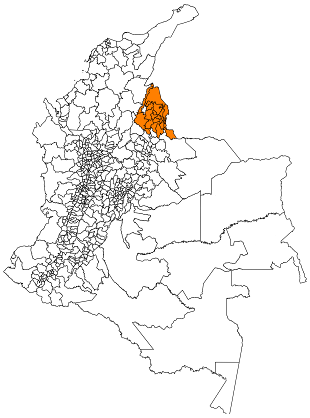
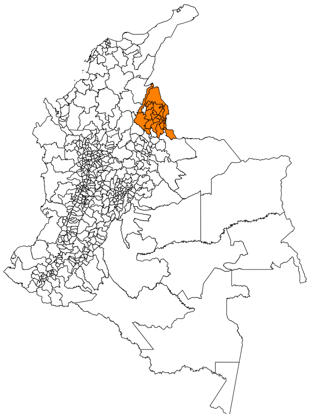
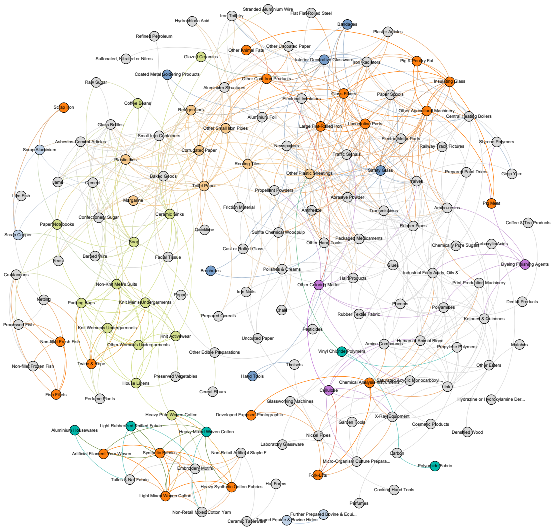
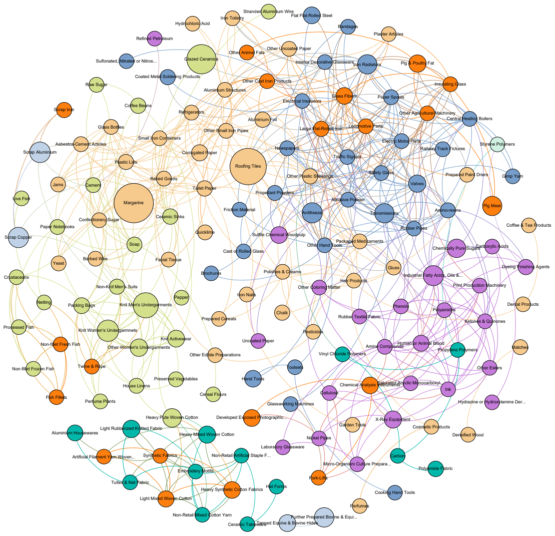
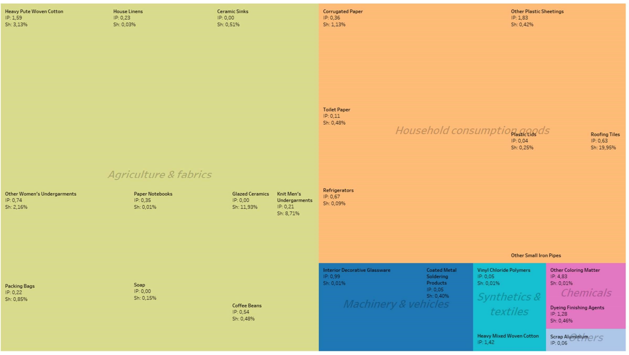
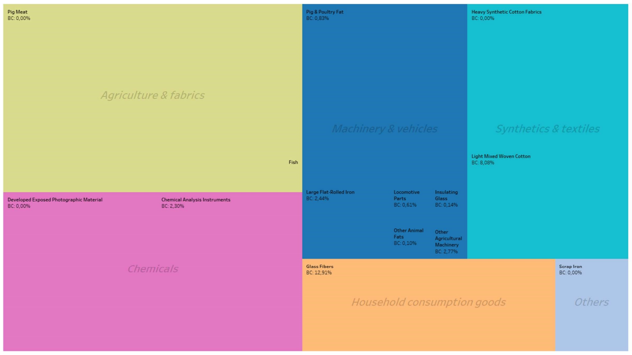
First period	Last period
	
	
	

Table 3.10: Evolution dashboard of region 12.

Analogously to Section 3.5.4 we show the final PS weighting the nodes of the graph according to the share of the region in the national production of the good. We stress the coherence between the initial and final representations of the PS in the dashboards for each region. For example, region 7 has a moderate presence in the Chemistry cluster which is reflected in the relative size of the nodes corresponding belonging to this cluster in the second column of Table 3.9; in contrast, region 12 has almost no presence in this cluster, which is reflected in the small size of those nodes in the corresponding graph in the second column of Table 3.9.

Finally, the tree diagram rendered in the second column of each dashboard shows the relative share of the region in the policy goods. This tree diagram can be regarded as the ultimate measure or diagnostic of effectiveness of the guided model. The boxes in these tree diagrams are weighted by the share of each region in the production of each good in the last period of the simulation.

3.6 Conclusions

In this paper we introduced two novel diversification strategies based on the p-innovation ecosystems (p-IE) model. We combined the ideas of complex contagion models, the p-IE model and complexity theory to incorporate a dynamic component into the p-IE model to assess and recommend long run diversification strategies for countries and regions. We do this through two versions of the model. The first one, the free version of the model, attempts to maximize the total complexity of the country's basket of goods, without defining the basket to be achieved. The second one or guided version, seeks instead to minimize the time required to develop a given set of target products defined by a policy maker. We also expand on the existing tools to understand the dynamics of innovation within a country by incorporating region effects, i.e., distance and boundary effects. Additionally, both, the free and guided versions of the model proposed provide detailed and systematic information about the regional and national strategies that should be followed by an economy pursuing long run policy goals.

We illustrated the potential applications of both versions of the model using the Colombian economy as an example. This work sheds light on certain features of both models at global (national) and local (regional) levels. Since both models impose constraints that force the economy to only develop economic activities that improve the average complexity of the system, we expect that the total amount of innovations achieved by each model should be very similar. For the same reasons, the total complexity of the country is bounded to increase in both models. However, ein the short run both models may have substantial differences, even at the global scale, depending on the selection of the long run targets in the guided model. In the case presented in this article it is reflected in the speed of convergence of the models (it took one period less in the guided case) and in the the rate at which the innovations are incorporated or absorbed by the economy.

The differences between both models are more ostensible at regional scale. The main source of differences in this case comes from the uneven distribution of initial endowments of industries of each economic area. These differences combined with the target set established in the guided model produces strong incentives to generate regional alliances (innovation ecosystems) that, in general, will differ from the ones proposed for the free version of the model. In our case of study, we found 12 as the optimal number of regions for the free model, whereas the guided model suggested a regionalization with 16 regions. This difference combined with the structural differences of the economic areas (municipalities) of Colombia make evident that these two models offer two qualitatively different alternatives. More interestingly we found that for the guided model, in contrast with the global dynamic, the evolution of the complexity at regional level may be dramatically distinct. For the case of study we had that the free version of the model designed regions so that, most of them, exhibited consistent improvement in their complexity levels. The benefits in terms of complexity for the guided model were more unbalanced and uneven. Certain regions suffered sustained complexity falls during the 5 periods, whereas other regions that, roughly, correspond to regions in the free model that presented negligible complexity improvement (e.g., region 10 in the free model and region 9 in the guided model). Thus, concretely, we found that the strategy of focusing the economic policy efforts in the development of the products suggested by the orange economy could generate incentives to the further divergence of the regional economies.

The guided model can however contribute to the policy discussion further as it allows the possibility to select a broad set of target industries and evaluate the effect of different policy interventions, such as the move toward an orange economy as discussed in this paper, or a move toward environmentally sustainable economic diversification (see Mealy and Teytelboym, 2020). Additionally, under certain circumstances the guided model can be used to attain a complexity improvement larger than the one obtained by the free model. Since the free model behaves like a greedy algorithm, seeking to maximize the complexity of the economy each period, it produces sub-optimal outcomes in the long run. This problem can be overcome setting up a new optimization problem in which the target set in the guided model is chosen such that the final complexity improvement is maximized. Also, following the same lines of reasoning, the target set can be determined such that the complexity dissimilarity among a given number of regions is minimized in the long run. This family of applications and modifications of the model opens up a new line of research in the area of optimal diversification strategies.

In both models, the qualitative dynamics followed by the regions can be classified into common categories for both models: (1) regions that evolve strengthening the links among their initial endowment of regions without enlarging its territory along the time, (2) regions that change its morphology adding or dropping areas seeking to find suitable new industrial associations to thrive. We provided some insights about this aspect zooming in certain regions in each model and studying variation of the relative distributions of the clusterings through the time. In the case of the Colombian economy, we found that the clusters in the guided model were more stable compared to the ones of the free model. This fact gains relevancy since the guided model generated more regions than the free model and, therefore, was more sensitive to possible exchange of areas between regions between periods . A deeper understanding of the evolution of the morphology of the regions obtained by the model are of central interest for regional policy development, since the policies that may be needed to strengthen industrial collaborations within a region may be considerably different when the needed links that forest the diversification lie beyond outside the original regional boundaries. Thus, knowing the nature of the evolution of each region can suggest how design policies tailored for each one of them, e.g., investing in infrastructure for regions with the tendency to elongate or designing collaboration incentives for industrial clusters. This will the be subject of a future work.

3.7 Appendix

3.7.1 Calibration of the number of simulations in the stepping-stones ranking

We calibrate the number of simulations of the random walkers through a convergence analysis of the ranking explained in 3.3.4. To this end, we propose a quantitative index that measures how far is the ranking to become stable. Let us denote by $IP_{i,n}$ the innovation potential of $g_i \in CG$ computed after n simulations of the random walkers ordered such that $IP_{i,n} \leq IP_{i+1,n}$. Notice that this indexing depends, a priori, on the number of simulations n. We want to find the number of simulations n such that this ordering does not change. If we denote by $\Delta IP_{i,n} = IP_{i,n+1} - IP_{i,n}$, it is clear that the order between i and $i+1$ is preserved if and only if

$$S_{i,n} = \frac{\Delta IP_{i,n} - \Delta IP_{i,n+1}}{IP_{i+1,n} - IP_{i,n}} \leq 1. \tag{3.18}$$

Thus, this criterion motivates us to define the Stability ratio

$$S_n = \max_i S_{i,n}. \tag{3.19}$$

So, it is clear that the ordering fo the innovation potential does not change if and only if $S_n < 1$ and that the ordering become more stable as S_n approaches to zero. Figure 3.17 shows the convergence analysis and shows how we find 3000 as a suitable number of simulations for this problem.

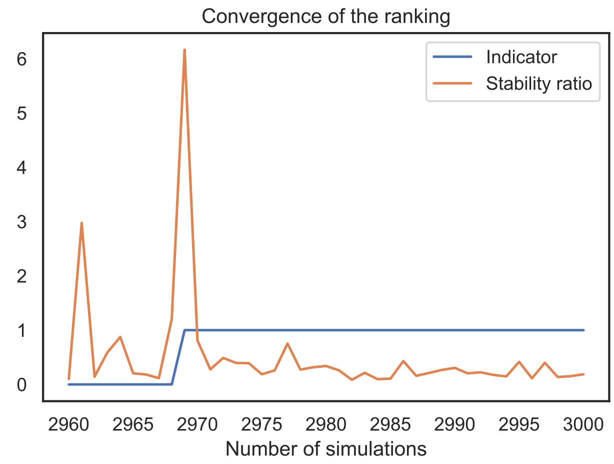


Figure 3.17: The indicator (in blue) is 1 when the ordering of the innovation potential coincides with the final ordering (3000 simulations) and 0 otherwise. The Stability ration S_n falls below 1 at $n = 2970$ and approaches systematically to 0 as n increases .

Since the number of stepping-stones decreases through the dynamic process the number of iterations required to ensure the stability of the ranking per period can be taking equal to 3000 through the whole process.

3.7.2 Distance function between clusterings

In Section 3.5 we introduced the distance function between clusterings R with the same number of clusters. A p-regionalization R_i of a spatial unit with n areas is a vector $(n_1, \cdots, n_p)$ with $\sum_{i=1}^{p} n_i = n$ and $n_i \geq 1$ for every $i = 1, \cdots, p$. Given two p-regionalizations $R_1 = (n_1, \cdots, n_p)$ and $R_2 = (m_1, \cdots, m_p)$ we define their distance as

$$d(R_1, R_2) = \sum_{i=1}^{p} \frac{|n_i - m_i|}{\frac{n_i+m_i}{2}} = 2 \sum_{i=1}^{p} \frac{|n_i - m_i|}{n_i + m_i}. \tag{3.20}$$

The value $d(R_1, R_2)$ measures the total cost for going from the configuration R_1 to the configuration R_2. Each one of terms of the sum in equation (3.20) accounts for the relative costs of dropping or adding more regions to a given cluster. Hence, the total cost is obtained adding up all these terms. Notice that the percentage change is taken with respect to the average of the size of the corresponding clusters. This choosing have two advantages: on the one hand, it guarantees that the symmetry of d implying that it actually defines a metric in the space of regional configurations; on the other hand, it penalizes the abrupt changes in the relative distribution of the areas among the clusters. The latter observation becomes clear when we notice that if $n_i >> m_i$ then $\frac{|n_i - m_i|}{\frac{n_i+m_i}{2}}$ behaves as 2. We also remark that $d(R_1, R_2)$ is always bounded from above by $2p$ and that this upper bound is sharp.

Conclusions

In this work, we introduced a new theoretical framework that exploits the role of the industrial interactions at regional level in the innovation processes to assess and recommend diversification strategies for economies. Our contributions are underpinned by the idea of the economic regions in which the interplay of industries exchanging knowledge and generating positive externalities creates suitable environments for the emergence of complex industries, this are the so-called *innovation ecosystems*. We devised a series of models to design, implement and diagnose innovation ecosystems. The main analytical tool introduced in this work is a new member of the p-regions clustering methods called the p-innovation ecosystems (p-IE) model. This model adds a new layer of complexity into the set of p-regions models by incorporating an underlying network structure that captures possible interactions between agents within the geographic areas. More precisely, this model consists of the aggregation of n areas into p spatially contiguous regions that (1) maximizes the number of relevant interactions among the industries within the same region; and (2) identify strategic relationships between industries within regions that maximize the probability of activating innovation processes that allow the region to jump to more complex relevant goods that are yet to appear in the economy.

We justified the robustness of our approach showing that other economic policies that may forest the creation of innovation ecosystems have inherent weaknesses and entail undesirable economic consequences. For this, we proposed a bargaining model in which we showed that if a government chooses a scheme of subsidies for relocating industries instead of designing optimal regions, then the information asymmetries between the firms and the government yields problems of rents extraction. Additionally, based on the p-IE model, we introduced two novel diversification strategies. We combined the ideas of complex contagion models, the p-IE model and complexity theory to incorporate a dynamic component into the p-IE model to assess and recommend long run diversification strategies for countries and regions. We do this through two versions of the model. The first one, the free version of the model, attempts to maximize the total complexity of the country's basket of goods, without defining the basket to be achieved. The second one, or guided version, seeks instead to minimize the time required to develop a given set of target products defined by a policy maker.

The contribution of this thesis to the existing literature is threefold: first of all, we contribute to the literature of economic complexity, proposing an alternative extension of the theories of Hausmann and Hidalgo to the context of economic regions. This extension is complementary to the developments of the evolutionary economic geography, since it provides a quantitative and tractable framework to exploit the findings of this latter theory for the development of economic policy. Secondly, we draw a first conceptual and theoretical link between the theory of economic complexity and the literature of quantitative region design, providing a new model with a solid economic foundation to the p-regions family. Finally, we contribute to the literature of economic growth and diversification strategies, introducing from a conceptual perspective the notion of economic region as a potential generator of economic ecosystems. This approach enable us to use regions as a new variable in the modelling of diversification strategies to pursue long run policy goals.

The main empirical contributions of this work are of descriptive and prospective nature. With the introduction of the p-IE model and its dynamic extensions we also devised a series of metrics to diagnose the current state and the evolution of the economic output of a given economy. Moreover, the metrics proposed in the framework of the model were designed such that they can be used to study the innovation processes and opportunities of regional economies that are not divided according to the regionalizations proposed by our models. We illustrate the duality of this empirical contributions showing how the diagnostics of the current state of an industrial region in terms of its complexity, betweenness centrality, innovation potential and other metrics are translated in terms of the outputs simulated by the model. As an example of this, we assess a guided diversification strategy proposed by the current Colombian administration (Economía naranja) comparing it with a model that let the industries interact and evolve freely over the time. We show that this guided diversification policy could reinforce the divergence process of the regional economies of the country, boosting the emergence of complex industries in regions with a significantly large initial endowment of capabilities and, at the same time, diminishing the overall complexity of the economy of certain less developed regions.

Another relevant feature of the models proposed in this thesis is the heterogeneity in the evolution of the morphology of the innovation ecosystems proposed. We summarize this behavior in two basic categories, (1) regions that evolve strengthening the links among their initial endowment of regions without enlarging its territory along the time, (2) regions that change its morphology adding or dropping areas seeking to find suitable new industrial associations to thrive. We provided some insights about this aspect zooming in certain regions in each model and studying variation of the relative distributions of the clusterings through the time. In the case of the Colombian economy, we found that the clusters in the guided model were more stable compared to the ones of the free model. This fact gains relevancy since the guided model generated more regions than the free model and, therefore, was more sensitive to possible exchange of areas between regions between periods. A deeper understanding of the evolution of the morphology of the regions obtained by the model are of central interest for regional policy development, since the policies that may be needed to strengthen industrial collaborations within a region may be considerably different when the needed links that forest the diversification lie beyond outside the original regional boundaries. Thus, knowing the nature of the evolution of each region can suggest how design policies tailored for each one of them, e.g., investing in infrastructure for regions with the tendency to elongate or designing collaboration incentives for industrial clusters. This will the be subject of a future work.

Bibliography

Alshamsi, A., Pinheiro, F. L., and Hidalgo, C. A. (2018). Optimal diversification strategies in the networks of related products and of related research areas. *Nature communications*, 9(1):1–7.

Arabia, S. (2016). Saudi arabia vision 2030. *Gazette, Riyadh Tuesday*, 26.

Arrow, K. J. and Lind, R. C. (1970). Uncertainty and the evaluation of public investment decisions. *The American Economic Review*, 60(3):364–378.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Technical report, Stanford.

Bahar, D., Hausmann, R., and Hidalgo, C. A. (2014). Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion? *Journal of International Economics*, 92(1):111–123.

Balassa, B. (1964). The purchasing-power parity doctrine: a reappraisal. *Journal of political Economy*, 72(6):584–596.

Baldwin, R. E. and Okubo, T. (2014). Tax competition with heterogeneous firms. *Spatial Economic Analysis*, 9(3):309–326.

Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

BC-SK-AB (2016). *The NWPTA- The Agreement*. http://www.newwestpartnershiptrade.ca/ [Accessed: 2017-03-06].

Binmore, K., Rubinstein, A., and Wolinsky, A. (1986). The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, pages 176–188.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Boschma, R. (2017). Relatedness as driver of regional diversification: A research agenda. *Regional Studies*, 51(3):351–364.

Boschma, R., Balland, P.-A., and Kogler, D. F. (2015). Relatedness and technological change in cities: the rise and fall of technological knowledge in us metropolitan areas from 1981 to 2010. *Industrial and corporate change*, 24(1):223–250.

Boschma, R., Frenken, K., Bathelt, H., Feldman, M., Kogler, D., et al. (2012). Technological relatedness and regional branching. *Beyond territory. Dynamic geographies of knowledge creation, diffusion and innovation*, pages 64–68.

Boschma, R., Minondo, A., and Navarro, M. (2013). The emergence of new industries at the regional level in s pain: A proximity approach based on product relatedness. *Economic geography*, 89(1):29–51.

Bottazzi, L. and Peri, G. (2003). Innovation and spillovers in regions: Evidence from european patent data. *European economic review*, 47(4):687–710.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Brekke, K. R. and Straume, O. R. (2004). Bilateral monopolies and location choice. *Regional Science and Urban Economics*, 34(3):275–288.

Brouwer, A. E., Mariotti, I., and Van Ommeren, J. N. (2004). The firm relocation decision: An empirical investigation. *The Annals of Regional Science*, 38(2):335–347.

Burbidge, J., Cuff, K., and Leach, J. (2006). Tax competition with heterogeneous firms. *Journal of Public Economics*, 90(3):533–549.

Bustos, S., Gomez, C., Hausmann, R., and Hidalgo, C. A. (2012). The dynamics of nestedness predicts the evolution of industrial ecosystems. *PloS one*, 7(11):e49393.

Camagni, R. and Capello, R. (2013). Regional innovation patterns and the eu regional policy reform: Toward smart innovation policies. *Growth and change*, 44(2):355–389.

Caragliu, A. and Nijkamp, P. (2016). Space and knowledge spillovers in european regions: the impact of different forms of proximity on spatial knowledge diffusion. *Journal of Economic Geography*, 16(3):749–774.

Centola, D. (2018). *How behavior spreads: The science of complex contagions*, volume 3. Princeton University Press.

Centola, D. and Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734.

Chino, A. (2016). Do labor unions affect firm payout policy?: Operating leverage and rent extraction effects. *Journal of Corporate Finance*, 41:156–178.

Ciabuschi, F., Dellestrand, H., and Kappen, P. (2012). The good, the bad, and the ugly: Technology transfer competence, rent-seeking, and bargaining power. *Journal of World Business*, 47(4):664–674.

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K. A., Davies, R., Bassett, K. L., et al. (1975). *Elements of Spatial Structure: A Quantative Approach*, volume 6. Cambridge University Press.

Commission of the European Communities (2001). *Ninth survey on state aid in the European Union*. COM(2001) 403.

Cova, T. J. and Church, R. L. (2000). Contiguity constraints for single-region site search problems. *Geographical Analysis*, 32(4):306–329.

De Groot, H. L., Poot, J., and Smit, M. J. (2016). Which agglomeration externalities matter most and why? *Journal of Economic Surveys*, 30(4):756–782.

Diodato, D., Neffke, F., and OClery, N. (2018). Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, 67(3):297–308.

DNP, D. (2019). Bases del plan nacional de desarrollo 2018-2022.

Dong, B., Gu, X., and Song, H. (2017). Capital market liberalization: Optimal tradeoff and bargaining delay. *The North American Journal of Economics and Finance*, 39:78–88.

Du, J. and Mickiewicz, T. (2016). Subsidies, rent seeking and performance: Being young, small or private in china. *Journal of Business Venturing*, 31(1):22–38.

Dupont, V. and Martin, P. (2005). Subsidies to poor regions and inequalities: some unpleasant arithmetic. *Journal of Economic Geography*, 6(2):223–240.

Duque, I. and Buitrago, P. (2013). *La economía naranja: una oportunidad infinita*. Inter-American Development Bank.

Duque, J. C., Anselin, L., and Rey, S. J. (2012). The max-p-regions problem. *Journal of Regional Science*, 52(3):397–419.

Duque, J. C., Church, R. L., and Middleton, R. S. (2011). The p-regions problem. p-. *Geographical Analysis*, 43(1):104–126.

Duque, J. C., Ramos, R., and Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, 30(3):195–220.

Fischer, M. M. (1980). Regional taxonomy: a comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*, 10(4):503–537.

Frenken, K. and Boschma, R. A. (2007). A theoretical framework for evolutionary economic geography: industrial dynamics and urban growth as a branching process. *Journal of economic geography*, 7(5):635–649.

Fujita, M., Krugman, P. R., Venables, A. J., and Fujita, M. (1999). *The spatial economy: cities, regions and international trade*, volume 213. Wiley Online Library.

Glover, F. (1977). Heuristics for integer programming using surrogate constraints. *Decision sciences*, 8(1):156–166.

Glover, F. (1989). Tabu searchpart i. *ORSA Journal on computing*, 1(3):190–206.

Glover, F. (1990). Tabu searchpart ii. *ORSA Journal on computing*, 2(1):4–32.

Greaker, M. (2003). Strategic environmental policy when the governments are threatened by relocation. *Resource and Energy Economics*, 25(2):141–154.

Gul, F. and Sonnenschein, H. (1988). On delay in bargaining with one-sided uncertainty. *Econometrica: Journal of the Econometric Society*, pages 601–611.

Han, S. and Leach, J. (2008). A bargaining model of tax competition. *Journal of Public Economics*, 92(5):1122–1141.

Haufler, A. and Stähler, F. (2013). Tax competition in a simple model with heterogeneous firms: How larger markets reduce profit taxes. *International Economic Review*, 54(2):665–692.

Hausmann, R., Hidalgo, C., Stock, D., and Yildirim, M. A. (2019). Implied comparative advantage.

Hausmann, R. and Hidalgo, C. A. (2011). The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A., and Yildirim, M. A. (2014). *The atlas of economic complexity: Mapping paths to prosperity*. Mit Press.

Hausmann, R. and Neffke, F. M. (2019). The workforce of pioneer plants: The role of worker mobility in the diffusion of industries. *Research Policy*, 48(3):628–648.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., and Hausmann, R. (2007a). The product space conditions the development of nations. *Science*, 317(5837):482–487.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., and Hausmann, R. (2007b). The product space conditions the development of nations. *Science*, 317(5837):482–487.

Imbs, J. and Wacziarg, R. (2003). Stages of diversification. *American Economic Review*, 93(1):63–86.

Jacobs, J. (2016). *The economy of cities*. Vintage.

Jung, J. and López-Bazo, E. (2017a). Factor accumulation, externalities, and absorptive capacity in regional growth: Evidence from europe. *Journal of Regional Science*, 57(2):266–289.

Jung, J. and López-Bazo, E. (2017b). Factor accumulation, externalities, and absorptive capacity in regional growth: evidence from europe. *Journal of Regional Science*, 57(2):266–289.

Keane, M. (1975). The size of the region-building problem. *Environment and Planning A*, 7(5):575–577.

Kim, H., Chun, Y., and Kim, K. (2015). Delimitation of functional regions using ap-regions problem approach. *International Regional Science Review*, 38(3):235–263.

Kreps, D. M. and Wilson, R. (1982). Sequential equilibria. *Econometrica: Journal of the Econometric Society*, pages 863–894.

Krugman, P. (1991a). Increasing returns and economic geography. *Journal of political economy*, 99(3):483–499.

Krugman, P. (1991b). Increasing returns and economic geography. *Journal of political economy*, 99(3):483–499.

Krugman, P. R. (1991c). *Geography and trade*. MIT press.

Krugman, P. R. (1997). *Development, geography, and economic theory*, volume 6. MIT press.

Lagendijk, A. (2006). Learning from conceptual flow in regional studies: Framing present debates, unbracketing past debates. *Regional Studies*, 40(4):385–399.

Laura, J., Li, W., Rey, S. J., and Anselin, L. (2015). Parallelization of a regionalization heuristic in distributed computing platforms–a case study of parallel-p-compact-regions problem. *International Journal of Geographical Information Science*, 29(4):536–555.

Li, W., Church, R. L., and Goodchild, M. F. (2014). The p-compact-regions problem. *Geographical Analysis*, 46(3):250–273.

Luelfesmann, C., Kessler, A., and Myers, G. M. (2015). The architecture of federations: Constitutions, bargaining, and moral hazard. *Journal of Public Economics*, 124:18–29.

Macmillan, W. and Pierce, T. (1994). Optimization modelling in a gis framework: the problem of political redistricting. *Spatial analysis and GIS*, pages 221–246.

Martin, R. (2010). Roepke lecture in economic geographyrethinking regional path dependence: beyond lock-in to evolution. *Economic geography*, 86(1):1–27.

Martin, R. and Sunley, P. (1996). Paul krugman's geographical economics and its implications for regional development theory: a critical assessment. *Economic geography*, 72(3):259–292.

Maya, J. L. (2013). Modo de desarrollo, organización territorial y cambio constituyente en el ecuador. *Problemas del desarrollo*, 44(174):214–216.

Mealy, P. and Teytelboym, A. (2020). Economic complexity and the green economy. *Research Policy*, page 103948.

Mehrotra, A., Johnson, E. L., and Nemhauser, G. L. (1998). An optimization based heuristic for political districting. *Management Science*, 44(8):1100–1114.

Midelfart-Knarvik, K. H. and Overman, H. G. (2002). Delocation and european integration: is structural spending justified? *Economic policy*, 17(35):321–359.

Moulaert, F. and Sekia, F. (2003). Territorial innovation models: a critical survey. *Regional studies*, 37(3):289–302.

Neary, J. P. (2001). Of hype and hyperbolas: introducing the new economic geography. *Journal of economic Literature*, 39(2):536–561.

Neffke, F. (2009). *Productive places: The influence of technological change and relatedness on agglomeration externalities*. Utrecht University.

Neffke, F., Hartog, M., Boschma, R., and Henning, M. (2014). Agents of structural change. *The role of firms and entrepreneurs in regional diversification. Papers in Evolutionary Economic Geography*, 14.

Neffke, F., Henning, M., and Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic geography*, 87(3):237–265.

O'Clery, N., Curiel, R. P., and Lora, E. (2019). Commuting times and the mobilisation of skills in emergent cities. *Applied Network Science*, 4(1):118.

O'Clery, N., Yildirim, M. A., and Hausmann, R. (2018). Productive ecosystems and the arrow of development. *arXiv preprint arXiv:1807.03374*.

Okubo, T. (2012). Antiagglomeration subsidies with heterogeneous firms. *Journal of regional science*, 52(2):285–299.

Openshaw, S. and Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment and planning A*, 27(3):425–446.

Osborne, M. J. and Rubinstein, A. (1990). *Bargaining and markets*. Academic press.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Ottaviano, G. (2003). Regional policy in the global economy: Insights from new economic geography. *Regional Studies*, 37(6-7):665–673.

Pellenbarg, P., van Wissen, L., and van Dijk, J. (2002). Firm relocation: state of the art and research prospects. Graduate School/Research Institute, Systems, Organisations and Management (SOM).

Randall, A. (2014). Probing the limits of risk-neutral government. *Journal of Natural Resources Policy Research*, 6(1):65–69.

Restrepo, D. E., Church, R., and Duque, J. C. (2020). The p-innovation ecosystems model. *arXiv preprint arXiv:2008.05885*.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2119–2171. Elsevier.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109.

Rubinstein, A. (1985). A bargaining model with incomplete information about time preferences. *Econometrica: Journal of the Econometric Society*, pages 1151–1172.

Schumpeter, J. A. et al. (1939). *Business cycles*, volume 1. McGraw-Hill New York.

She, B., Duque, J. C., and Ye, X. (2017). The network-max-p-regions model. *International Journal of Geographical Information Science*, 31(5):962–981.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, 70(1):65–94.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, pages 312–320.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.

Ulltveit-Moe, K. H. (2007). Regional policy design: An analysis of relocation, efficiency and equity. *European Economic Review*, 51(6):1443–1467.

Vickrey, W. (1961). On the prevention of gerrymandering. *Political Science Quarterly*, 76(1):105–110.

Vogel, R. M. (2000). Relocation subsidies: Regional growth policy or corporate welfare? *Review of Radical Political Economics*, 32(3):437–447.

Wang, S. (2004). From special economic zones to special technological zones. *Changing China: A Geographic Appraisal*, pages 137–156.

Wei, Y. D. (2007). Regional development in china: transitional institutions, embedded globalization, and hybrid economies. *Eurasian Geography and Economics*, 48(1):16–36.

Wydick, B. (2007). *Games in economic development*. Cambridge University Press.

Zhu, S., He, C., and Zhou, Y. (2017). How to jump further and catch up? path-breaking in an uneven industry space. *Journal of Economic Geography*, 17(3):521–545.

Zoltners, A. A. and Sinha, P. (1983). Sales territory alignment: A review and model. *Management Science*, 29(11):1237–1256.