Exploring the Relationship between Geography, Agglomeration Economies, and Firm Performance in Tanzanian Manufacturing

An Honors Thesis for the Department of Economics.

Mia Ellis

Abstract: Agglomeration economies are thought to lead to benefits for clustered firms, an idea that often underpins industrial development schemes. Though the effects and determinants of these agglomerations have been studied thoroughly in many developed countries, there is relatively little work of this kind in the developing country context. Identifying the determinants of firm clusters in Tanzania can help in understanding how agglomeration economies can be formed and sustained. One of the primary theoretical determinants of agglomeration is natural advantage. This paper makes use of data from the 2013 Census of Industrial Production as well as data on geography from Henderson et al. (2017) to test how well geographic characteristics of a district in Tanzania are able to explain agglomeration rates of manufacturing firms in that district. It finds that geographic variables can explain up to 35% of the variation in firm clustering across districts; however, the relationship between geography and agglomeration varies across the agglomeration rates of different sectors and sector-pairings. The paper concludes with a cursory analysis of the relationship between agglomeration and firm performance. It finds some evidence for a positive relationship between agglomeration and productivity, and demonstrates that for certain measures of agglomeration, agglomeration's relationship to productivity varies by firm characteristics.

Tufts University, 2018

Table of Contents

I. Introduction

Tanzania is considered one of Africa's rising stars, and in recent years has seen impressive growth rates (Page 2017). Unfortunately, development of its industry is lagging and slowing progress toward its goal of becoming a middle-income country by 2025 (Rodrik 2014, Page 2017). Manufacturing has the potential to become a high productivity sector, especially given Tanzania's natural resource abundance (augmented recently in light of new gas and mineral resources). If the manufacturing industry's potential is fully realized, it could provide good-paying jobs to many people, reduce poverty, and create sustainable growth in the country. The development of industrial clusters is often suggested a potential way to support growth in the manufacturing industry (MITM 2010).

The link between industrial clusters and firm performance is one that has been studied extensively, and there exists significant evidence of the benefits of agglomeration economies for firms (Ellison & Glaeser 1999, Callois 2008, Combes et al. 2012, Howard et al. 2014). Though the empirical literature on agglomeration economies in developing countries is somewhat limited, there is evidence that agglomeration has benefits for firm performance (Yoshino 2010). Marshall (1920) has argued that industrial agglomerations can affect firm performance in three major ways. The first is through the lowering of input costs; the second is through the increase in the pool of skilled labor; and the third is through technology and knowledge spillovers. These benefits can accrue to firms located in general industrial clusters, clusters of one specific sector, and clusters of two sectors (co-agglomerations). They are likely to be strongest in situations where the firms in the cluster are related enough that they have common inputs, labor needs, or technology.

While this topic has not been studied extensively in Tanzania, the development of industrial clusters is nonetheless a major focus of Tanzania's industrial development policy (MITM 2010). Beginning in 2002, there have been efforts to develop special economic zones (SEZs) and export processing zones (EPZs), following the example of many industrializing countries (Kinyondo et al. 2016). These zones offer favorable laws and regulations with the purpose of attracting investment and new businesses, and may allow firms to benefit from industrial agglomerations. However, very few of these zones in Tanzania are actually operational, they struggle to attract firms, and it has been difficult to identify any measurable benefits (Kinyondo et al. 2016, Newman & Page 2017).

Despite the struggles of SEZs in Tanzania, the fundamental idea of developing industrial clusters where firms can benefit from agglomeration economies is important. It is therefore useful to identify the drivers of agglomeration, which allows us to understand how industrial clusters form and are sustained. This information could help cluster development policies to align with the already-existing incentives for firms to cluster together. Natural advantage is often cited as a possible determinant of industrial clusters, the idea being that some areas might have geographic characteristics that create natural advantages and result in agglomerations. For example, proximity to a harbor may incentivize exporting firms to cluster there. On the other hand, the location of important natural resources might incentive firms that use these resources as inputs to cluster around the resource.

This paper tests for the existence of a relationship between geographic characteristics of a district and manufacturing firm clustering in that district in Tanzania. It uses data from two main sources, the 2013 Industrial Census in Tanzania, and geographical data from Henderson et al. (2017). In addition to identifying the role of geography in driving manufacturing firm clusters in

Tanzania, this paper also tests whether the findings of Henderson et al. (2017) regarding the spatial distribution of economic activity in developing countries holds true for Tanzania.

They find that agriculture-related geographic variables were important in explaining the spatial distribution of economic activity in early-urbanizing countries, but that trade-related geographic variables play a relatively more important role in predicting the distribution of economic activity of late-urbanizing countries. As a result, economic activity in late-urbanizing (developing) countries is not as centered around areas favorable for agriculture; instead, they tend to have their economic activity concentrated on coasts and to have a less developed hinterland, leading to significant spatial inequality in the distribution of economic activity. This paper compares findings on the distribution of manufacturing activity in Tanzania to Henderson et al. 's (2017) prediction. Finally, the paper presents a cursory analysis testing for the existence of a relationship between agglomerations and firm productivity.

The analysis finds that geographic variables can explain up to 36% of the variation in firm clustering across districts. However, the relationship between different geographical characteristics and clustering varies significantly across sectors. Additionally, the paper finds that the distribution of manufacturing activity in Tanzania is not quite what was predicted by Henderson et al. (2017); instead, there exists significant clusters of manufacturing activity throughout the country. The paper's cursory analysis of agglomeration and firm performance finds evidence for a positive relationship between agglomeration and productivity, and demonstrates that for certain measures of agglomeration, agglomeration's relationship to productivity varies by firm characteristics.

The rest of the paper is organized as follows. Section 2 presents the literature on the relationship between natural advantage and agglomeration, and section 3 discusses the data used.

Section 4 presents the methodology and primary results, and section 5 presents a cursory analysis of the relationship between agglomeration and firm characteristics. Section 6 is a brief discussion, and section 7 concludes.

II. Literature Review

There are two commonly identified reasons why industrial clusters form. The first is the benefit of agglomeration economies. This is the argument that clusters form because firms identify the benefits of agglomeration (lower input costs, better access to skilled labor, and knowledge transfers) and therefore choose to locate in a cluster. The second is natural advantage, which argues that location fundamentals are responsible for clustering. Most studies on this issue find evidence for both causes. The construction of agglomeration measures varies across the literature; section 3 describes some of the differences and carefully defines the measures of agglomeration and co-agglomeration used in this paper.

Ellison and Glaeser (1999) test for the relationship between natural advantage and agglomeration in the United States. They include the cost of inputs such as electricity, natural gas, coal, agricultural and livestock products, and lumber, as measures of natural advantage. They also include variables intended to measure transport costs, which are related to export/import intensity and being on a coast. They find that their measures of natural advantage predict around 20% of industrial pairwise co-agglomeration, and conjecture that all natural advantages are likely responsible for about 50% of geographic concentration.

Also in 1999, Kim studies the relationship between natural resource endowments and firm clusters in the United States from 1880-1987. He specifically tests whether the existence of agriculture, tobacco, timber, petroleum, and mineral resources can explain agglomeration. Kim finds that these variables, along with factors related to the availability of labor and capital, can explain between 70-86% of the variation in agglomeration, on average across sectors. The explanatory power of these factors does decline over time, and Kim suggests that as lower transportation costs increase input mobility, natural advantage may play less of a role in explaining agglomeration, while spillover effects may become more important.

Rosenthal and Strange (2001) also attempt to understand the determinants of agglomeration. They look at both the impact of agglomeration externalities (labor market pooling, shared knowledge, etc.) and natural advantage on geographic concentration. Their controls for natural advantage include energy per $ shipment, natural resources per $ shipment, and water per $ shipment. They also proxy for transport costs with a variable for inventories per $ shipment. They find that natural advantage and transport costs significantly affect agglomeration at the state level, but have little effect at lower levels of geography. They find relatively low R-squared values, which suggest that their proxies only explain a small fraction of variation in agglomeration.

The research done by Rosenthal and Strange (2001) raises an interesting point, which is the possibility that geography only explains agglomeration at broader geographic boundaries, and cannot explain clustering at smaller scales. This intuitively makes sense, as geography is likely consistent across districts and therefore would not be responsible for heterogeneity in clustering within a district. It is therefore likely that other factors such as agglomeration externalities or industrial policy, explain the finer details of agglomerations. Lu and Tao (2009) reinforce this idea with their study on the trends and determinants of agglomeration in China. They find that resource endowments related to agriculture and mining have positive and significant effects on agglomeration, but that these effects are stronger and somewhat more significant at larger geographical boundaries.

Howard et al. (2015) look at the relationship between agglomeration externalities, natural advantage, and clustering in Vietnam. Similar to other papers in the literature, their measures of natural advantage are entirely based on costs. They test the relationship using several measures of agglomeration, including their own co-location index, the pairwise co-agglomeration index established by Ellison and Glaeser (1997), and a measure of absolute agglomeration between sectors. They find that when measuring absolute co-agglomeration (clusters measures in terms of absolute size) as opposed to co-agglomeration (clusters measured in terms of their deviation from general industrial concentration), natural advantage plays a much bigger role in explaining variation.

Finally, Henderson et. al. (2017) looks at the relationship between geography and the distribution of overall economic activity around the world. This is slightly outside of the scope of the literature on agglomeration economies, because the outcome variable of interest is general economic activity rather than clusters of industrial activity based off of firms or employees. However, it still presents important insights for this research. They regress the distribution of economic activity as measured by lights on a set of geographic variables divided into agriculture, trade, and base variables.

Overall, they find that geography alone can explain 47% of the variation in the distribution of lights. They find that agriculture-related geographic variables play a larger role in explaining the distribution of economic activity than trade-related geographic variables, but that the trade-related variables play a relatively more important role in explaining the distribution of economic activity in late-urbanizing countries than in early-urbanizing countries. As mentioned in the introduction, this is because activity in early-urbanizing countries tended to center around areas of agricultural production, whereas lower transport costs for late-urbanizing countries allowed for greater mobility of agricultural goods and also led to an increased importance of low transport costs, which led to relatively greater concentration of activity in low transport-cost areas. This means that on average, for late-urbanizing countries, the distribution of economic activity is more unequal than in early-urbanizing countries, with it being more concentrated on coasts and less so in the hinterland.

Based on the literature, it is commonly found that natural advantage plays a significant role in explaining agglomeration, and that it is relatively more important in explaining agglomeration at broader spatial boundaries. This research contributes to this literature in a few major ways. First, most studies use indirect measures of natural advantage, which are often related to costs of resources, utilities, or transport assumed to proxy for natural advantages in the region. Low R-squared values are common in these studies, which suggest that the included variables do not sufficiently explain the variation in agglomeration – potentially because the proxies used to measure natural advantage do not effectively capture the true natural advantage in a region. This paper instead follows the approach of Henderson et al. (2017) and uses explicit measures of geography to measure natural advantage.

These explicit measures of geography are useful because they are arguably exogenous from agglomeration. In most of the literature, proxies for natural advantage based on prices could very well be endogenous – as more firms cluster in an area, this could lead to the agglomeration benefit of lower transport costs or lower input costs, which in turn could influence the natural advantage proxies used throughout the literature. For example, the value of natural resources per shipment could be related to the existence of natural resources in the area (which would truly be natural advantage), but could also be due to the benefits of agglomeration economies. This paper therefore benefits from the use of exogenous geographic variables, and it can be argued that the regression results are truly causal.

It is important that all of the variables included in this analysis are truly measures of natural geography; for example, if the binary variable for close proximity to a natural harbor were a binary variable for being within 25 kilometers of an actual harbor, this would not necessarily be exogenous. However, each variable included in the analysis is a measure of natural geography in the district, and we can therefore assert that there is no risk of reverse causality between natural advantage and agglomeration. Furthermore, the measures of agglomeration control for overall industrial concentration (described in section 3) so there is not a risk that the geographic variables in this analysis are just explaining industrial concentration. Therefore, it is reasonable to argue that the relationship between natural advantage and agglomeration in this paper is causal.

Additionally, most of the literature conducts analysis using sector or sector-pairings as observations. This ignores some of the potential complexity in relationship between natural advantage and agglomeration, because the results do not show how the effect of natural advantage might vary across sector or sector-pairings. Instead, this paper conducts analysis using districts as observations, regressing different rates of agglomeration and co-agglomeration in each district on the geographic characteristics of that district. This allows for us to see how the effect of natural advantage varies across different measures of agglomeration, and especially to see that the relationship between one geographic characteristic and clustering might vary with different measures of agglomeration.

III. Data & Variables

a. Data Sources

This paper uses data from two different sources. The first is the 2013 Census of Industrial Production, which contains information on 12,792 manufacturing firms around the country. There were two levels of stratification in the census design. The first was a take-all stratum, in which all establishments with ten or more employees as well as some smaller establishments in regions with less than a hundred firms in the same activity, were surveyed. The second was a sampling stratum of firms with less than 10 employees, in regions where 80 or more establishments were found in the same activity. This sampling focused on four major activities: grain milling, tailoring, welding, and furniture. It is therefore not a true census, and sampling weights were used in the construction of all variables and regression analyses, where appropriate.

The second source of data used in this paper is a dataset from Henderson et al. (2017) containing geographical information for Tanzania. These variables can be split into three groups: base variables; variables affecting trade; and, variables affecting agriculture. The base variables include a measure of the stability of malaria transmission and ruggedness. The trade variables include distance to the nearest coast, natural harbor, large lake, and navigable river, as well as binary variables for whether a region is very close to a coast, natural harbor, large lake, or navigable river. Finally, the agricultural variables include land suitability for agriculture, temperature, precipitation, length of growing period, and elevation, as well as biome type.

The geographic variables are defined by a latitude and longitude point, whereas the firm-level data is assigned to a region, district, and ward. In order to merge these datasets, I first map the geographic data by their latitude and longitude point on an administrative map of Tanzania, and then attach them to the district in which they are located. I then collapse the data in order to aggregate this information to the district level. For districts that have more than one point, the resulting information is an average of all of the points contained by the district. In the cases that this results in the average value of a binary variable, I recode every value greater than zero to be a one. For example, consider the case of the binary variable for an area being coastal. If a district contains three latitude and longitude points, only one of which is coastal, then the average would be .33. However, the district is still coastal by definition. Therefore, I recode the binary variable coastal to have a value of one.

I use the census data to calculate measures of agglomeration and co-agglomeration (described below) and then aggregated the data to the district level. I then merged the two datasets. Because the geographic data is set up according to a grid, districts that are small and close together pose a risk of not each having a latitude-longitude coordinate in them. As a result, the geographic data only matches up to 147 districts in the industrial census, out of a total 163 included in the census (there are 169 districts total in the country). Each observation of geographic variables does match to a district included in the census, so the missing districts are only a result of missing geographic data (e.g. the geographic data does not match up with any of the six districts excluded by the census). Two major districts with missing geographic data were Kinondoni and Ilala, which make up two-thirds of the districts in Dar es Salaam (the included district being Temeke). Excluding the firms in these districts results in a loss of potentially important information, and given that Dar es Salaam covers a small geographic area, I assign the geographic data from Temeke to both Kinondoni and Ilala, allowing them to be included in the analysis. This results in 149 district-level observations of cluster rates and geographic characteristics.

b. Choice of Explanatory Variables

Because of the low number of districts in Tanzania, there are not many observations available for the analysis and this significantly limits its statistical power. Even if the geographic data had perfectly matched to the census data, this would still only provide 163 observations. Therefore, in order to preserve as much statistical power as possible, the analysis makes use of just six of the geographic variables discussed in Henderson et al. (2017). It looks at the relationship between agglomeration and a district being coastal, being within 25 kilometers of a natural harbor, within 25 kilometers of a large lake, average monthly precipitation, land suitability (defined as the probability that a grid is cultivated), and an index of ruggedness. Table 1 presents summary statistics for these variables.

These six variables were chosen based on intuitive reasoning of the geographic characteristics most likely to impact agglomeration and the characteristics most likely to be important in the Tanzanian context. Though Henderson et al. (2017) finds that agriculture-related geographic characteristics are much more important in explaining the distribution of economic activity than the trade-related characteristics, this paper hypothesizes that trade-related geographic variables will play a more important role in explaining industrial clusters. Tanzania is a coastal country, and it contains several lakes on its borders with other countries (Lakes Victoria, Tanganyika, and Nyasa). Both these coasts and lakes are home to important ports for inter-regional and international trade, and therefore may attract firms that engage in trade. Similarly, being within 25 kilometers of a natural harbor may proxy for lower transport costs.

While I expect that the trade-related geographic characteristics will be most important in explaining the distribution of manufacturing clusters, it also possible that agriculture-related variables could affect industrial clusters for sectors that are closely related to agriculture by providing an important source of inputs (such as food manufacturing). Amongst the agricultural variables, Henderson et al. (2017) finds that land suitability is the largest contributor in explaining the variation in the distribution of economic activity. I therefore choose to include land suitability as my measure of agricultural conditions, a choice that is further reinforced by the fact that land suitability exhibits greater variation than the other agricultural variables within the country.

Henderson et al. (2017) suggests that precipitation could lead to increased activity in an area due to its positive effect on agriculture (therefore increasing access to agriculture-based inputs), but could also lead to decreased activity as rainy conditions may be less desirable. This may be especially true in the context of the Tanzanian manufacturing sector. Many of the micro firms in the country operate outdoors and in backyards, and increased precipitation could make these activities more difficult. Finally, ruggedness could affect both the availability of agriculture-related inputs and trade costs by increasing transport costs.

c. Measuring Agglomeration

In this paper, agglomeration is considered to be a rate of firm clustering, net of overall industrial clustering. Essentially, it is a measure of how the concentration of a specific sector or sector-pairing deviates from what we would expect from overall industrial concentration in a district. For example, consider a district that contains 4% of all industrial firms. If this district also contains 4% of firms in given sector, then we would not consider this to be an agglomeration because the concentration is simply what we would expect from the distribution of overall industrial activity. However, if the district contains 10% of firms in the sector, then we would consider that sector to be agglomerated in the district. Similarly, if the district contained just 1% of firms in the sector, we would consider that sector to be dis-agglomerated in the district.

This paper explores a few different agglomeration measures. First, it looks at industry-wide and sector-specific measures of agglomeration, and then at measures of co-agglomeration for select sectors. Co-agglomeration is similar to agglomeration, but represents the extent to which two sectors are clustered together, net of overall industrial concentration. There is an extensive literature surrounding strategies for developing the most appropriate measures of clustering (Yoshino 2010). Such measures can be divided into two broad categories: continuous measures based on measured distances between firms; and discrete measures based on shares aggregated to some geographic boundary. This paper is limited to discrete measures, as the census data lacks the location specificity required for a continuous measure.

Agglomeration measures in the literature can be further divided into two types of measures: those that are based on employment shares; and, those that are based on firm shares. In a country like Tanzania, where labor is mostly unskilled, it is likely that employers (rather than employees) are the driving force of agglomeration benefits (Howard et al. 2016). Additionally, as most firms in Tanzania are micro and small sized, an employment-based measure of agglomeration would underestimate the true concentration of firms. Therefore, this paper makes use of firm-based measures of agglomeration. Finally, as discussed above, any measure of agglomeration should take into account overall industrial concentration.

The majority of the literature on agglomeration utilizes indexed measures of agglomeration for industries aggregated across districts, meaning that the units of observation are industries or industry-pairings. They then calculate explanatory variables based off of prices for different factors used by said industries. These price-based measures are taken to proxy for natural advantage as well as the more typical characteristics of agglomeration economies such as lower input prices, labor pooling, and knowledge transfer. This study differs significantly from the described literature in that it does not aggregate agglomeration rates or natural advantage measures across districts. Instead, it calculates distinct measures of agglomeration for each district and uses district-level factors in attempt to explain said measures of agglomeration.

As stated, this paper makes use of three different types of agglomeration measures. The first is a measure of overall industrial agglomeration in a district, given by $agg_i = (x_i - \bar{s} + 1) * \frac{1}{A_i}$ where x_i is the share of all manufacturing firms in district i, $\bar{s}$ is the average share of manufacturing firms across all districts, and A_i is the area of district i. This measure tells us how much more (or less) concentrated firms are in district *i* than they are on average across districts in Tanzania. This difference is weighted by district area to account for the fact that a larger district will likely have a greater share of firms than a smaller district simply due to size, and the higher share might not indicate true geographic concentration.

The term also includes a plus one, which is important due to the weighting by district size. Consider a case of two districts, in both of which the concentration of firms is .02 less than the average concentration of firms across all districts (e.g. dis-agglomerated). The difference in both districts is the same, but one district is large and one is small. Given that the share of firms is the same, we would assume that the true degree of firm clustering in the small district is greater than in the large district. However, weighting by district size would result in a measure of agglomeration that is greater in the large district than in the small district. This produces a contradiction, because we expect that the small district should have a greater (less negative) value of agglomeration. This example makes clear that when weighting by district size, negative values of agglomeration produce problematic results. Therefore, the one is added to the term in order to ensure that it is strictly positive, and such issues are avoided.

The second is a measure of sector-specific agglomeration, given by $agg_{i,m} = s_{i,m} - x_i$, where $s_{i,m}$ is the share of an sector m's firms in district *i* and x_i is the share of aggregate manufacturing firms in district i. This essentially measures how concentrated sector m is in district i, net of overall industrial concentration, producing a measure of sectoral agglomeration. This measure is positive if the sector *m* is more concentrated in district *i* than overall industry (e.g. agglomerated), is equal to zero if the concentration of sector *m* is equal to that of overall industry, and negative if sector m is less concentrated in district *i* than overall industry (e.g. dis-agglomerated).

In this case, it is not necessary to weight by district size because the difference is within the district – so the x_i term essentially controls for district size. For example, consider two districts: one is a large district with 10% of all firms and 14% of firms in a given sector, and the other is a small district with 2% of all firms and 6% of firms in a given sector. If we did not difference out the rate of overall industrial concentration, x_i, then we would inaccurately believe that the sector in the larger district is more agglomerated, when in reality the sector is equally agglomerated in the two districts. However, because we account for the fact that larger district has a greater share of overall firms, we negate this risk and therefore do not need to weight by district size.

Finally, the third measure I construct is a co-agglomeration measure given by $agg_{d,m,n} = (s_{im} - x_i + 1)(s_{in} - x_i + 1)$, where s_{im} is the share of sector *m*'s firms in district *i*, s_{in} is the share of sector *n*'s firms in district i, and x_i represents the industry-wide firm share in district i. For the same logic described above, it is not necessary to weight by district size. The addition of one to both terms is done in order to ensure that neither is negative, so that the measure is strictly increasing as the share of firms in either sector in district *i* increases. For example, consider the case where two sectors are both dis-agglomerated in district *i,* meaning that their shares of firms are less than the share of overall industrial firms. Without the addition of the one to both terms, then the measure of co-agglomeration would be the product of two negative terms and would therefore be positive. This would incorrectly indicate that the two sectors were co-agglomerated in district *i,* when the reality is that they are both dis-agglomerated. Given that each term is the difference between shares, each is bounded between zero and two and the resulting rate of co-agglomeration is bounded between zero and four. A higher value indicates a greater degree of co-agglomeration.

I calculate each type of measure using the total 12,796 firms in the census. While the geographic analysis is restricted to firms in the 149 districts for which there exists geographic data – of which there are only 10,806 – the construction of the agglomeration measures is most accurate if it uses information from all of the existing firms. If there are no firms in a given sector in a district, I assign a missing value to that agglomeration rate rather than assigning a minimum value. This potentially loses some information, because I am not considering what geographic characteristics might discourage firms from locating in a district altogether; rather, I am just considering what geographic characteristics incentivize clustering, conditional on the fact that firms in the sector being considered exist in the district. I next aggregate these measures to the district level and eliminate the 14 districts for which there is no geographic data, and then standardize each measure so that the regression coefficients will be easy to interpret. Table 2 presents summary statistics for these agglomeration measures.

Figure 1 is a heat map illustrating the distribution of population across Tanzania. It is useful as a reference when looking at figures 2-7, which is a set of heat maps showing how each measure of agglomeration and concentration varies across districts in Tanzania. These maps are constructed using the standardized versions of each cluster measure. The degree of clustering is broken up by quintiles, and darker shades represent a more clustered district. I create these maps for agglomeration measures of overall industry and of the following five sectors: food; apparel; wood excluding furniture; fabricated metals; and, furniture. These sectors were chosen for two reasons. The first is that they are by far the largest sectors in the country in terms of number of firms. The manufacture of food products contains 40% of all manufacturing firms; apparel is the second largest sector with 27% of firms; and furniture is the third largest with 14% of firms. Fabricated metals and wood excluding furniture consist of 7.8% and 3.7% of all manufacturing firms, respectively.

The second reason for the choice of these sectors is that they are some of the highest-potential sectors in the country identified by a 2013 World Bank report on light manufacturing in Tanzania. This report highlights the apparel, wood and wood products, and agro-processing sectors as those having significant potential for job creation and growth, and many of the recommendations in this report are related to developing clusters within these sectors (Dinh & Monga 2013). Furthermore, several of these sectors are specifically targeted by the Tanzanian Integrated Industrial Development Strategy (IIDS) 2025. The agro-processing and apparel sectors in particular are two of the primary sectors focused upon by the IIDS, and there have already been efforts to establish clusters in these sectors (MITM 2010). The strategy also places importance on the potential of light industry and specifically metal and minerals processing (ibid).

Figures 2-7 show that the distribution of clusters varies significantly across sectors. While the highest rates of overall industrial agglomeration are dispersed throughout the country and clearly located in small districts, the sector-specific heat maps show significantly different distributions across sectors. The food sector is mostly clustered in a band running from the north to the south through the middle of the country. The apparel sector is highly agglomerated in the north and north-west, while the wood excluding furniture sector is clustered throughout the south. The districts in which the fabricated metals sector is most clustered are scattered throughout the country, and the furniture sector is most agglomerated in the south.

This information leads to a few important lessons. First, the majority of the country seems to be home to some sort of cluster. Even those areas that have a significantly lower rate of overall industrial concentration may contain relevant clusters. For example, figure 2 shows that firms are not particularly concentrated in the center-west of the country; however, figures 4 show that there does exist a significant degree of agglomeration of firms in the apparel sector in that area. Finally, some sectors appear to be mostly agglomerated in one continuous section of the country (food, apparel, furniture), while others have discrete clusters existing throughout the country (wood excluding furniture, fabricated metals). This is important because it may be more complicated and therefore costlier to create optimal circumstances for sectors with more scattered clusters.

Figures 8a-c further reinforce the idea that cluster distributions are heterogeneous across overall industry and specific sectors. The heat maps show the distribution of co-agglomeration rates for ten different sector pairs. For example, looking at the food sector pairings shows that the food sector is co-agglomerated with the other sectors in distinctly different areas. It is most heavily co-agglomerated with the apparel sector in the north, with the wood excluding furniture sector in the mid-south, in a more scattered pattern with the fabricated metals sector, and is co-agglomerated with the furniture sector in a large block of southern districts. This indicates that the co-agglomerative dynamics of one sector may differ significantly across sector pairings. The other heat maps in figures 8b-8c confirm that this is the case across sector pairings.

IV. Methodology & Results

a. Results for Clusters of All Firm Sizes

This paper makes use of ordinary least squares (OLS) regressions to examine the relationship between geographic variables and different measures of firm clustering. Though I run numerous iterations of this model, the general form of the relationship is given by the following equation:

Eq. 1 $$cluster_i = \beta_0 + \sum \beta_j \alpha_{m_i} + \epsilon_i$$

I first test for the relationship between geography and industry-wide firm agglomeration, where $cluster_i$ is the industry-level measure of agglomeration for a district *i,* and α_{m_i} is a vector of the geographic variables of interest measured at the district level. These include three binary variables – whether a district is coastal, within 25 kilometers of a large lake, and within 25 kilometers of a natural harbor – and standardized versions of precipitation, land suitability, and ruggedness. This standardization is intended to allow for easier interpretation of the regression coefficients. I run Shapley decompositions for each regression, which show the percent contribution of each explanatory variable to the total R-squared. This is useful in allowing us to understand how big of a role each explanatory variable plays in explaining the variation in rates of clustering.

Table 3, column 1 presents the results of this regression. I next test for the relationship between geography and sector-specific firm agglomeration for each of the five sectors previously identified. In this form, $cluster_i$ is the measure of sector-specific agglomeration for a sector s in district i. This outcome variable measures the rate of firm agglomeration in a specific sector, net of the overall industrial concentration in the region. These results are presented in table 3, columns 2-6. Table 4 presents a Shapley decomposition that shows the relative contribution of each geographic variable to the overall R-squared.

Next, I test for the relationship between geography and sector co-agglomeration, where $cluster_i$ is the pairwise co-agglomeration measure for two sectors x and y in district i. This tests for the ability of geography to explain why two different sectors might co-locate. I run this test for co-agglomeration measures of pairs of the five sectors identified above. Tables 5a and 5b present the results of these regressions, and table 6 presents the Shapley decompositions of these regressions.

These results show that geography explains between 1-35% of the variation in agglomeration and co-agglomeration rates. This suggests that geography can play an important role in cluster development, but that this importance varies significantly across sectors and sector pairings. The results further show that the effects of geographic characteristics on clustering vary in terms of significance and sign across agglomeration and co-agglomeration measures for different sectors and sector pairings, indicating that the relationship between geography and clustering varies by sector. For example, coastal status is negatively associated with increased agglomeration of the food sector, but positively associated with agglomeration of overall industry and the apparel, wood excluding furniture, fabricated metals, and furniture sectors.

The Shapley decompositions presented in table 4 show that coastal status is, on average, the biggest contributor in explaining agglomeration rates. However, the other explanatory variables play important roles in explaining the variation of agglomeration rates across sectors – for example, ruggedness explains 47% of the variation in overall industrial agglomeration, while precipitation explains 43% of the variation in furniture sector agglomeration. This tells us that the importance of geography varies across different types of agglomeration, further supporting the idea that the relationship between geography and clustering varies across different measures of agglomeration. Furthermore, these results tell us that trade-related geographic variables play a dominant role in explaining the variation in manufacturing firm agglomeration, as compared to agriculture-related geographic variables.

The co-agglomeration results in tables 5a and 5b further reinforce the idea that the relationship between geography and agglomeration is not consistent across sectors. For example, no variable exhibits a consistent effect across the regression results. Precipitation is positively correlated with co-agglomeration for all measures that it is significantly related to, but it is only significant for three pairings. None of the other explanatory variables even exhibit a consistent sign across sectors. Furthermore, the Shapley decompositions in table 6 illustrate the irregularity in the role that each geographic variable plays in explaining the variation in co-agglomeration. Similar to the agglomeration results, the decompositions show that the most important variable, on average, is coastal status. However, the other explanatory variables also play varyingly important roles in explaining the variation of different co-agglomeration measures. Once again, the trade-related geographic characteristics are relatively more important in explaining variation than agriculture-related characteristics.

The primary lesson from these results is that the relationship between geography and clustering varies significantly for different measures of clustering across different sectors and sector pairings. Geography can offer important insights into the determinants of firm clusters, but these insights may not be transferrable across different types of clustering. Furthermore, the importance of geography varies significantly across different cluster measures. For example, geography explains just 13% of the variation in apparel sector agglomeration, but 32% of the variation in food sector agglomeration. Therefore, to develop a deeper understanding of how and why industrial clusters form, analysis should take into account the specific context of the sector or sector pairing in question.

Doing so can enable sector-specific trends in the relationship between geography and agglomeration to be identified, in spite of the seemingly heterogeneous relationship between geography and firm clusters. For example, the results suggest that increased precipitation is consistently positively associated with clustering in the furniture sector. This is true for both furniture sector agglomeration and co-agglomerations. This could be explained by a tendency of the furniture industry to cluster around areas with more forests, if higher average precipitation facilitates wood production. This would be an example of geography creating natural advantage through access to inputs.

Furthermore, the rate of agglomeration for the furniture sector is negatively correlated with proximity to both natural harbors and large lakes. This might suggest that furniture sector firms are not attracted to trading hubs, perhaps because most products are sold domestically. In another example, clustering of food sector firms, both in terms of agglomeration and co-agglomeration measures, is consistently negatively associated with coastal status. This fits into the logic of Henderson et al. (2017), which suggests that activity dependent on agriculture is more likely to be located in the hinterland of a country. Future research into the dynamics of firm clusters should be cognizant of potentially important sectoral contexts, and should not treat overall industrial firm clusters as homogenous.

b. Results for Firms with 10 or more Employees

The Tanzanian manufacturing industry is dominated by micro firms. Firms with ten or more employees make up just 2% of all manufacturing firms in the country. Therefore, it is possible that the relationship between geography and clustering is different for clusters defined just by firms with ten or more employees. Using the same construction of agglomeration rates as discussed above, this section analyzes the relationship between geography and clusters defined by firms with ten or more employees. Given that there are only five such firms in the apparel sector, this analysis is limited to overall industry and the food, wood excluding furniture, fabricated metals, and furniture sectors. There are not enough observations to analyze co-agglomeration patterns, so the analysis is further limited to just agglomeration rates. Table 7 shows the results of regressing the geographic variables on agglomeration rates, and table 8 presents the Shapley decompositions of these regressions.

In general, it appears that coastal status is the most significant geographic variable in explaining clusters of firms with ten or more employees; however, its sign is not consistent across sectors. Though many of the geographic variables lose significance or change slightly in magnitude in this regression when compared to the results of agglomerations of all firm sizes, the results are generally consistent. The only contradictory finding is that coastal status is significantly positively correlated to increased agglomeration of all firms in the wood excluding furniture sector, but significantly negatively associated with increased agglomeration of firms with ten or more employees in the same sector. The Shapley decompositions confirm that once again, coastal status is the most important explanatory variable, on average. In fact, the role of each explanatory variable in explaining the variation in agglomeration is quite similar to their roles in explaining the agglomeration of all firms, presented in table 4. The only major differences arise in the fabricated metals and furniture sector; however, these results should not be given much weight as geography does not explain any of the variation in the agglomeration rates of these sectors.

These results suggest that the effect of geography on firm agglomeration for larger firms is generally similar to its effect on firm agglomeration of all firm sizes; however, there is some evidence that there exist differences in the role of geography. It is worth noting that the lack of significance of many of the geographic variables shown in table 7 may be a result of the low number of observations; the wood excluding furniture, fabricated metals, and furniture sectors are all located in less than 20 districts. With more observations, we would be better able to know how the effect of geography on firm clusters varies for clusters defined by different firm sizes. The low number of districts that contain large firms raises another point, which is that this may indicate that there is not much work to be done by the government in order to create industrial clusters of large firms. Instead, for most sectors, the large firms are already clustered together in a small number of districts and the government can just work to support these already-existing clusters.

V. Agglomeration & Firm Characteristics

The fundamental idea underpinning this research is that agglomerations can be an important source of benefits for firms, and can help promote industrial development. There is existing literature that supports this idea. For example, in Ethiopia, Abebe et al. (2017) make use a natural experiment and find that the presence of a large, foreign firm in a region leads to increased total factor productivity for nearby firms. They note that both labor flows and learning by observation play a role in this knowledge transfer. Chhair and Newman (2014) use data from Cambodia to examine the effect of agglomerations on firm productivity, specifically through the mechanisms of competition and spillovers. They find strong, negative effects of increased competition on productivity, largely for formal and manufacturing firms. However, they also find that there exist productivity spillovers from agglomeration, largely for informal and manufacturing firms. This may be important in the Tanzanian context, where the majority of firms in the manufacturing sector are small and informal.

Also in Ethiopia, Siba et al. (2012) uses panel data to demonstrate that localization economies lead to lower prices and increase productivity of firms in the cluster; however, they do not find any relationship between the clustering of firms that produce different products and productivity. Similarly, evidence from UNIDO (2009) suggests that clustering of related firms has a strong, positive impact on firm-level productivity. This implies that it is important to look at sector-specific and co-sector agglomerations as well as overall industrial agglomeration.

In general, overall industrial agglomeration could have benefits for firms through its effects on factors that affect firms regardless of sector, such as lower costs of transport. Sector-specific and co-agglomerations are more likely to have benefits for firms through their impact on common resources. For example, firms in the same sector clustered together may benefit from increased access to common inputs and labor needs. Co-agglomerations are likely to have a positive impact on firms in sectors that are at least somewhat related to one another. For example, the wood excluding furniture and furniture sectors likely have common inputs, types of labor, and technologies. Therefore, firms in these sectors could benefit from co-agglomerations. On the other hand, the apparel and fabricated metal sectors likely have very different inputs, labor types, and technologies. As a result, firms are less likely to benefit from co-agglomeration of these two sectors.

One of the major problems underlying studies of agglomeration and firm performance is endogeneity. The primary cause of this endogeneity is the existence of reverse causality, which arises from the possibility that an area is more productive and that leads to firms locating in the area. It is also possible that some omitted variable such an industrial policy can both create agglomerations and benefit firm performance. It is therefore difficult to establish causality without making use of some natural experiment, randomization, or panel data.

This paper conducts a cursory regression analysis of the relationship between agglomeration and firm performance, in an attempt to understand whether there exists a significant relationship between agglomeration and firm productivity in Tanzania. I include interaction terms of agglomeration rates and firm characteristics in this regression, to examine whether the relationship between agglomeration and productivity might vary by firm characteristics. For example, micro firms (defined as those with fewer than ten employees) may not be in a position to take advantage of some of the theoretical benefits of agglomerations, such as increased access to skilled labor. On the other hand, some findings in the literature suggest that informal firms benefit more from agglomeration economies than formal firms (Chhair & Newman 2014). Due to issues of endogeneity discussed above, this analysis is no way purports to identify a causal relationship. It simply attempts to explore the existence of possible correlations between agglomerations, firm characteristics, and firm productivity. This relationship is modeled by the following equation:

Eq. 2 $$y_{i,d} = \beta_0 + \beta_1 ag_d + \beta_2 micro_i + \beta_3 (micro_i * ag_d) + \beta_4 lic_i + \beta_5 (lic_i * ag_d) + \varepsilon$$

where $y_{i,d}$ is firm productivity measured by value-added per worker for firm *i* in district d, ag_d is the rate of agglomeration in district d, $micro_i$ is a dummy variable indicating that a firm is micro (less than 10 employees), and lic_i is a dummy variable indicating that a firm has a license (proxy for formal status). Value-added per worker is defined as the value-added of a firm divided by the number of paid employees, where value-added is calculated as the difference between gross output and total intermediate consumption. Standard errors are clustered at the district level. I test this relationship using measures of both agglomeration and co-agglomeration. The measures of clustering used and the sample of firms included in the regression are the same as those that were used for the examination of the relationship between geography and clustering, so that the results are more comparable. Results of these regressions are found in tables 9-10b.

Table 9 shows that overall industrial agglomeration is not significantly correlated to value-added per worker, nor are either of the interaction terms. This is reasonable, considering that many of the theoretical benefits of agglomeration come from shared needs. Across different sectors, the relationship between agglomeration or co-agglomeration and firm productivity is consistently positive, for all those coefficients that are significant. While co-agglomerations do not appear to be significantly related to productivity for most sector pairings, it is positively correlated to increased productivity for the wood excluding furniture and furniture pairing. This is in line with the previous logic that co-agglomerations will benefit firms if the co-agglomerated sectors have linkages. These results support the idea that clusters may potentially benefit firms located in the clusters, but reveal the fact that the significant of this relationship is not consistent across different types of clusters.

The interaction terms are significant across several agglomeration and co-agglomeration measures, and are both consistently negative. If the results were causal, this would suggest that micro-sized firms benefit less from agglomeration economies than larger firms; however, the net effect is still positive. Similarly, the negative sign on the interaction between licensing and agglomeration and co-agglomeration rates would suggest that licensed firms benefit less from agglomeration economies than non-licensed firms, though once again, the net effect of agglomeration would likely be positive. This is consistent with the finding in the literature that informal firms benefit more from agglomeration economies.

These results suggest that while agglomerations may not always be significantly related to productivity, when significant, the relationship is consistently positive. There is evidence that the relationship between agglomeration and productivity varies by firm characteristics such as micro and formal status, and this trend is actually consistent across all measures. These findings present evidence that agglomeration can be positively correlated to productivity, but that the existence of this relationship varies across sector and sector pairings. Furthermore, it presents evidence that the relationship between agglomeration and productivity varies by firm characteristics. These findings offer a somewhat different lesson from those of the relationship between geography and clustering, because in this case the signs of each of the explanatory variables of interest are consistent across types of clusters.

VI. Discussion

There are a few weaknesses in this paper to discuss. The first is the use of a discrete measure of agglomeration. Discrete measures suffer from the fact that they define agglomeration at an arbitrary border and therefore pose the risk of underestimating clusters. For example, if there is a large industrial cluster in which half of the firm lie on one side of a border and the other half on the other, the resulting measure of agglomeration will cut the cluster into half. The heat maps in figures 2-8c present evidence that this issue exists, for some sectors. For example, the food and furniture sectors are most heavily co-agglomerated in several districts bordering one another, in the south of the country. Analysis of these co-agglomeration patterns fails to take into the fact that these sectors are co-agglomerated across districts. Unfortunately, without precise information on firm location, this issue is not solvable.

Furthermore, the discrete measures of agglomeration and concentration used in this paper lose significant variation by aggregating to district-level measures, rather than having firm-level measures based on a firm's proximity to surrounding firms. Also, spatial autocorrelation is potentially a concern. Unfortunately, there is no way to correct for these issues in this paper. However, future research would benefit from access to specific firm-location data.

Another weakness is the lack of statistical power combined with the source of the geographic data. Because the geographic data was not administratively collected, it was not necessarily assigned to populated areas. Though I attempted to conduct analyses at the ward level, this proved to be impossible because the geographic data was primarily attached to wards with very few firms. With only 149 observations at the district level, the analysis lacks statistical power. This is what led to the decision to exclude most of Henderson et. al. 's (2017) geographic variables, some of which may have provided interesting insights into the relationship between the geography and firm clustering. Furthermore, excluding most of the variables used by Henderson et al. (2017) makes it more difficult to compare results, as the results in this paper may be underestimating the importance of agriculture-related geographic variables. Finally, it is problematic that this paper excludes 14 districts from its analysis. Because of this, we cannot be sure that the results hold true for all of Tanzania. Though the districts were excluded randomly, by the arbitrary grid of latitude and longitude points defining the geographic variables, it is still possible that some of them contained relevant agglomerations.

This paper finds that trade-related geographic variables explain a significantly greater degree of the variation in manufacturing firm clusters than agriculture-related variables. This finding differs from the findings of Henderson et al. (2017), who find that agriculture variables are consistently more important than trade variables in explaining the distribution of economic activity. This difference may arise because this paper's analysis excludes many of the variables used by Henderson et al. (2017). However, it is also possible that this difference is due to the outcome variable in question. Henderson et al. (2017) looked at the distribution of overall economic activity, as measured by lights, whereas this analysis looks at the distribution of manufacturing clusters. It is reasonable that trade-related variables play a bigger role in explaining variation in manufacturing clusters than they do overall economic activity, given the importance of trade for many manufacturing firms.

Another insight from this research, in relation to the findings of Henderson et al. (2017), is that it reveals significant within-country heterogeneity in the effect of geography on industrial activity. For example, though Henderson et al. (2017) finds that coastal status leads to increased economic activity around the world, this paper finds that it is negatively related to firm clusters in many cases. This is likely explained by the fact that clustering of specific sectors is driven by distinctive characteristics of the sector in question. For example, some sectors might locate in areas because of lower transport costs, while others might locate in areas with sector-specific inputs.

The heterogeneous relationship between coastal status and firm clusters found by this paper, and the heat maps showing the distribution of different types of manufacturing clusters lead to another comparison. Henderson et al. (2017) argue that because transport-related geography played a relatively bigger role in explaining the distribution of economic activity in late-urbanizing countries than in early-urbanizing countries, economic activity in these late-urbanizers tends to be clustered on coasts and the hinterlands of developing countries therefore have relatively less activity. As such, they argue that the attempts of many developing countries to follow the industrial development path of developed countries may be misplaced.

However, the results from this paper show that this is clearly not the case for the distribution of manufacturing activity Tanzania. One possible reason for this difference is that Tanzania has several important regional borders on which there are large lakes, ports, and other trading hubs, which may lead to firms clustering in those border districts rather than in coastal districts. Another explanation is that important forms of natural advantage, such as mineral resource deposits or forests, are located throughout the country and cause clusters to form around those resources rather than on the coasts. Either way, these results suggest that industrial development policies in Tanzania would not be mistaken in focusing on promoting activity in the hinterlands of the country.

These results suggest that the distribution of manufacturing clusters in Tanzania does not perfectly follow either of the trends described in Henderson et al. (2017). Unlike early-urbanizing countries, trade-related geographic variables play a dominant role in explaining Tanzania's distribution of manufacturing clusters. However, unlike late-urbanizing countries, the distribution of activity in Tanzania is not centered on the coast. These findings represent a first step in achieving a nuanced understanding of how industrial clusters can form and be sustained in Tanzania. Knowing which geographic characteristics lead to firms clustering in a district is important in aligning any cluster development schemes with natural advantages.

VII. Conclusion

This research finds that there is a significant relationship between geography and firm clustering in the Tanzanian manufacturing industry, and in most cases, geography explains between 10-30% of the variation in clusters. A cursory analysis suggests that agglomeration and firm productivity are positively correlated, and there is evidence that this relationship varies by firm characteristics. Though the sign of coefficients is consistent, the significance of this relationship between agglomeration and productivity varies greatly across different sector and sector pairings.

In general, this research consistently finds heterogeneity in the dynamics and effects of agglomeration economies. The distribution of firm clusters are quite different across different sectors, as are co-agglomeration patterns for different sector pairings, and the effect of geography on clustering varies significantly by sector and sector pairings. However, when considering just one sector, such as the furniture sector, it is possible to identify consistent trends in the role of geography in explaining clusters. This suggests that when considering the dynamics of manufacturing clusters, a nuanced approach should be taken that is cognizant of sector-specific and sector pair-specific dynamics.

The heterogeneity in the relationship between geography and clusters should motivate additional investigation in order to better understand the mechanisms underlying these results. Future research should conduct case studies for the sectors analyzed here, and work to develop a theoretical explanation for each of the relationships between geography and firm clustering identified in this paper, based on the characteristics of the sector being considered. There are also numerous ways to improve the set of natural advantage variables being considered. For example, given the importance of regional trade in Tanzania, it would be valuable for future analyses to look at the impact of distance to borders. It would also be useful to look at how the distribution of specific natural resources, such as minerals, forests, and agricultural goods might affect cluster development. In general, conducting this analysis with a set of geographic variables more tailored to the Tanzanian context, rather than those used to analyze the distribution of global activity, would be beneficial.

Furthermore, additional research is needed to test for the existence of a causal relationship between agglomeration and firm performance, and to test how the effect of agglomerations might vary across a wider range of firm characteristics. This would help to reveal reasons why some types of agglomerations economies do not benefit some types of firms. Ideally, this research would be able to identify some of the underlying mechanisms, such as knowledge spillovers, through which agglomerations benefit manufacturing firms in Tanzania.

Figures & Tables

Figure 1: Population Distribution

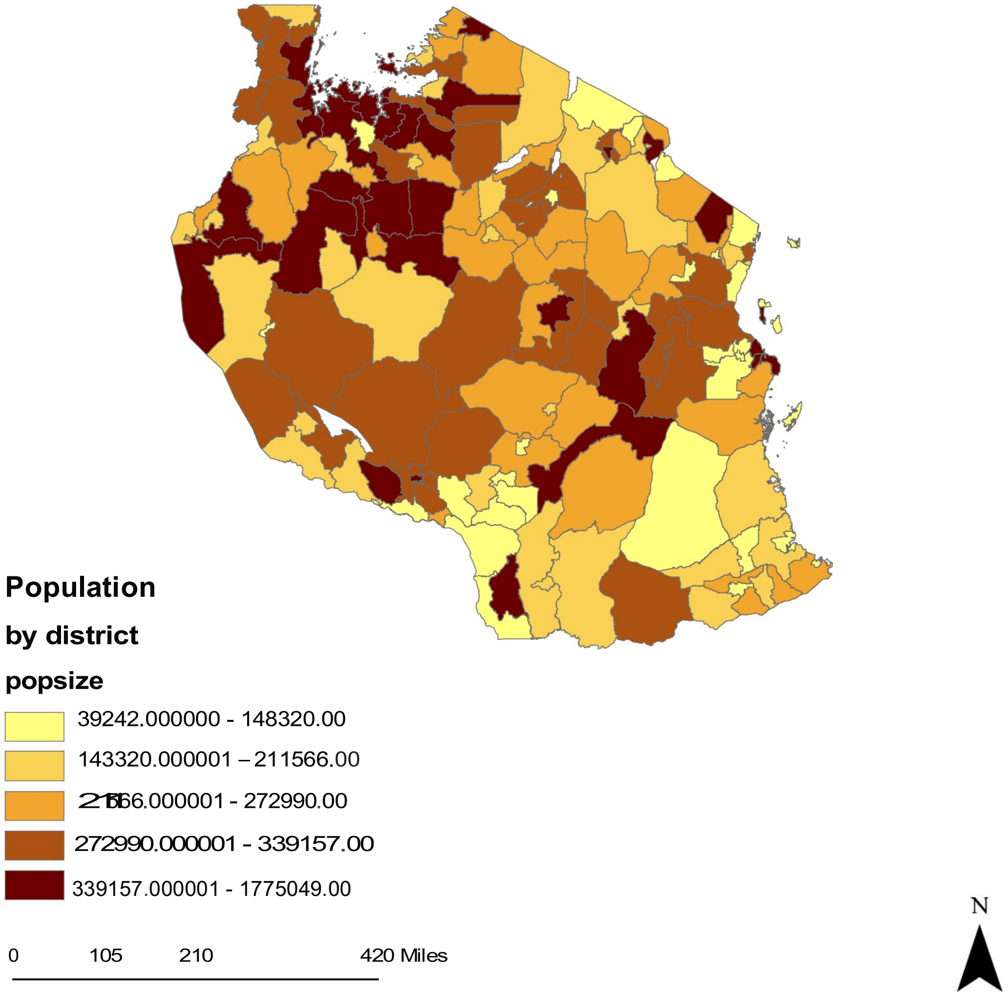

Figure 2: Overall Industrial Agglomeration

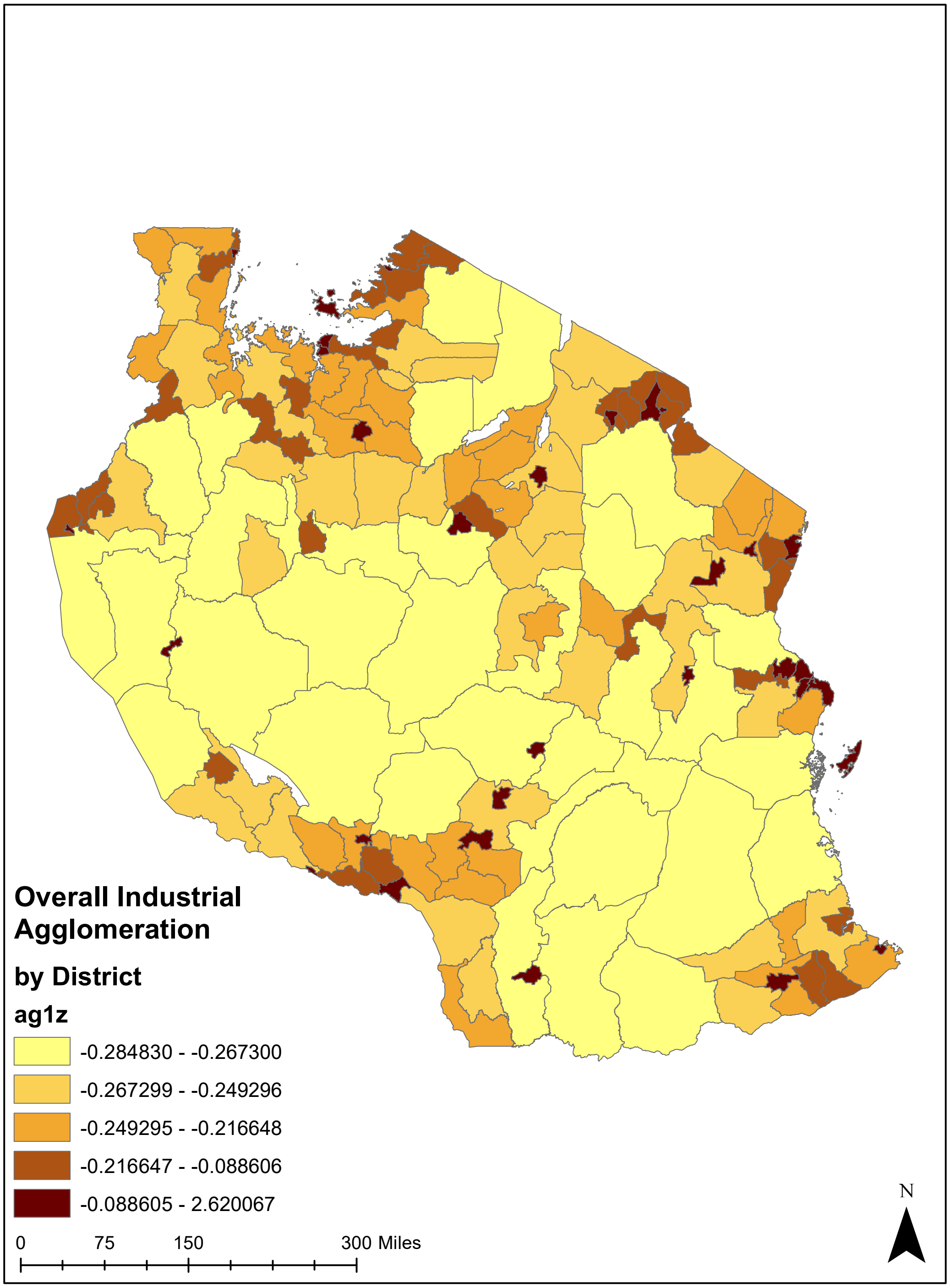

Figure 3: Food Sector Agglomeration

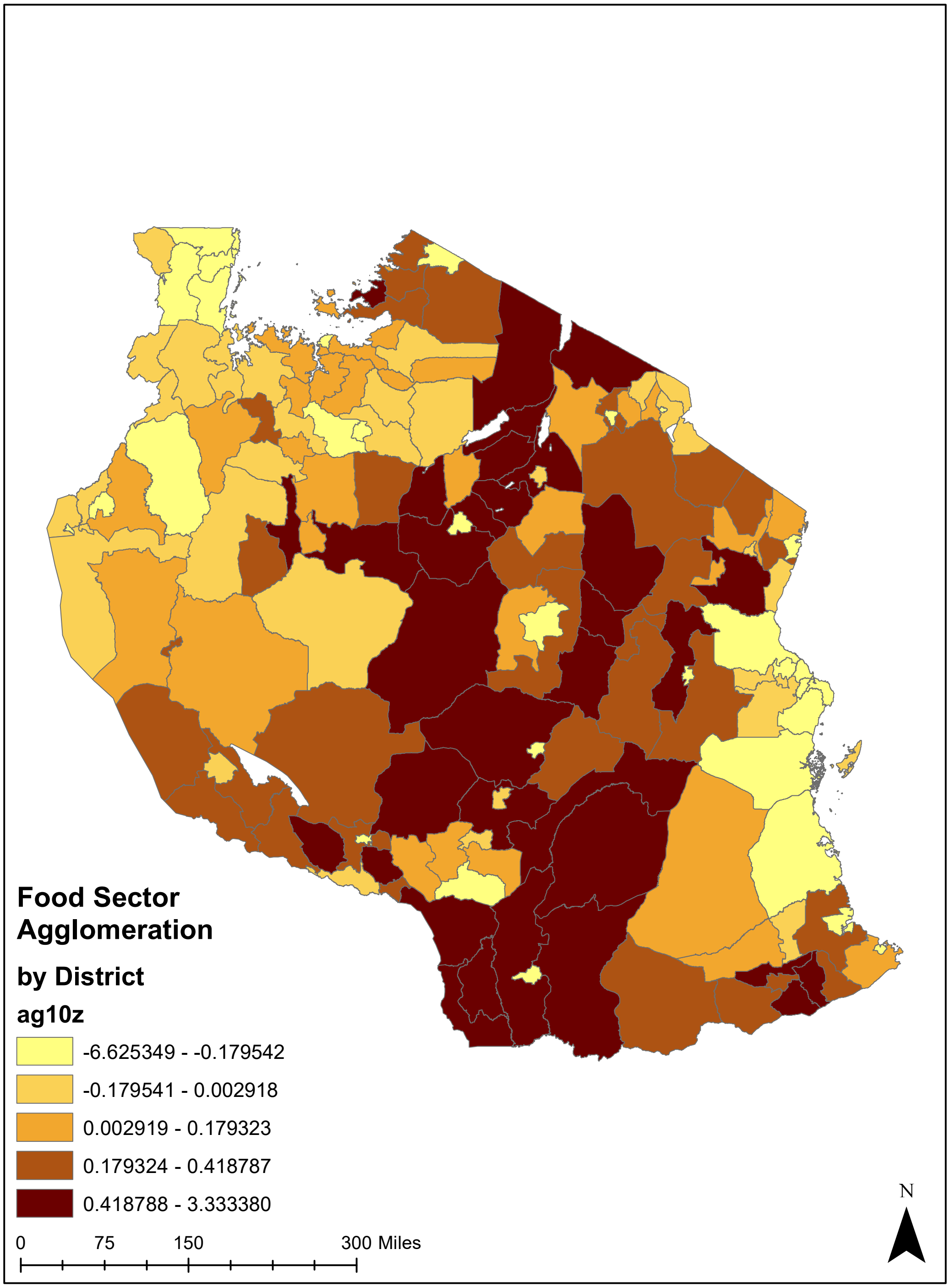

Figure 4: Apparel Sector Agglomeration

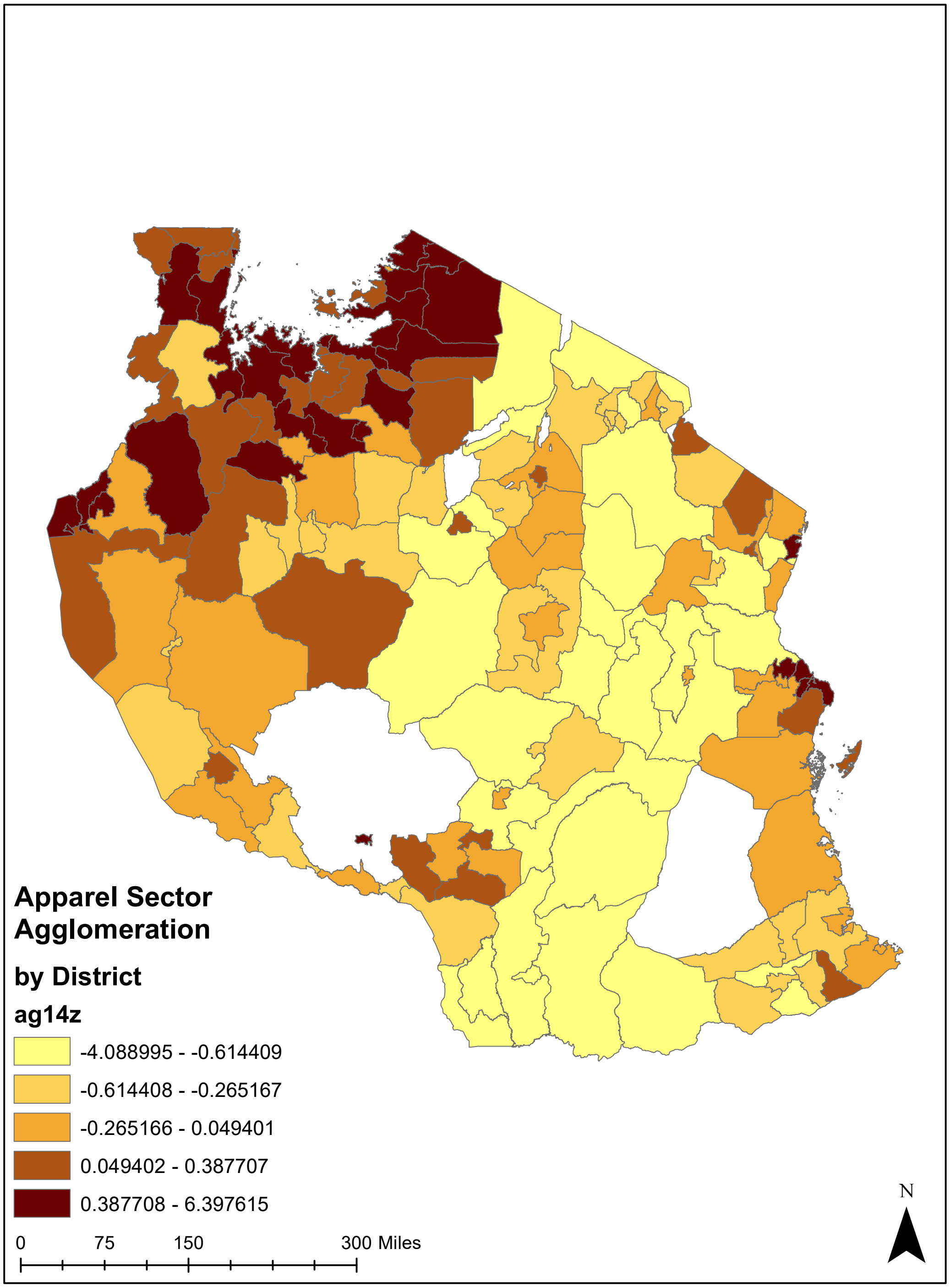

Figure 5: Wood excl. Furniture Sector Agglomeration

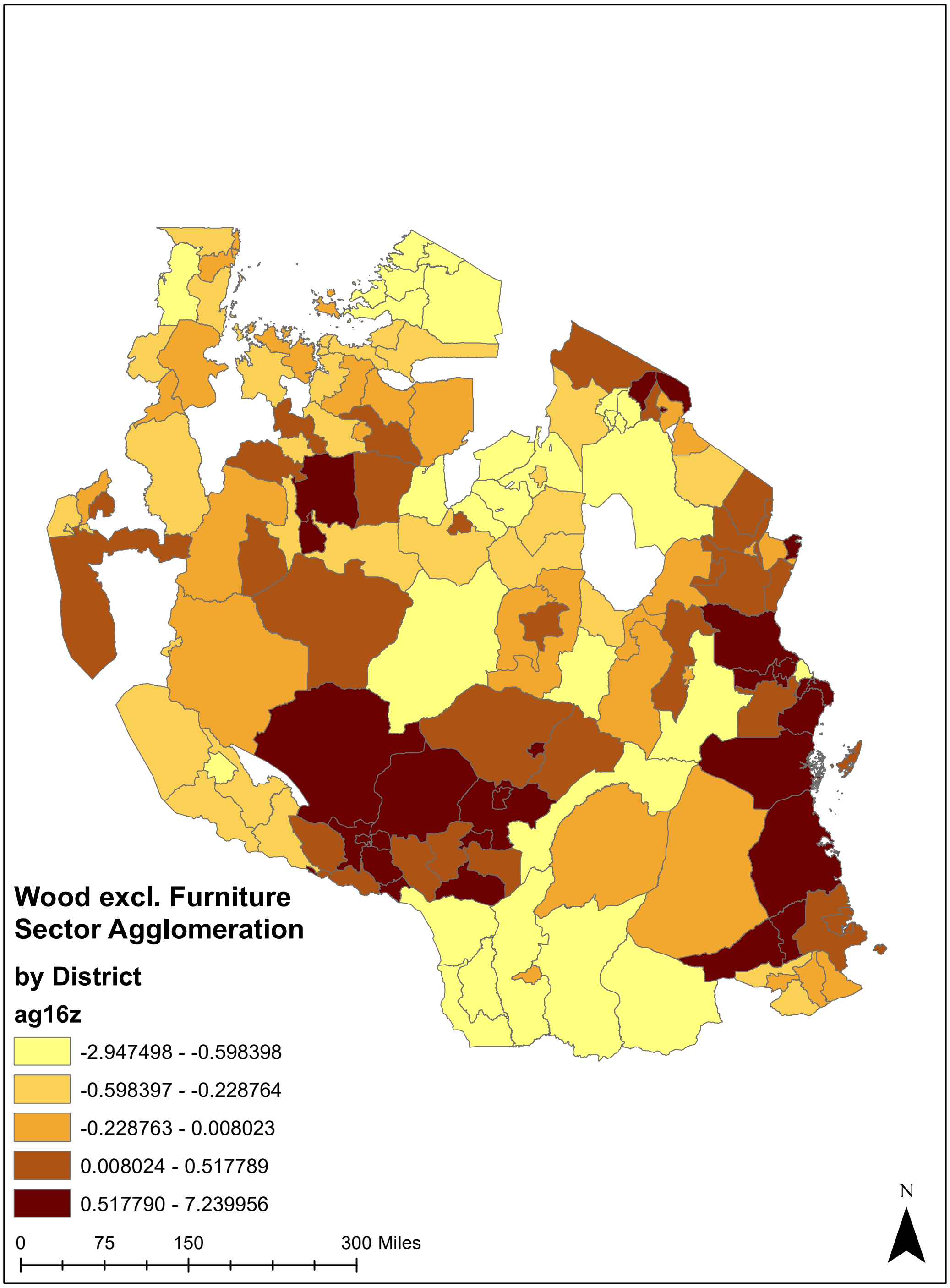

Figure 6: Fabricated Metals Sector Agglomeration

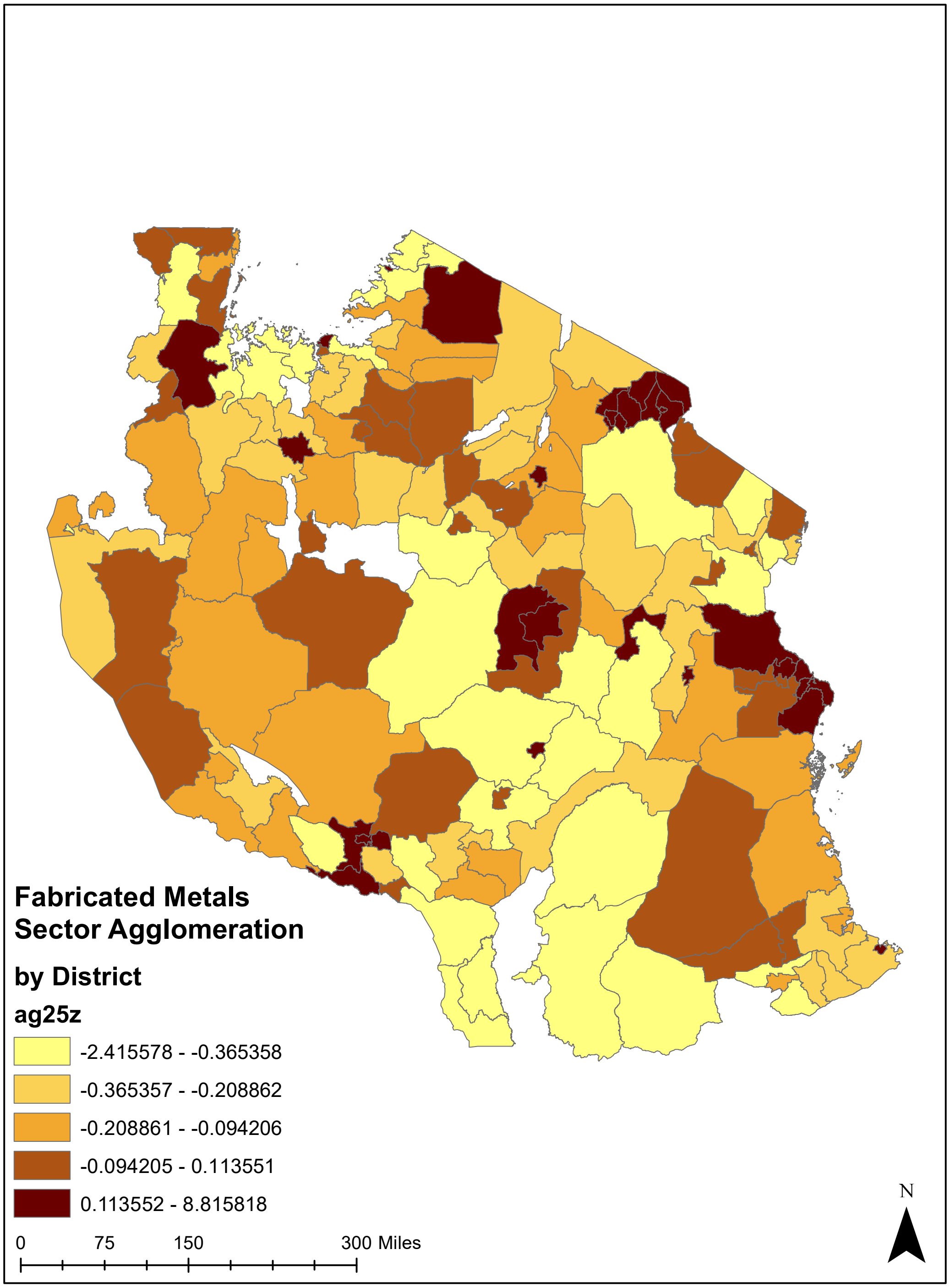

Figure 7: Furniture Sector Agglomeration

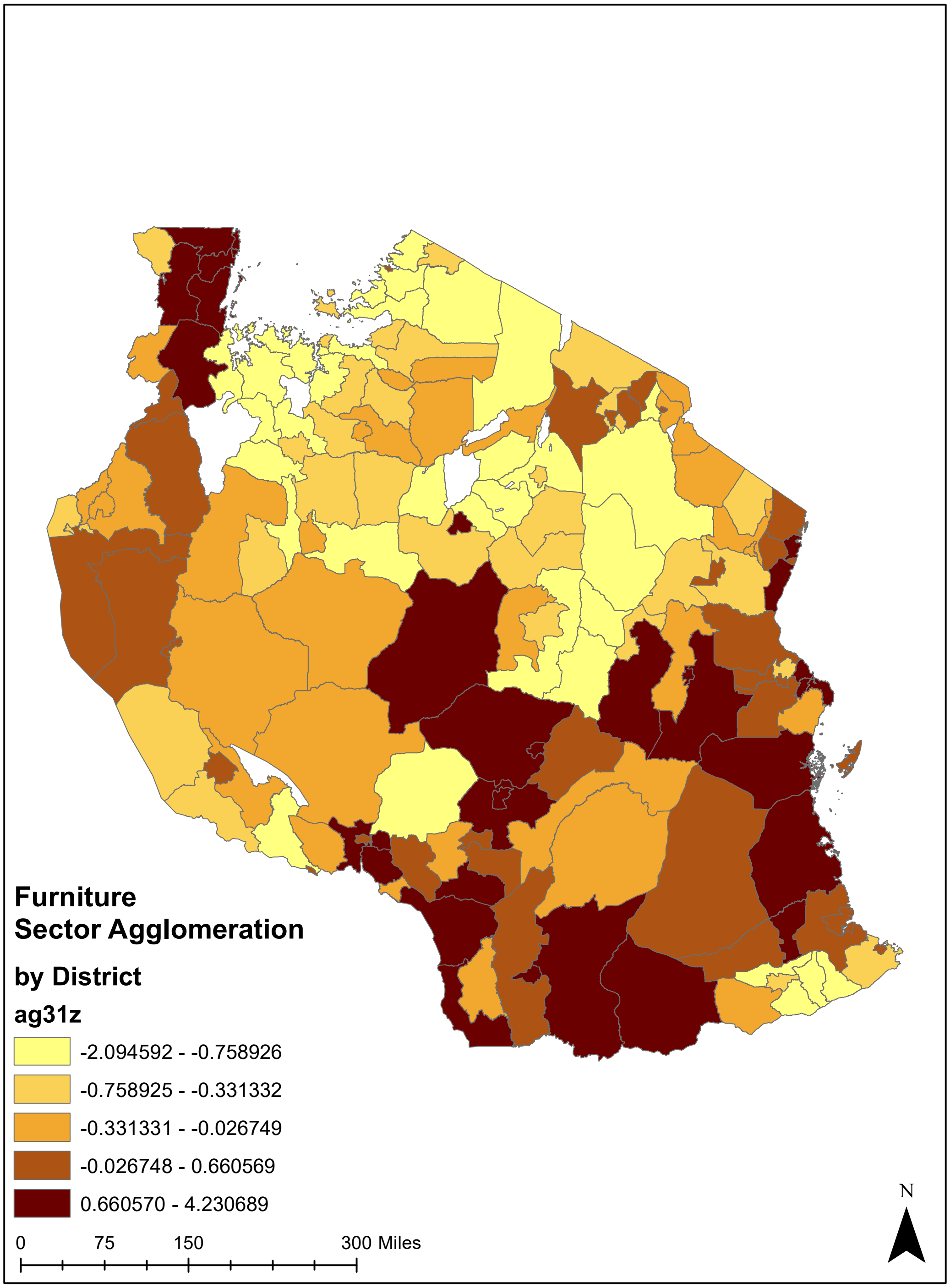

Figure 8a: Co-Agglomerations

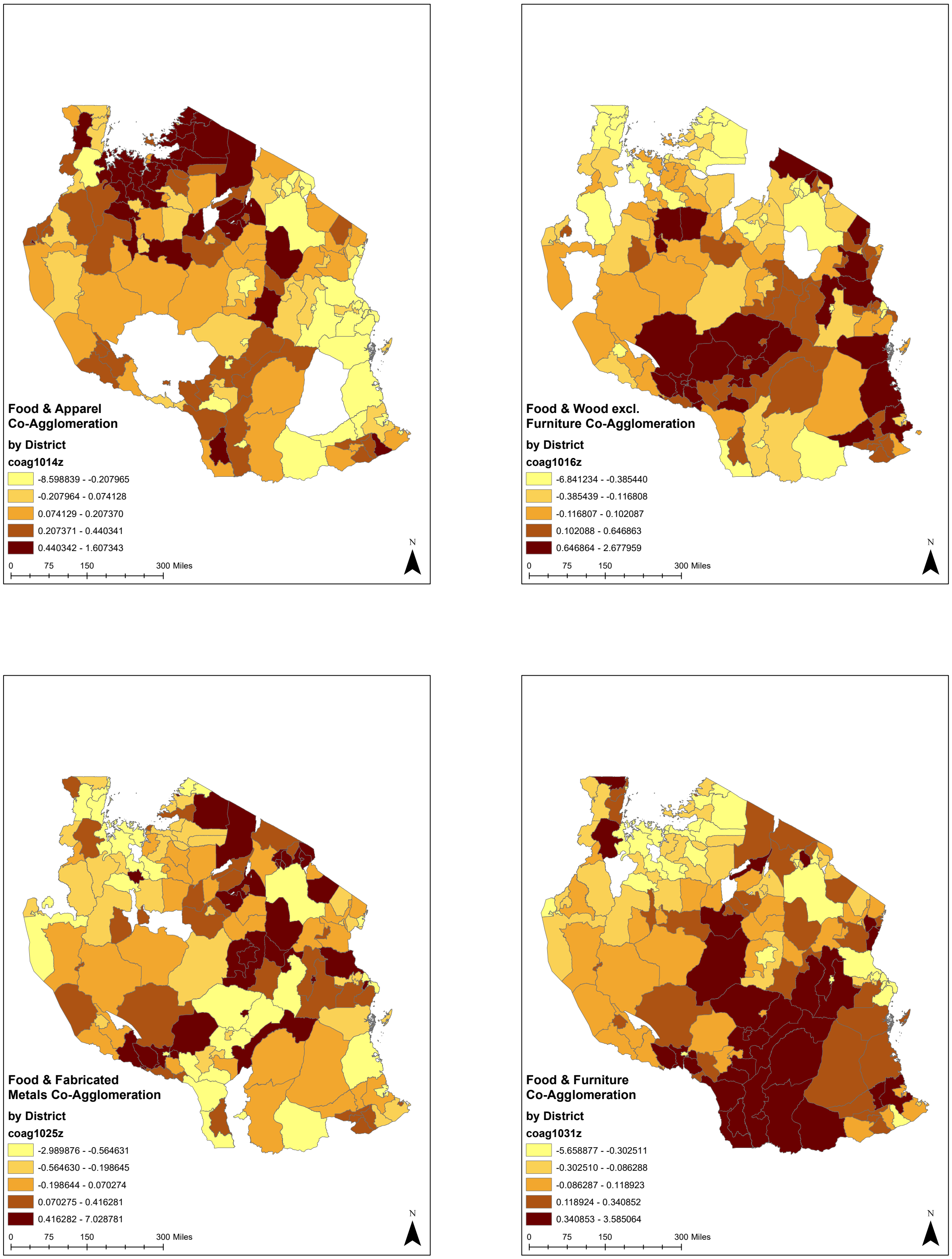

Figure 8b: Co-Agglomerations

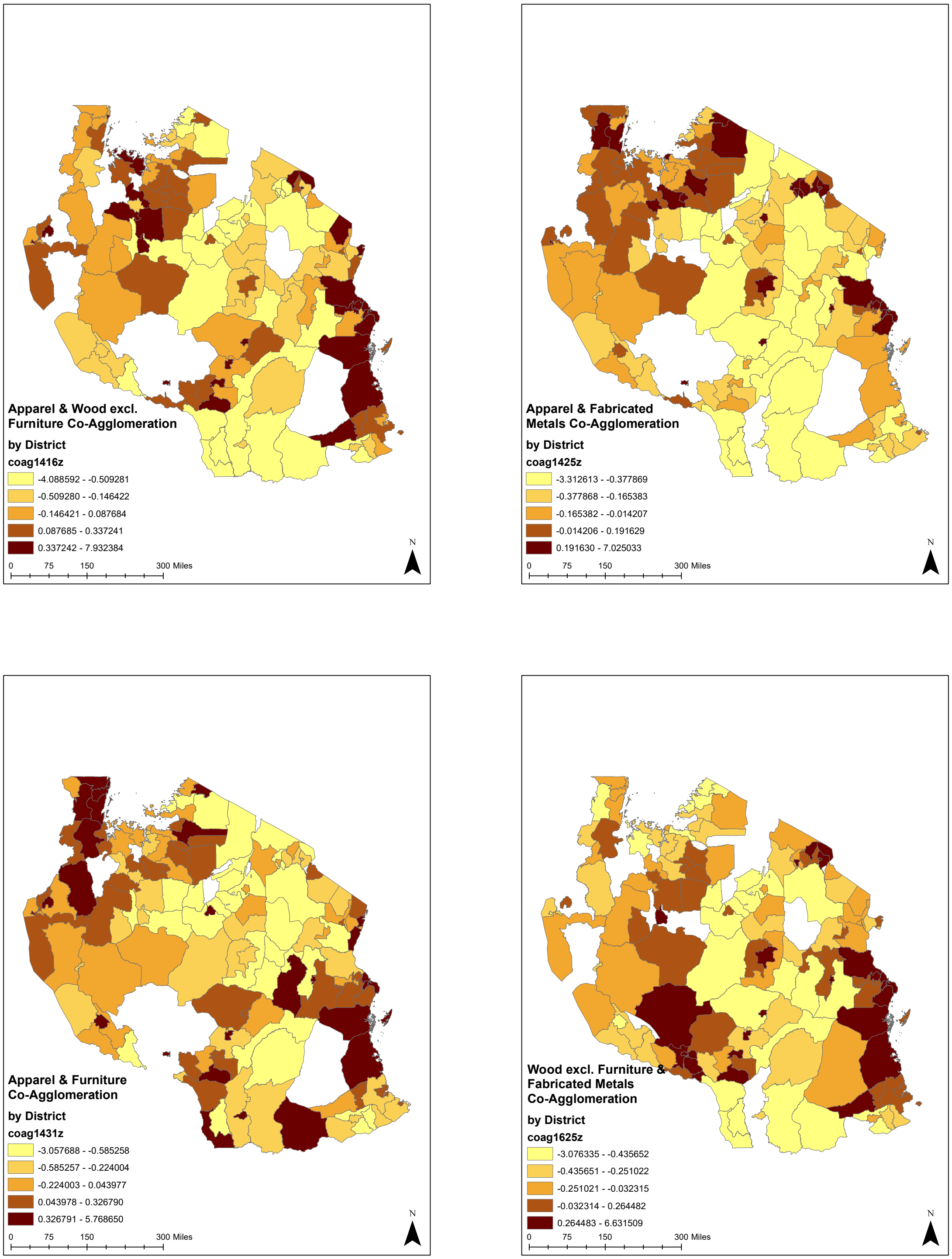

Figure 8c: Co-Agglomerations

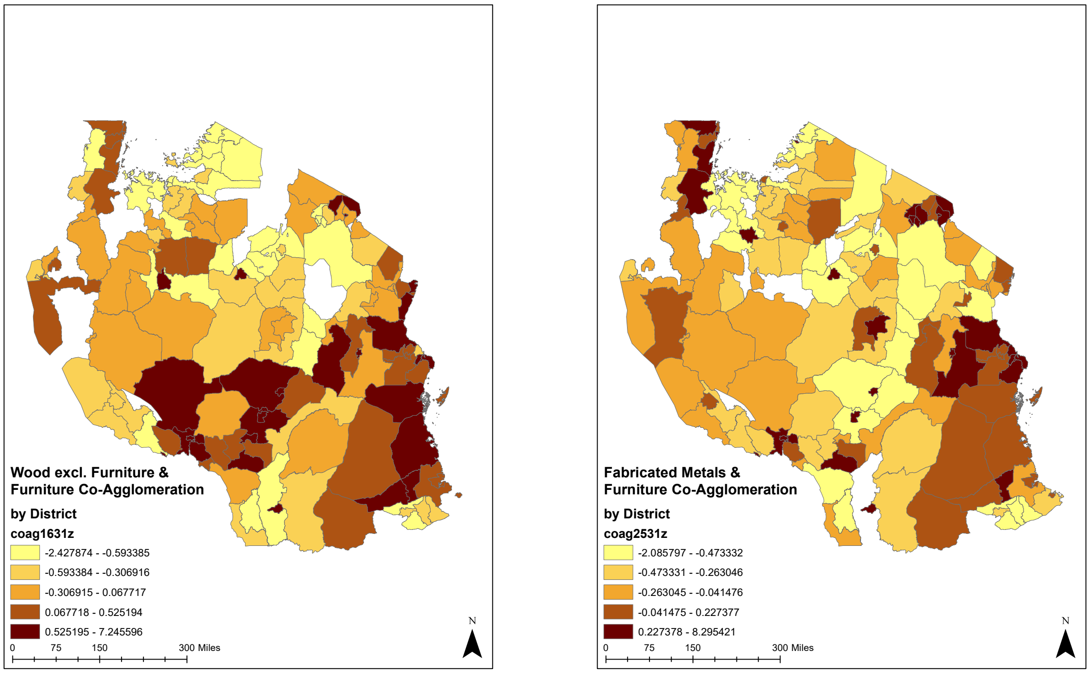

Table 1: Geographic Variables at District Level

Variable	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
District is coastal (binary)	149	0.13	0.33	0	1
<25km away from natural harbor (binary)	149	0.08	0.27	0	1
<25km away from lake (binary)	149	0.17	0.37	0	1
Monthly total precipitation (1960-90 average, cm)	149	8.55	2.39	4.69	18.25
Land suitability (probability that land is cultivated)	149	0.70	0.18	0.33	0.99
Ruggedness (000s of index)	149	1.92	2.38	0	21.20

Notes: The measures of monthly total precipitation, land suitability, and ruggedness are standardized in the regression analysis to allow for easier interpretation of coefficients, but their non-standardized summary statistics are presented here. Land suitability is defined as the probability that land is cultivated, based off of measures of climate and soil.

Table 2: Standardized Agglomeration and Co-Agglomeration Measures

Agglomeration Measures	Obs	Mean	Std. Dev.	Min	Max
Overall	149	0	1	-0.42	8.64
Food	145	0	1	-6.48	3.20
Apparel	137	0	1	-4.15	6.60
Wood excl. Furniture	130	0	1	-2.99	7.42
Fabricated Metals	138	0	1	-2.38	8.86
Furniture	141	0	1	-2.13	4.04
Food - Apparel	137	0	1	-8.56	1.25
Food - Wood excl. Furniture	130	0	1	-7.01	2.14
Food - Fabricated Metals	138	0	1	-3.25	7.66
Food - Furniture	141	0	1	-5.58	3.50
Apparel - Wood excl. Furniture	123	0	1	-4.07	8.02
Apparel - Fabricated Metals	130	0	1	-3.22	6.96
Apparel - Furniture	134	0	1	-3.03	5.87
Wood excl. Furniture - Fabricated Metals	126	0	1	-3.02	6.62
Wood excl. Furniture - Furniture	129	0	1	-2.42	7.33
Fabricated Metals - Furniture	135	0	1	-2.05	8.35

Notes: All measures of agglomeration and co-agglomeration are standardized, in order to allow for easier interpretation of regression coefficients. These summary statistics show the clustering rates for all districts included in the geographic analysis.

Table 3: Industry and Sector Agglomeration

VARIABLES	(1) Overall	(2) Food Products	(3) Apparel	(4) Wood excl. Furniture	(5) Fabricated Metals	(6) Furniture
Coastal	0.34***	-2.25***	1.34***	1.42***	1.97***	1.09***
	(0.089)	(0.29)	(0.32)	(0.32)	(0.30)	(0.28)
<25km away from harbor	-0.15	1.28***	-0.87**	-0.49	-1.25***	-0.64**
	(0.10)	(0.31)	(0.34)	(0.33)	(0.32)	(0.30)
<25km away from lake	0.17**	0.027	0.34	-0.48**	-0.20	-0.33*
	(0.066)	(0.19)	(0.21)	(0.21)	(0.20)	(0.19)
Precipitation	0.0077	-0.031	-0.018	0.17**	-0.023	0.40***
	(0.025)	(0.077)	(0.091)	(0.081)	(0.078)	(0.076)
Land suitability	-0.0096	0.087	-0.0024	-0.092	-0.059	-0.20***
	(0.025)	(0.074)	(0.085)	(0.079)	(0.077)	(0.072)
Ruggedness (000s of index)	0.12***	0.019	-0.081	0.13	0.016	-0.0086
	(0.026)	(0.076)	(0.12)	(0.085)	(0.078)	(0.073)
Constant	-0.22***	0.16*	-0.18*	-0.029	-0.100	-0.017
	(0.029)	(0.086)	(0.097)	(0.092)	(0.090)	(0.084)
Observations	149	145	137	130	138	141
Adjusted R-squared	0.160	0.317	0.113	0.217	0.257	0.288

Notes: Each column reports the estimated coefficients for equation 1 with agglomeration rates for different sectors. Robust standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 4: Shapley Decomposition for Agglomeration Regressions

	Overall	**Food Products**	**Apparel**	**Wood excl. Furniture**	**Fabricated Metals**	**Furniture**
Coastal	30.6	77.12	63.8	53.94	73.98	30.44
<25km away from harbor	4.18	11.87	15.06	5.12	14.62	3.97
<25km away from lake	11.95	0.68	9.51	17.18	3.79	4.67
Precipitation	4.96	0.36	0.22	11.91	0.36	43.01
Land Suitability	1.04	7.63	3.02	6.12	6.18	17.25
Ruggedness (000s of index)	47.27	2.32	8.39	5.73	1.08	0.66

Notes: Each column shows the percentage contribute of the individual explanatory variables to the overall R-squared value for a given measure of agglomeration.

Table 5a: Co-agglomeration

VARIABLES	(1) Food – Apparel	(2) Food – Wood excl. Furniture	(3) Food - Fabricated Metals	(4) Food - Furniture	(5) Apparel – Wood excl. Furniture
Coastal	-1.98***	-1.04***	0.39	-1.92***	1.80***
	(0.29)	(0.34)	(0.32)	(0.31)	(0.33)
<25km away from harbor	1.11***	0.86**	-0.53	1.08***	-0.79**
	(0.30)	(0.35)	(0.34)	(0.33)	(0.34)
<25km away from lake	0.39**	-0.39*	-0.32	-0.16	-0.19
	(0.19)	(0.22)	(0.21)	(0.21)	(0.21)
Precipitation	-0.093	0.12	-0.10	0.20**	0.096
	(0.081)	(0.086)	(0.083)	(0.083)	(0.089)
Land suitability	0.12	-0.0073	0.030	-0.014	-0.037
	(0.076)	(0.084)	(0.081)	(0.079)	(0.087)
Ruggedness (000s of index)	0.020	0.15*	0.065	0.017	-0.0036
	(0.11)	(0.090)	(0.082)	(0.081)	(0.14)
Constant	0.059	0.12	0.040	0.17*	-0.11
	(0.087)	(0.098)	(0.095)	(0.092)	(0.098)
Observations	137	130	138	141	123
Adjusted R-squared	0.329	0.117	0.016	0.229	0.221

Notes: Each column reports the estimated coefficients for equation 1 with co-agglomeration rates for different sector pairings. Robust standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 5b: Co-agglomeration

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Wood excl. Furniture – Fabricated Metals	Wood excl. Furniture - Furniture	Apparel – Fabricated Metals	Apparel - Furniture	Fabricated Metals - Furniture
Coastal	1.99***	1.58***	2.22***	1.62***	1.97***
	(0.31)	(0.30)	(0.30)	(0.30)	(0.29)
<25km away from harbor	- 1.28***	-0.99***	-1.14***	-0.68**	-1.25***
	(0.33)	(0.32)	(0.32)	(0.31)	(0.31)
<25km away from lake	-0.049	0.036	-0.40**	-0.51***	-0.27
	(0.21)	(0.20)	(0.19)	(0.19)	(0.20)
Precipitation	-0.018	0.21**	0.088	0.30***	0.10
	(0.087)	(0.089)	(0.076)	(0.078)	(0.078)
Land suitability	-0.046	-0.12	-0.084	-0.16**	-0.12
	(0.082)	(0.080)	(0.075)	(0.074)	(0.075)
Ruggedness (000s of index)	-0.056	-0.089	0.069	0.11	0.0078
	(0.11)	(0.11)	(0.080)	(0.080)	(0.075)
Constant	-0.14	-0.13	-0.078	-0.028	-0.086
	(0.095)	(0.093)	(0.088)	(0.087)	(0.088)
Observations	130	134	126	129	135
Adjusted R-squared	0.267	0.240	0.351	0.333	0.309

Notes: Each column reports the estimated coefficients for equation 1 with co-agglomeration rates for different sector pairings. Robust standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 6: Shapley Decomposition for Co-Agglomeration Regressions

	Food – Apparel	Food – Wood excl. Furniture	Food - Fabricated Metals	Food - Furniture	Apparel – Wood excl. Furniture
Coastal	67.83	32.59	10.56	70.35	78.21
<25km away from harbor	9.33	16.68	23.12	12.27	7.88
<25km away from lake	8.15	14.19	34.42	0.68	3.87
Precipitation	1.67	11.23	22.35	11.4	2.59
Land Suitability	10.12	2.47	1.1	2.05	4.44
Ruggedness (000s of index)	2.91	22.83	8.45	3.24	3.02
	Wood excl. Furniture – Fabricated Metals	**Wood excl. Furniture - Furniture**	**Apparel – Fabricated Metals**	**Apparel - Furniture**	**Fabricated Metals - Furniture**
Coastal	73.4	58.32	73.64	48.8	68.77
<25km away from harbor	15.14	9.92	8.85	4.9	12.23
<25km away from lake	0.99	0.4	8.52	12.15	4.64
Precipitation	0.25	12.55	1.57	21.09	2.44
Land Suitability	5.78	11.97	6.47	10.25	10.92
Ruggedness (000s of index)	4.45	6.83	0.95	2.8	1.01

Notes: Each column shows the percentage contribute of the individual explanatory variables to the overall R-squared value for a given measure of co-agglomeration.

Table 7: Industry and Sector Agglomeration of 10+ Employees

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Overall	Food Products	Wood excl. Furniture	Fabricated Metals	Furniture
Coastal	0.75***	-1.86***	-1.43**	0.35	0.73
	(0.21)	(0.34)	(0.52)	(2.02)	(0.87)
<25km away from harbor	-0.29	1.21***	-0.40	-2.38	-1.18
	(0.24)	(0.41)	(0.47)	(2.37)	(1.49)
<25km away from lake	0.40**	0.012	-0.49	-0.49	-0.18
	(0.17)	(0.23)	(0.78)	(1.60)	(1.05)
Precipitation	0.014	-0.12	0.31	0.37	0.027
	(0.064)	(0.088)	(0.20)	(1.23)	(0.30)
Land suitability	-0.022	-0.013	-0.10	-0.29	-0.0062
	(0.065)	(0.091)	(0.22)	(0.61)	(0.29)
Ruggedness (000s of index)	0.27***	0.074	-0.064	-0.16	-0.056
	(0.067)	(0.15)	(0.40)	(0.46)	(0.48)
Constant	-0.33***	0.051	0.26	0.13	-0.055
	(0.076)	(0.098)	(0.22)	(1.12)	(0.39)
Observations	113	57	18	12	19
Adjusted R-squared	0.170	0.414	0.296	-0.364	-0.336

Notes: Each column reports the estimated coefficients for equation 1 with agglomeration rates defined by clusters of firms with ten or more employees. Robust standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 8: Shapley Decomposition for Agglomeration Regressions, 10+ Employees

	Overall	Food Products	Wood excl. Furniture	Fabricated Metals	Furniture
Coastal	29.99	70.99	60.85	28.09	52.96
<25km away from harbor	3.63	9.37	3.19	33.94	29.19
<25km away from lake	13.13	0.9	2.32	5.3	4.9
Precipitation	4.5	4.6	14.11	7.52	1.47
Land Suitability	1.75	7.55	7.59	16.08	4.69
Ruggedness (000s of index)	47	6.59	11.94	9.06	6.79

Notes: Each column shows the percentage contribute of the individual explanatory variables to the overall R-squared value for a given measure of agglomeration.

Table 9: Relationship between Agglomeration & Value-added per Worker

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Overall	Food Products	Apparel	Wood excl. Furniture	Fabricated Metals	Furniture
Agglomeration	3,884	2,592***	-13,077	928	2,290***	3,884
	(3,114)	(302)	(14,050)	(1,085)	(714)	(3,114)
Micro	-24,130**	-2,389**	-51,811	-6,712**	-5,591***	-24,130**
	(9,269)	(1,104)	(47,541)	(3,269)	(1,652)	(9,269)
Micro * Agg	-4,721	-1,850***	13,698	-139	-1,579*	-4,721
	(3,113)	(314)	(14,152)	(1,128)	(933)	(3,113)
License	1,663***	593	2,210**	745***	1,472***	1,663***
	(485)	(361)	(1,011)	(282)	(349)	(485)
License * Agg	-75.2	-564***	-766**	-967***	117	-75.2
	(177)	(179)	(308)	(110)	(292)	(177)
Constant	26,886***	4,430***	54,008	9,380***	7,959***	26,886***
	(9,253)	(1,074)	(47,200)	(3,231)	(1,624)	(9,253)
Observations	2,502	1,566	1,200	1,329	1,711	2,502
Adjusted R-squared	0.045	0.134	0.067	0.097	0.073	0.045

Notes: Each column reports the estimated coefficients for equation 2; the agglomeration variable and interaction variables across each column are calculated with the agglomeration rate for the sector identified by the column heading. Robust standard errors, clustered at the district level, are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 10a: Relationship between Co-Agglomeration & Value-added per Worker

VARIABLES	(1) Food - Apparel	(2) Food – Wood excl. Furniture	(3) Food – Fabricated Metals	(4) Food - Furniture	(5) Apparel – Wood excl. Furniture
Co-agglomeration	3,409	5,032	-975	3,801	-14,697
	(3,569)	(4,304)	(2,018)	(2,975)	(15,285)
Micro	-25,044**	-28,232**	-20,082***	-22,275***	-50,537
	(9,902)	(12,094)	(7,018)	(8,408)	(45,565)
Micro * Co-agg	-3,954	-5,760	1,590	-4,556	15,373
	(3,654)	(4,324)	(2,068)	(2,915)	(15,313)
License	1,698***	1,932***	1,566***	1,707***	1,158***
	(373)	(473)	(449)	(427)	(441)
License * Co-Agg	12.5	-310	-277	-173	-389***
	(312)	(230)	(260)	(178)	(145)
Constant	27,398***	30,669**	22,667***	24,951***	52,794
	(9,884)	(12,065)	(6,992)	(8,395)	(45,502)
Observations	3,994	3,549	3,793	4,191	2,619
Adjusted R-squared	0.048	0.048	0.040	0.046	0.073

Notes: Each column reports the estimated coefficients for equation 2; the co-agglomeration variable and interaction variables across each column are calculated with the co-agglomeration rate for the sector pairing identified by the column heading. Robust standard errors, clustered at the district level, are in parentheses.
* p<0.1, ** p<0.05, *** p<0.01.

Table 10b: Relationship between Co-Agglomeration & Value-added per Worker

VARIABLES	(1) Wood excl. Furniture – Fabricated Metals	(2) Wood excl. Furniture - Furniture	(3) Apparel – Fabricated Metals	(4) Apparel - Furniture	(5) Fabricated Metals - Furniture
Co-agglomeration	1,578	2,037***	-12,820	-6,372	1,213
	(1,284)	(603)	(13,907)	(7,988)	(1,349)
Micro	-4,827***	-5,313***	-44,056	-35,349	-6,451***
	(1,726)	(1,551)	(38,482)	(30,945)	(1,791)
Micro * Co-agg	-789	-1,388**	13,556	6,772	-329
	(1,218)	(630)	(13,973)	(8,015)	(1,486)
License	656**	894***	1,456**	1,864***	1,080***
	(261)	(272)	(702)	(639)	(274)
License * Co-Agg	-583***	-187	-650***	-44.5	-349**
	(133)	(272)	(212)	(211)	(143)
Constant	6,955***	7,525***	46,577	37,908	8,926***
	(1,724)	(1,497)	(38,311)	(30,811)	(1,746)
Observations	2,824	3,203	2,477	2,853	3,006
Adjusted R-squared	0.136	0.088	0.063	0.050	0.111

Notes: Each column reports the estimated coefficients for equation 2; the co-agglomeration variable and interaction variables across each column are calculated with the co-agglomeration rate for the sector pairing identified by the column heading. Robust standard errors, clustered at the district level, are in parentheses.
* p<0.1, ** p<0.05, *** p<0.01.

References

Abebe, G., M. McMillan., and M. Serafinelli (2018). 'Foreign Direct Investment and Knowledge Diffusion in Poor Locations: Evidence from Ethiopia'. NBER Working Paper Series, 24461.

Callois, J. M. (2008). 'The two sides of proximity in industrial clusters: The trade-off between process and product innovation'. Journal of Urban Economics, 63(1):146{162.

Chhair, S. and C. Newman (2014). 'Clustering, Competition, and Spillover Effects: Evidence from Cambodia'. WIDER Working Paper 2014/065. Helsinki: UNU-WIDER.

Combes, P. P., G. Duranton, L. Gobillon, D. Puga, and S. Roux (2012). 'The Productivity Advantages of Large Cities: Distinguishing Agglomeration From Firm Selection'. Econometrica, 80(6):2543{2594.

Dinh, H. T., and C. Monga (2013). 'Light Manufacturing in Tanzania: A Reform Agenda for Job Creation and Prosperity. Directions in Development--Private Sector Development'. Washington, DC: World Bank.

Ellison, G., and E. L. Glaeser (1997). 'Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach'. Journal of Political Economy, October 1997, 105(5), pp. 889 927.

Ellison, G., E. L. Glaeser. (1999). 'The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?'. American Economic Review, vol. 89, no. 2, pp. 311–316.

Henderson, V., T. Squires, A. Storeygard, and D. Weil (2017). 'The global distribution of economic activity: nature, history, and the role of trade'. Quarterly Journal of Economics 133, 357–406.

Howard, E., C. Newman, J. Rand, and F. Tarp (2014). 'Productivity-enhancing manufacturing clusters – evidence from Vietnam'. UNU-Wider Working Paper N. 2014/71.

Howard, E., C. Newman, and F. Tarp (2016). ‘Measuring industry coagglomeration and identifying the driving forces’. Journal of Economic Geography, 16(5), 1055-1078.

Kim, S. (1999). ‘Regions, resources and economics geography: Sources of U.S. regional comparative advantage, 1880-1987’. Regional Science and Urban Economics, 29: 1-32.

Kinyondo, A., C. Newman, and F. Tarp (2016). ‘The Role and Effectiveness of Special Economic Zones in Tanzania’. WIDER Working paper 2016/122. Helsinki: UNU-WIDER.

Lu, J., & Z. Tao (2009). ‘Trends and determinants of China's industrial agglomeration’. Journal of Urban Economics, 65(2), 167-180.

Marshall, A. (1920). ‘Principles of Economics’. Macmillan, London.

Ministry of Industry, Trade and Marketing (MITM) (2010). ‘Integrated Industrial Development Strategy 2025’. United Republic of Tanzania: Dar es Salaam.

Newman C. and J. Page (2017). ‘Industrial clusters: The case for Special Economic Zones in Africa’. WIDER Working Paper Series 015. World Institute for Development Economic Research (UNU-WIDER).

Page, J. M. (2016). ‘Industry in Tanzania: Performance, prospects, and public policy’. WIDER Working Paper No. 2016/5. Helsinki: UNU-WIDER.

Rodrik, D. (2014). ‘An African Growth Miracle?’. NBER Working Paper 2018. Cambridge, MA: NBER.

Rosenthal S., and W. C. Strange (2001). ‘The Determinants of Agglomeration’. Journal of Urban Economics, 50(2), 191-229.

Siba, E. et al. (2012). ‘Enterprise Agglomeration, Output Prices, and Physical Productivity: Firm-Level Evidence from Ethiopia’. WIDER Working Paper 2012/085. Helsinki: UNU-WIDER.

UNIDO (2009). 'Industrial Development Report 2009: Breaking In and Moving Up - New Industrial Challenges for the Bottom Billion and the Middle-Income Countries'. United Nations Publications.

Yoshino, Y. (2010). 'Industrial clusters and micro and small enterprises in Africa: from survival to growth'. Washington, DC: World Bank.