Determining the Differences that Matter: Development and Divergence in US States over 1850-2010

Faculty Research Working Paper Series

Penny Mealy
Oxford University

J. Doyne Farmer
Oxford University

Ricardo Hausmann
Harvard Kennedy School

September 2018

RWP18-030

Determining the differences that matter: Development and divergence in US states over 1850-2010*

Penny Mealy[†] J. Doyne Farmer[‡] Ricardo Hausmann[§]

May 18, 2018

Abstract

Understanding the differences between rich and poor places is complicated by the fact that places differ from each other in numerous ways. In this paper, we show how a dimension reduction algorithm can unveil hidden patterns in US census data and consistently yield useful insights into the type of economic activities that separate rich and poor states over 160 years of development history. Moreover, we find this approach has a unique ability to shed light on the dynamics of evolving landscapes and changes in relevance of particular types of activities, such as the shift from manufacturing to high skill services that occurred in the US over the last 40 years. Our results have important implications for the decline of the rustbelt and the reversal of US regional income convergence from 1980 onwards.

*This project was supported by Partners for the New Economy and the Oxford Martin School project on the Post-Carbon Transition. We are also grateful Frank Neffke, Ljubica Nedelkoska, the CID Growth Lab Research Group and the INET Complexity Research Group for their valuable discussions and feedback.

†Institute for New Economic Thinking at the Oxford Martin School, Smith School for Enterprise and the Environment, and St Edmund Hall. E-mail: penny.mealy@inet.ox.ac.uk.

‡Institute for New Economic Thinking at the Oxford Martin School, Mathematical Institute, and Nuffield College.

§Center for International Development, Harvard Kennedy School of Government

1 Introduction

How can we better understand the differences between rich and poor places? A key challenge for economists and economic geographers alike is that places differ to each other in many ways. Take Massachusetts and Mississippi: One state has a strong employment concentration in high tech and financial services, while the other is more concentrated in agricultural activities. But they also differ in terms of their natural resources, military facilities, manufacturing industries and concentrations of retail outlets and religious organisations. How can we know whether some of these differences matter more for economic performance than others? And what if we wanted to compare more than two places? How might we analyse the differences between Massachusetts, Mississippi and Michigan – or the entire set of US states?

One way to overcome the high dimensionality of making such comparisons is to ignore the *type* of industries or occupations a place is concentrated in, and just examine the extent to which a place is more or less diversified. And indeed, the topic of specialisation vs diversification has received significant attention within the economic geography literature, with mixed findings in terms of the relationship with economic performance (Attaran, 1986; Quigley, 1998; Duranton and Puga, 2000; Kemeny and Storper, 2015). More recently, scholars have suggested that when it comes to regional development, the 'what' might be more important than the 'scale' of specialisation (Kemeny and Storper, 2015; Martin et al., 2018).

In this paper, we demonstrate an alternative methodology for analysing differences in employment concentrations across places. We draw on a dimension reduction algorithm that has been previously applied to country export data to calculate the country-based Economic Complexity Index (ECI) and product-based Product Complexity Index (PCI). This approach, which proved to be successful in explaining cross-country differences in per capita GDP and economic growth was originally advanced on the intuition that more prosperous places would be likely to export a large number of products that few other countries would be able to export (Hidalgo and Hausmann, 2009). However, the ECI was later shown to be mathematically unrelated to the number ('diversity') of products a country can export competitively (Kemp-Benedict, 2014). Instead, the value of the ECI and PCI metrics was suggested to lie in their ability to reduce the high-dimensional space of country-export specialisations into two single-dimension vectors that best describe the similarities in the type of products that countries export (Mealy et al., 2017).

As we discuss further in Section 2, the ECI provides an ordering that places countries with similar exports close together in the ranking, and those with dis-similar exports far apart. Analogous to the Dewey Decimal system for placing books with similar topics close together in library shelves, the ECI provides a useful way to arrange countries in terms of the similarity in *what* they export. The corresponding PCI index helps identify the type of exports that countries at either end of the ECI spectrum have in common. That is, it indicates what type of exports distinguish countries with high ECI scores from countries with low ECI scores.

This paper investigates this new interpretation of the ECI and PCI metrics in the context of regional economic development. We apply the Hausmann et al. (2014) algorithm to US census data on industrial and occupational employment concentrations across US states over 160 years of development history. For all periods in which we have state income data (1880-2010), we find that this approach consistently provides useful insights into the types of industries and occupations that underpin places of higher (and lower) economic prosperity. Moreover, in congruence with results for country-export data (Hausmann et al., 2014), we find these measures to be significantly predictive of future economic growth.

Applying this approach to census data over such a long time period also yields a further interesting discovery not yet documented in applications to country-export data: the ECI and PCI metrics appear to have a unique ability to capture the dynamics of evolving economic landscapes. Over the 1970-1980 period, the PCI of manufacturing activities declines substantially, while high skill services show a marked rise. It is important to stress that the dimension reduction algorithm takes in no information about types of industries or occupations, income or wages. It only operates on the pattern of employment concentrations across US states. However, this pattern turns out to be surprisingly informative of the changing relevance of particular *types* of specialisations in a given economic context.

We show that this shift influences states' ECI trajectories in different ways. Although rust-belt states and eastern states had very similar ECI scores leading up to 1980 (suggesting they were specialised in similar types of activities), in 1980 the ECI of these two groups diverge – signalling very different future economic prospects. We find this new divergence between these groups of states is well explained by their differences in education, which were much higher in east coast states than in rust belt states. While educational differences did not strongly distinguish between states at either end of the ECI spectrum in previous periods, these differences have become particularly important for economic success in the new knowledge economy (Glaeser et al., 2014).

Finally, we suggest that the dimension reduction tools described in this paper could provide a new, complementary analytical angle onto the reversal in the pattern of state income convergence that occurred from 1980 onwards. While this 'great inversion' has been studied from numerous standpoints (Storper, 2018), we are not aware of literature that has explored income convergence through the lens of these dimensions reduction tools. Complementing the Krugman Index (or dissimilarity index) (Krugman, 1991, 1993), which has previously been applied to investigate regional income convergence (Kim, 1998; Martin et al., 2018), we find the changes in the PCI and ECI over this period to be particularly informative of a new spatial divide separating US regions. Not only do we observe a bifurication in the PCI of manufacturing and high skill services over this period (Moretti, 2012), we also observe the error bars on these mean values becoming much further apart, suggesting that these activities have become much more strongly spatially segregated. While investigating the drivers of regional income convergence is beyond the scope of this paper, these preliminary findings align with existing theories linking slowing regional convergence to the spatial segregation of skills, which have become increasingly concentrated in cities (Glaeser and Saiz, 2003; Berry and Glaeser, 2005; Moretti, 2012).

The remainder of this paper is organised as follows. Section 2 reviews the ECI and PCI metrics and explains their use as dimension reduction tools. Section 3 describes the data and approach for applying the ECI and PCI metrics to state employment concentrations. Section 4, 5, 6 detail our results, and Section 7 concludes.

2 Understanding the ECI and PCI metrics as dimension reduction tools

The ECI and PCI metrics were originally introduced to infer information about cross-country differences in productive capabilities or 'know-how' from export data (Hidalgo and Hausmann, 2009; Hausmann et al., 2014). Their calculation involves applying an algorithm to a binary country-product matrix M, with elements M_{cp} indexed by country c and product p. $M_{cp} = 1$ if country c has a *'revealed comparative advantage'* (RCA) > 1 in product p and 0 otherwise. RCA is calculated using the Balassa (1965) index

$$RCA_{cp} = \frac{x_{cp}/\sum_p x_{cp}}{\sum_c x_{cp}/\sum_c \sum_p x_{cp}}, \tag{1}$$

where x_{cp} is country c's exports of product p.

Note that this M matrix is a $c \times p$ dimensional object, which specifies each of the p export specialisations for every country c.

We can calculate how many products a country has RCA in (its *diversity*) by summing across the rows of the M matrix (denoted d_c). Similarly, we can count how many countries have RCA in a given product (its *ubiquity*) by summing aross the columns of the M matrix (denoted u_p). That is,

$$d_c = \sum_p M_{cp} \tag{2}$$

and

$$u_p = \sum_c M_{cp}. \tag{3}$$

While the diversity and ubiquity measures provide one way to compare the number of export specialisations across countries, they don't tell us anything about the *type* of exports countries are specialised in. However, the ECI and PCI metrics can be particularly helpful in this regard.

In Mealy et al. (2017), it was shown that the ECI vector provides a one-dimensional ranking that places countries with similar types of exports close to each other in the ordering and countries with dis-similar types of exports far apart. While this interpretation is explained in detail in Mealy et al. (2017), we provide a brief overview here. The ECI is defined as the eigenvector associated with the second largest eigenvalue of the matrix

$$\widetilde{M} = D^{-1}S, \tag{4}$$

where D is the diagonal matrix formed from the diversity vector, and S is a matrix whose rows and columns correspond to countries and whose entries are given by

$$S_{cc'} = \sum_p \frac{M_{cp}M_{c'p}}{u_p}. \tag{5}$$

One can think of S as a symmetric similarity matrix, corresponding to how similar two countries' exports baskets are. Each element $S_{cc'}$ sums up the scarcity (one divided by ubiquity) of each product that countries c and c' both have RCA in.

Drawing on the work of Shi and Malik (2000), it can be shown that the ECI vector exactly minimises

$$\frac{\sum_{cc'} (y_c - y_{c'})^2 S_{cc'}}{\sum_c y_{c'}^2 d_c}, \tag{6}$$

subject to the constraint

$$\sum_c y_c d_c = 0, \tag{7}$$

where y_c and $y_{c'}$ represent the ECI scores of country c and c' respectively, and d_c denotes the diversity of country c. That is, the ECI vector is the exact solution to the problem of assigning real numbers y_c to each country c to minimise the sum of the squared the distances between countries, where distances are weighted by the similarity matrix S. The constraint $\sum_c y_c d_c = 0$ forces country ECI scores to take on positive and negative values, and to have a fairly balanced distribution above and below zero.[1] From this perspective, we can see that countries with high ECI scores have similar types of exports to countries with other high ECI scores, and dis-similar types of exports to countries with low ECI scores.

The associated PCI measure, which is symmetrically defined as the eigenvector associated with the second largest eigenvalue of the transpose of the $\widetilde{M}$ matrix, helps us identify the types of products that countries at one end of the ECI spectrum have *in common*. Products with high (low) PCI scores are more likely to be exported by countries with high (low) ECI scores, and less likely to be exported by countries with low (high) ECI scores.

In Hidalgo and Hausmann (2009) and Hausmann et al. (2014), it was shown that countries with high ECI scores tend to be advanced countries that are more likely to export technologically sophisticated ('complex') products, while countries with low ECI scores tend to be less developed countries that tend to export less technologically sophisticated products. As such, the application of the ECI and PCI measures to country-trade data offers strong empirical support for developmental theories based on 'technological capabilities' (Lall, 1992, 2000; Lall et al., 2006; Sutton and Trefler, 2016), which contend that countries possessing more advanced technological capabilities are more likely to grow and prosper.

3 Application to US States: 1850-2010

Can the interpretation of ECI and PCI as dimension reduction tools give us new insights to regional development in the US? We explore this question by drawing on US census data over the period 1850-2010 from the Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2017).

IPUMS provides decadal data on state employment in 3-digit industries and occupations. While regular changes in industry and occupation classification schemes have historically complicated longitudinal analysis, the IPUMS project has recoded census data into several sets of 'standardised' occupation classification schemes.

[1] It also forces the ECI vector to be exactly orthogonal to the country diversity vector.

We draw on the IND1950 and OCC1950 codes (based on the US Census 1950 classification scheme) to analyse the 1850-1950 period, the OCC1990 and IND1990 codes to analyse the 1950-1990 period and the OCC2010 and IND2010 codes to analyse the 2000-2010 period.[2] Unfortunately the 1890 period is missing due to a fire that destroyed these records. We also omit the year 1940 due to significant data inconsistencies (most likely due to the Second World War.)[3]

To apply the Hausmann et al. (2014) algorithm to the IPUMS data, we first construct binary matrices W of state employment concentrations for each time period. These matrices are analogous to the binary M country-product matrix for export data. We calculate separate W matrices for state industrial employment concentrations (based on the harmonised IND codes) and state occupational employment concentrations (based on the harmonised OCC codes).

Each element W_{si} relates to a state's location quotient in industry or occupation i, given by

$$LQ_{si} = \frac{e_{si}/\sum_i e_{si}}{\sum_s e_{si}/\sum_s \sum_i e_{si}}, \tag{8}$$

where e_{si} is the number of people employed in industry or occupation i in state s. Analogous to the approach taken for country-export data, we let $W_{si} = 1$ if $LQ_{si} > 1$ and $LQ_{si} = 0$ otherwise.

We then define $\widetilde{W}$ as

$$\widetilde{W} = \widetilde{D}^{-1}\widetilde{S}, \tag{9}$$

where $\widetilde{D}^{-1}$ is the diagonal matrix formed from the diversity vector $d_s = \sum_i W_{si}$ of state employment concentrations (i.e the row sums of W) and $\widetilde{S}$ is a symmetric state similarity matrix whose rows and columns correspond to states, and whose entries are given by

$$S_{ss'} = \sum_i \frac{W_{si}W_{s'i}}{d_s}. \tag{10}$$

For each time period, we calculate a *State Industrial Complexity* (SIC) index by finding the eigenvector associated with the second-largest eigenvalue of $\widetilde{W}$ based

[2]While the IPUMS project makes the OCC1950 and IND1950 harmonised codes available for the entire 1850-2000 period, our analysis suggests that only backward projection of these harmonised codes (1850-1950) is likely to be appropriate. Using them to project into the future does not account for the introduction of new industry or occupation codes.

[3]The number of US states also changes over the study period. In 1850, the United States comprised of 31 states and experienced subsequent additions until 1960. Our analysis includes data on all states available in each census year.

on industrial employment concentrations. We calculate the corresponding *Industry Complexity* (IC) index by finding the second-largest eigenvalue of the transpose of the $\widetilde{W}$ matrix. We apply the same approach to the $\widetilde{W}$ matrix of state occupational employment concentrations to calculate the *State Occupational Complexity* (SOC) index and *Occupation Complexity* (OC) index.

The SIC (and SOC) indices for states are analogous to the ECI for countries based on export data. They give a one-dimensional ranking of states that places states with similar types of industrial (occupational) employment concentrations close to each other in the ordering and states with di-similar employment concentrations far apart. The IC (and OC) indices for industries (and occupations) are similarly analogous to the PCI index for exported products.

4 Explaining Differences in Economic Prosperity

In this section, we explore the relationship between the SIC and SOC indices and state per capita income over all periods for which we have income data. As the Gross Domestic Product (GDP) measure was only developed in the 1960s, we draw on estimates of state personal income per capita from Easterlin (1960), US Department of Commerce (1962) and Klein (2013) for the earlier 1880-1950 years. We unfortunately do not have income data for the 1850-1870 period. More information about these income sources and the differences between state personal income and state GDP can be found in Appendix A.1.

Not surprisingly, the SIC and SOC indices are highly correlated. Over the 1880-2010 period, the Pearson correlation coefficient between these two measures lie between 0.74 and 0.93 (see Fig 7 in Appendix A.2). This indicates that states that have similar industrial employment concentrations also tend to have similar occupational concentrations.

In Table 1, we examine the relationship between the indices and state per capita income. Model (1) and (2) regress state per capita income on the SOC index (with and without state fixed effects), Model (3) and (4) perform the same analysis with the SIC index, and Model (5) and (6) include both SIC and SOC variables in the regression. While the SIC and SOC indices are both strongly positively correlated with state per capita income over the 1880-2010 period, we find that the SOC index consistently has a slightly greater ability to explain variance in state per capita income than the SIC index.

We suggest two possible reasons for the comparably stronger association between the SOC index and state per capita income. In the periods prior to 1910 the US

Table 1: Regression analysis comparing the relationship between state per capita income and the SIC and SOC indices

	Dependent variable:					
	Log state income per capita (1880-2010)					
	(1)	(2)	(3)	(4)	(5)	(6)
State Occupational Complexity	0.230***	0.114***			0.241***	0.067***
	(0.010)	(0.013)			(0.019)	(0.025)
State Industrial Complexity			0.196***	0.079***	−0.012	0.042**
			(0.012)	(0.010)	(0.021)	(0.018)
Constant	5.136***	4.868***	5.136***	4.778***	5.136***	4.840***
	(0.062)	(0.075)	(0.066)	(0.074)	(0.062)	(0.077)
Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects?	No	Yes	No	Yes	No	Yes
Observations	593	593	593	593	593	593
Adjusted R^2	0.986	0.992	0.982	0.991	0.986	0.992
Note:						*p<0.1; **p<0.05; ***p<0.01

census only asked respondents to report their occupation – no separate industry question was posed. While it is usually straightforward to infer the appropriate industry form the occupational response, in some cases it is less clear. As such, industry coding for these earlier periods is less precise (Rondander, 1999). However, the stronger relationship between the SOC index and income per capita is not only limited to the period prior to 1910; we find this pattern to extend over the entire 1880-2010 period.

A second possible reason which could explain this phenomenon in the later years is the evolution from sectoral to functional specialisation. Rather than being home to an industry's entire value chain, cities and regions have experienced an increasing tendency to be specialised in functions (Duranton and Puga, 2005; Rossi-Hansberg et al., 2009) (or even tasks (Acemoglu and Autor, 2011)), with firms locating management and administrative headquarters in larger cities and plants or production activities in other areas. As such, the SOC index based on state occupational employment concentrations may better capture this spatial separation of activities along the industrial value chain. From this perspective, the index's closer alignment with per capita income could suggest that functional, rather than sectoral specialisations have stronger implications for a region's average level of economic prosperity.[4] For these reasons, we focus the most of the analysis in the remainder of this paper on the SOC index.

It is important to highlight that the high adjusted R^2 is primarily driven by the

[4] Martin et al. (2018) similarly suggest that functional structure may be more important for productivity growth than sectoral specialisation.

inclusion of the year fixed effects, which alone account for 97.3% of the variance in state per capita income. However, without year fixed effects, the SOC index alone still captures a fairly large proportion of the variance in state per capita income. As shown in the year-on-year correlations in Figure 1, the SOC index consistently accounts for at least 24% of the variance in state per capita income over the 1880-2010 period, and in some years explains up to 86%.

In Table 2, we also test whether the close association between the SOC index and state per capita income is robust to a number of controls. We find that the relationship remains strongly significant, even with the inclusion of variables relating to state population, occupational diversity (d_s) and each state's proportion of overseas and interstate workforce.

Table 2: Regression analysis testing the robustness of the relationship between SOC and state per capita income to a series of control variables

	Dependent variable:					
	Log state income per capita (1880-2010)					
	(1)	(2)	(3)	(4)	(5)	(6)
State Occupational Complexity	0.173***	0.126***	0.141***	0.052***	0.119***	0.075***
	(0.015)	(0.012)	(0.013)	(0.011)	(0.017)	(0.011)
Log State Pop	−0.108***	−0.215***			−0.063***	−0.162***
	(0.011)	(0.028)			(0.010)	(0.024)
State Occupational Diversity	0.004***	0.004***			0.002***	0.003***
	(0.001)	(0.0003)			(0.001)	(0.0003)
Proportion overseas workforce			1.428***	1.644***	1.204***	1.283***
			(0.163)	(0.170)	(0.153)	(0.159)
Proportion interstate workforce			0.388***	0.075	0.246***	0.104
			(0.065)	(0.100)	(0.061)	(0.095)
Constant	6.382***	7.716***	4.660***	4.536***	5.477***	6.747***
	(0.166)	(0.408)	(0.047)	(0.052)	(0.152)	(0.346)
Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects?	No	Yes	No	Yes	No	Yes
Observations	593	593	593	593	593	593
Adjusted R^2	0.989	0.994	0.990	0.994	0.991	0.995
Note:						*p<0.1; **p<0.05; ***p<0.01

Population, overseas and interstate workforce data are also drawn from the IPUMS database (Ruggles et al., 2017). Proportion of overseas and interstate workforce statistics are calculated on the basis of the IPUMS 'Birthplace' variable, which indicates the US state or foreign country in which a worker was born.

Table 3 explores the relationship between a state's SOC index, and its future economic growth rate. In Model (1), we regress states' decadal compound annual growth rate against their SOC scores and control for their initial level of per capita income. We find that states with higher SOC tend to have higher future growth rates. In Models (2), (3) and (4), we show that this result is also robust to a series of control variables.

Table 3: SOC Growth Regression Analysis

	Dependent variable:			
	Compound Annualised Growth Rate per capita (by decade): 1880-2010			
	(1)	(2)	(3)	(4)
State Occupational Complexity	0.002**	0.002*	0.002***	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
Initial Log Income Per Capita	−0.023***	−0.026***	−0.024***	−0.026***
	(0.003)	(0.004)	(0.004)	(0.004)
Log State Pop		−0.001		−0.001
		(0.001)		(0.001)
State Occupational Diversity		0.0001		0.0001
		(0.00003)		(0.00003)
Proportion Overseas Workforce			−0.0005	−0.002
			(0.010)	(0.010)
Proportion Interstate Workforce			0.005	0.004
			(0.004)	(0.004)
Constant	0.130***	0.150***	0.134***	0.147***
	(0.017)	(0.022)	(0.020)	(0.023)
Year Fixed Effects?	Yes	Yes	Yes	Yes
Observations	542	542	542	542
Adjusted R^2	0.875	0.875	0.875	0.875
Note:				*p<0.1; **p<0.05; ***p<0.01

Let us pause briefly to digest what these regression results mean in light of our interpretation of the SOC index as a one-dimensional ordering that places states with similar types of occupational specialisations close together in the ranking and those with dis-similar types of occupational specialisations far apart. The strong and robust relationship between the SOC index and state per capita income portrayed in Table 2 suggests that over the 1880-2010 period, states with similar *types* of occupational concentrations tend to have similar levels of per capita income.

This finding is, in a sense, both trivial and profound. It is trivial if one considers it in the two-region setting. Imagine two states that had identical economic structures. It would probably be fairly surprising if they *didn't* have very similar levels of per capita income. Now imagine 51 states with economic structures having varying degrees of similarities. It is not immediately obvious that if you collapse this high dimensional space of state occupational specialisations into a single ordering reflecting relative state similarities, this sequence would correlate so strongly with the ordering of state per capita income. And yet, this result holds over the 130 year period for which we have state income data.

Returning to our two-region setting - suppose we did find two states that had identical economic structures, but *different* levels of per capita income. Most economists would expect some degree of convergence in incomes - the state with lower per capita income should be likely to grow relatively faster, and 'catch up' with the living standards enjoyed by the other state (see for example Kim (1998)). And indeed, this basic intuition is reflected in our growth regression analysis: states that have a higher SOC index than one would expect given their current per capita income (i.e. look more similar to wealthier states) do tend to grow faster over the subsequent decade.

How does a measure that captures the *type* of occupational specialisations differ to measures that reflect the *number* or *spread* of state occupational specialisations? In Figure 1, we compare the SOC index to state occupational diversity (d_s), the inverse Herfindahl-Hirschman Index (HHI) and Shannon's Entropy in terms of their relationship with state per capita income over the 1880-2010 period. These three measures each capture related notions of the extent to which a state's employment distribution is diversified over different occupational categories. State occupational diversity simply sums up the number of occupations for which a state has a $LQ > 1$. The inverse HHI sums up the square of each state's employment share s_i in each occupation i (and takes the inverse), and is given by

$$Inv\ HHI = \frac{1}{\sum_i s_i^2}. \tag{11}$$

Shannon's entropy treats state employment shares as a probability distribution and calculates the degree of uncertainty that a given worker in a state will be employed in a particular occupation

$$Entropy = -\sum_i s_i \log(s_i). \tag{12}$$

For each of these measures, higher (lower) values indicate greater (lesser) state employment diversification. As shown in Figure 1, states with higher diversification tended to have higher levels of per capita income up to around 1960. However, this association starts breaking down from 1970 onwards. In contrast, the SOC index continues to be strongly positively correlated with state per capita income over the entire 1880-2010 period. This suggests that over time (most likely due to reductions in transport and communication costs (Gaspar and Glaeser, 1998; Glaeser and Ponzetto, 2007; Michaels, 2008; Michaels et al., 2013; Leamer and Storper, 2014)) the *type* of occupations states are specialised in has become more important in distinguishing between richer and poorer places.

Figure 1: Comparison of the relationship between log state per capita income and Herfindahl-Hirschmann index (HHI), Occupational Diversity, Entropy and the SOC index over the 1880-2010 period.

5 What separates rich and poor states?

Relative to poorer places, what type of occupations are richer places concentrated in? Recall from Section 2 that in the context of country-export data, the PCI metric helps us identify the types of products that countries at one end of the ECI spectrum have in common. In this section, we draw on the Occupational Complexity (OC) index to identify the types of occupations that states with high (low) SOC scores have in common.

It is important to emphasise that the tendency for places with high (low) SOC scores tend to be more (less) prosperous is *not in-built* into these metrics. The algorithm only operates on the binary pattern of state employment concentrations in the W matrix, and takes in no further information on income, wages, skill level, or occupational classification (such as agricultural, manufacturing or services). However, as we demonstrated in the previous section, the SOC index tends to consistently distinguish between states with higher and lower levels of income - and this strong and robust result reflects the differences in the types of occupations concentrated in richer and poorer places.

To test this even further, we calculate a measure that explicitly captures an occupation's concentration in prosperous places. Inspired by the *'PRODY'* income-content measure for exported products (Hausmann et al., 2007), we calculate an *Occupational Income Content* ($OccY$) measure, which calculates the average per capita income of states that have a $LQ > 1$ in a particular occupation. That is, for each occupation i, $OccY_i$ measures the average prosperity of states that are concentrated in it, and is given by

$$OccY_i = \frac{1}{\sum_s W_{si}} \sum_s W_{si} Y_s, \tag{13}$$

where W_{si} relates to elements in the binary W matrix of state occupational concentrations, and Y_s relates to the per capita income of state s.

Figure 2 shows the Pearson R correlation coefficient for the year-on-year correlations between each occupation's $OccY$ and Occupational Complexity index over the 1880-2010 period. This strong relationship is on par with the year-on-year correlations between SOC and state per capita income over the same period (see Figure 1), and provides further empirical evidence that the occupational differences separating rich and poor regions are well captured by the OC measure.[5] In Appendix A.4, we show that this relationship is robust to a series of controls including occupational wage, educational requirement and total employment.

[5] A more subtle point is that, unlike the ordering given by $OccY$, which is directly related to per capita income, the OC index reflects the similarity gradient specifying the direction along which the similarity of occupations (in terms of the locations they are concentrated in) changes as much as it possibly can. To our knowledge, there is no theory that explains why this should be the case. As such, it opens the door for new developments in location theory to explain this robust empirical pattern.

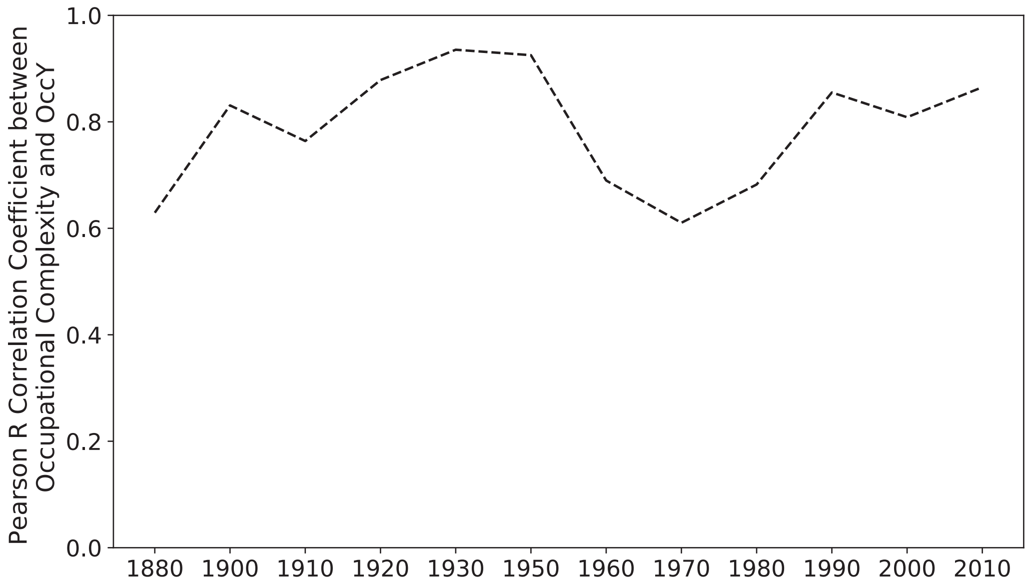

Figure 2: Pearson R correlation coefficient for the relationship between Occupational Income Content (OccY) and Occupational Complexity over the 1880-2010 period.

In Table 4, we show the highest and lowest ranked occupations in terms of their OC index for 2010. A cursory comparison of the top and bottom ranks suggests that, not surprisingly, prosperous places tend to be concentrated in high-skill services, while less prosperous places are underpinned by industrial activities. However, reflecting the surrounding ecosystem of activities that co-evolve with a region's tradable activities, a number of lower skilled occupations, such as baggage porters also occupy the top-ranks.[6]

While one might be tempted to refine the SOC and OC indices by activities that are typically non-traded from the W matrix, we show in Appendix A.6 that this reduces the ability of the SOC index to explain variance in state income per capita and predict future growth. The non-tradable supporting ecosystem of increasingly specialised activities turn out to be particularly important in distinguishing between rich and poor places. And indeed it is worth noting that even Adam Smith

[6]This result is consistent with Shutters et al. (2016), who examined the concentration of 'creative' and knowledge-intensive cities in US metropolitan statistical areas (MSAs). They found a tendency for 'creative' urban economies to specialise in both creative and non-creative occupations and suggested that this could be because particular non-creative occupations complement the tasks performed by creative occupations.

(1776) found cause to highlight the unique presence of particular activities that are *only* able to exist in prosperous places:

> *"There are some sorts of industry, even of the lowest kind, which can be carried out nowhere but in a great town. A porter for example, can find employment and subsistence in no other place."*

OC Rank	Occupation
1	Lawyers, and judges, magistrates, and other judicial workers
2	Actors, Producers, and Directors
3	Financial Analysts
4	Securities, Commodities, and Financial Services Sales Agents
5	Accountants and Auditors
6	Editors, News Analysts, Reporters, and Correspondents
7	Software Developers, Applications and Systems Software
8	Computer Scientists and Systems Analysts/Network systems Analysts/Web Developers
9	Personal Financial Advisors
10	Baggage Porters, Bellhops, and Concierges
11	Managers, nec (including Postmasters)
12	Taxi Drivers and Chauffeurs
13	Medical Scientists, and Life Scientists, All Other
14	Writers and Authors
15	Managers in Marketing, Advertising, and Public Relations
16	Architects, Except Naval
17	Other Business Operations and Management Specialists
18	Parking Lot Attendants
19	Management Analysts
20	Designers
…	…
434	Molders and Molding Machine Setters, Operators, and Tenders, Metal and Plastic
435	Chefs and Cooks
436	Industrial Production Managers
437	Paper Goods Machine Setters, Operators, and Tenders
438	Stock Clerks and Order Fillers
439	Industrial and Refractory Machinery Mechanics
440	Clergy
441	First-Line Supervisors of Production and Operating Workers
442	Industrial Truck and Tractor Operators
443	Woodworking Machine Setters, Operators, and Tenders, Except Sawing
444	Metal workers and plastic workers, nec
445	Laborers and Freight, Stock, and Material Movers, Hand
446	Other production workers including semiconductor processors and cooling and freezing equipment operators
447	Assemblers and Fabricators, nec
448	Welding, Soldering, and Brazing Workers
449	Millwrights
450	Driver/Sales Workers and Truck Drivers
451	Grinding, Lapping, Polishing, and Buffing Machine Tool Setters, Operators, and Tenders, Metal and Plastic
452	Cutting, Punching, and Press Machine Setters, Operators, and Tenders, Metal and Plastic
453	Extruding, Forming, Pressing, and Compacting Machine Setters, Operators, and Tenders

Table 4: Pearson R correlation coefficient for the relationship between Occupational Income Content (OccY) and Occupational Complexity over the 1880-2010 period.

6 New insights into shifting economic landscapes

How has the type of occupations underpinning prosperous places has changed over time? And what implications might this have for long term regional development? In this section we investigate the evolution of SOC and OC indices over the 1850-2010 period.

Examining the trajectory of detailed occupational OC scores over such a long time horizon is complicated by changes occurring in the harmonised occupational classification scheme in 1950-1960, and 1990-2010. As detailed occupations across these classifications are not directly comparable, we look at the average OC values of detailed occupations falling into broader categories that we can compare across the entire period.[7]

In Figure 3, we show the average OC scores of occupations falling into three broad categories: professional, technical and finance occupations (including management, scientific, financial and other professional and technical activities), service occupations (including retail and administrative activities) and production (construction, extraction and manufacturing) occupations. Figure 3 reveals a fairly striking shift in OC scores from 1980 onwards. Professional, technical and finance occupations experience a distinct rise in their average OC scores, while production activities see a marked decline. The distance between these average OC scores also become much larger from 1980 onwards, as shown by the error bars which represent the 95% confidence interval around the mean values.

[7] More about this mapping of detailed occupations into broader categories can be found in Appendix A.3.

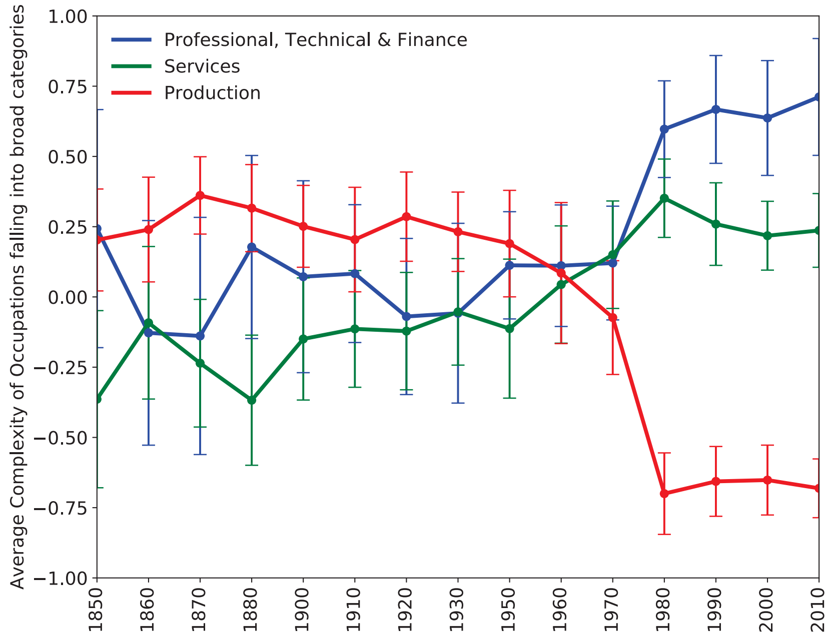

Figure 3: Average OC scores of Production, Services and Professional, Technical and Finance Occupations. Error bars relate to the 95% Confidence Interval.

It is important to note that this distinct change does not reflect changes in the occupation classification (these changes occurred in over 1950-1960, and 1990-2000). As we show in the Appendix, this structural shift in OC scores can also not be well accounted for by changes in occupational employment shares in these broad categories (as shown in Figure 13, this evolution was much smoother). Moreover, a very similar shift is evident when analysing metropolitan (rather than state) level data (see Figure 12), and when analysing industry complexity index scores (see Figure 11).

Before we digest the OC dynamics further, we first examine the corresponding evolution of state SOC scores over the 1850-2010 period. Figure 4 shows four key groups of states that experienced similar SOC development trajectories. While California and a number of East-coast states maintained consistently high SOC scores over the entire period (Panel A), Southern agricultural-based states tended to have comparatively low SOC scores (Panel B). Panel C depicts many resource-rich states located in the US's North-West Mid-West region, which have shared similar SOC trajectories, but perhaps the most striking SOC dynamics occur in rust-belt states in Panel D. While rust-belt states had relatively high SOC scores (comparable to states in Panel A) for most of the earlier development phase, the

1970-1980 period sees their SOC scores decline dramatically.

Figure 4: Evolution of State Occupational Complexity over time for a selection of US states with similar SOC trajectories.

Why did the OC and rust-belt SOC scores experience such a significant shift in 1970-1980? This particular decade is historically interesting on a number of fronts. The 1973 oil-crisis, which led to a quadrupling of oil prices, severely impacted US automotive businesses that were concentrated in rust-belt states (Bresnahan and Ramey, 1993; Lee and Ni, 2002). With higher fuel prices and severe gas shortages, domestic demand for 'gas gazzling' vehicles, which places like Detroit were famous for, shifted to more efficient models imported from Japan (Carrillo, 2004). However, for states endowed with coal, oil and gas reserves, the oil-crisis generated a short-lived boom, as higher energy prices stimulated mining and drilling activities around the country (Black et al., 2005; Allcott and Keniston, 2017).

At the same time, computers began proliferating the workforce - augmenting the productivity of higher skill occupations specialised in cognitive tasks and reducing demand for labour concentrated in routine and manual activities (Autor et al., 2003; Berger and Frey, 2016). However, skill-biased technological change was not the only force impacting US labour markets. As highlighted by (Autor, 2015, p10)

> *"Many forces distinguish the labour markets of these two epochs of 1940-1980 and 1980-2010. A partial list would include changes in the relative supply of college and non-college labour, rising trade penetration offshoring and globalization of production chains, declines in labour union penetration, the changing 'bite' of the minimum wage, and certain shifts in tax policy. Of course, many of these factors combine and interact as well such that attributing changes to a single cause would be foolish."*

Heeding Autor's words, we do not attempt to identify causal channels for the changes in SOC and OC scores in this paper. However, we do seek to reflect on the insights that these metrics give us about the changes in the US economic landscape over this period.

As reflected by their high SOC scores, both rust-belt states and East-coast (and California) states were concentrated in relatively similar types of occupations in the period leading up to 1980. However in 1980, the SOC scores of these two groups of states diverge – signalling different future economic prospects for these states. The changes in OC scores in Figure 3 help identify the nature of these differences. East-coast states (and California) become much more concentrated in high-skill professional services, while rust-belt states remained largely specialised in manufacturing.

We stress again that the algorithm for calculating SOC and OC metrics is blind to the type (or classification) of occupations states are concentrated in – it only operates on the binary pattern of 1's and 0's in the W matrix and finds the direction in which states and occupations are maximally separated, in terms of the similarity in their employment concentrations. Yet this pattern of similarities appears particularly useful for identifying changes in relevance (or 'fitness') of particular activities for generating prosperity – without having any a priori theoretical intuition.

Why were the rust-belt states less able to transition to high-skill service activities like their East-coast counterparts? Figure 5 suggests that differences in education had a lot to do with it.[8] As shown in Panel A of Figure 5, which plots 1970 SOC scores against 1970 average state educational attainment levels, education does not well separate states at either ends of the SOC spectrum. However, a different picture emerges in 1980. Panel B shows that 1980 SOC scores are well explained by differences in 1970 average state educational attainment levels. And,

[8] Entrenched unionism and lack of competition in rustbelt states' labour and output markets may also have played a key role (Alder et al., 2014)

indeed as we show in Figure 6, which shows the year-on-year Pearson correlation coefficient between the SOC index and average state educational attainment for all years education data is available, the relationship between the SOC index and statement education attainment shows a marked shift in 1980. While it comes as no surprise that in the new knowledge economy, education clearly matters – we find that the SOC and OC indices have a unique ability to pin-point what has become an increasingly important divide (Berry and Glaeser, 2005).

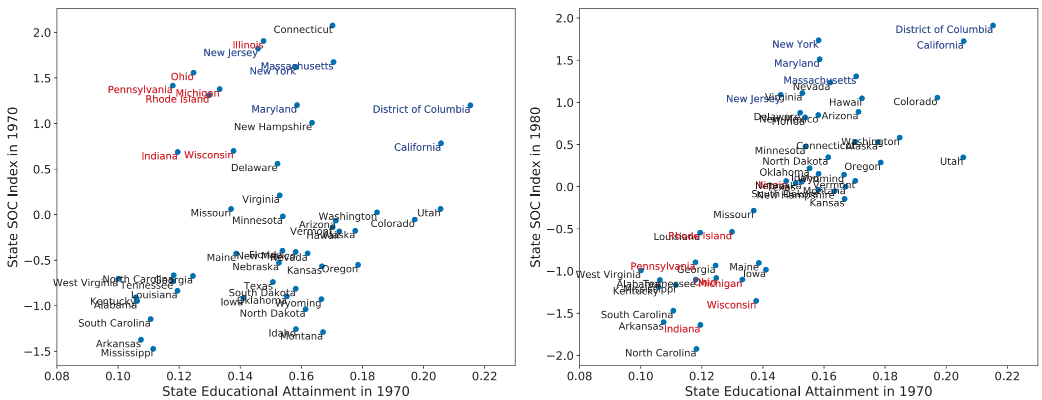

Figure 5: Panel A: Relationship between State SOC index in 1970 and State Educational Attainment in 1970. Panel B: Relationship between the SOC index in 1980 and State Educational Attainment in 1970. Rust-belt states are labelled in red, and East-coast states (and California) that consistently maintained high SOC scores are labelled in blue. State educational attainment are drawn from the IPUMS database (Ruggles et al., 2017) and calculated as the proportion of the state workforce having completed at least one year of college as given by the 'EDUC' variable.

Figure 6: Pearson R correlation coefficient for the relationship between the SOC index and average state educational attainment over time.

One further important topic that these dimension tools could shed light on is regional income convergence. Although considerable economic disparities have long separated US states, throughout most of the 20th century poorer states were tending to grow faster relative to richer states (Barro and Sala-i-Martin, 1992). However, this process slowed substantially from 1980 onwards. A number of scholars have suggested that this phenomenon could be underpinned by the increasing returns to skills and innovation, which have become much more concentrated in cities (Glaeser and Saiz, 2003; Berry and Glaeser, 2005). As highlighted by Moretti (2012), p 106,

> *"It is almost as if, starting in the 1980s, the American economy bifurcated. On the one side, cities with little human capital and traditional economies started experiencing diminishing returns and stiff competition from abroad. On the other, cities rich in human capital and economies based on knowledge-intensive sectors started seeing increasing returns and took full advantage of globalised markets."*

Interestingly, this 'bifurification' is well captured by the OC and SOC dynamics over this period. In Figure 3, not only do we observe a new polarisation of production and high skill services at either end of the PCI spectrum in 1980 (suggesting the activities became much more strongly spatially segregated), we also observe the error bars becoming much further apart.

Undertaking an in-depth study of the drivers of regional income convergence and divergence is beyond the scope of this paper. However, we believe that the SOC and OC indices could be useful additions to the current set of analytical tools available to investigate this phenomenon, complementing existing measures such as the Krugman Index of 'dissimilarity' (Krugman, 1991, 1993), which has been used to examine regional convergence in the US (Kim, 1998) and the UK (Martin et al., 2018).[9] [10] Moreover, to the extent that convergence and divergence may depend on the type of economic activities that richer and poorer states are specialised in, the metrics we have described here could be particularly advantageous.

7 Conclusion

This paper has used a dimension reduction algorithm, which has previously been applied to country-export data (Hidalgo and Hausmann, 2009; Hausmann et al., 2014), to study differences in state employment concentrations over 1850-2010. This algorithm reduces the high-dimensional space of state specialisations into two single-dimension vectors that capture relative similarities in the *type* of economic activities states are concentrated in. In particular, the SOC index examined in this paper places states with similar occupations close together in the ranking and states with dissimilar occupations far apart. The corresponding OC index helps identify the types of occupations that distinguish states with high SOC scores from states with low SOC scores.

Despite operating only on the binary pattern of state location quotients in different occupations, we find that the resulting SOC and OC indices provide remarkably consistent insights into the types of activities distinguishing rich and poor places over the 130 year period for which state income data is available. The SOC index also outperforms measures of diversification, which become less useful in explaining differences in per capita income from 1930 onwards. It also proves to be a useful predictor of future economic growth.

While these findings align with intuition based on a two-region setting (two states with similar occupations should converge to similar levels of per capita incomes), we are not aware of any theories which explain why the linear ordering of states along the dimension captured by the SOC index should correlate so strongly with the ordering of state per capita incomes. New theories are needed to explain this robust empirical pattern.

[9] We make comparisons to the Krugman Index over the 1850-2010 period in Appendix A.7.

[10] Berry and Glaeser (2005) also utilise a dis-similarity index and an isolation index to show that skilled workers have become more segregated across cities. While these results are consistent with the OC dynamics (which also hold at metro level, as shown in Figure 11), the OC measures have the advantage of not requiring information or a priori intuition on skilled workers.

We also find that in this particular setting, the SOC and OC indices demonstrated a unique ability to capture changing relevance of particular types of productive activities as the US economy underwent a fundamental structural shift in 1970-1980. While this finding suggests that these metrics could provide a new analytical lens to study structural change and regional convergence patterns more generally, future work is needed to verify this in other contexts undergoing similar patterns of change. The UK regional economy could be a particularly interesting test case (Martin et al., 2018).

Finally, we believe the bifurification of SOC and OC scores in 1980 warrants further attention. The divergence in the SOC trajectories and economic prospects of rustbelt states remains pertinently relevant today. While more research is required to understand whether the *type* of differences separating regions could influence a nation's capacity for inclusive growth, we believe that these new metrics could help shed light on this important question.

A Appendix

A.1 Historical State Income Data

Studying economic development over such a long stretch of history is complicated by the fact that state Gross Domestic Product (GDP) is only available from 1963 onwards. For previous years, we draw on estimates of state personal income per capita. Although GDP and personal income typically show very similar trends, the measures do have some key differences in their calculation. Personal income represents the value of economic activity received by people in a given economy. It includes wages, benefits, proprietor income, dividends, interest, rent and transfer payments (such as social security). GDP represents the value added in production by the labour and capital located in a state. It is calculated by summing up a state's gross output (sales or receipts and other operating income, commodity taxes and inventory change), minus its intermediate inputs (consumption of goods and services from other US industries or imports)(Landefeld and Marcus, 2006).

Unfortunately, state-level estimates of personal income for these earlier years are not available from a single source. Moreover, to our knowledge, no estimates are available for 1850-1880. We consequently draw on multiple sources to arrive at state-level estimates for the years 1880-1950. For the years 1880-1910 we draw on estimates available in Klein (2013). For 1920, we employ Easterlin (1960)'s estimates. For the period 1930-1950, we draw on a statistical abstract from the US Department of Commerce (1962).

A.2 Correlation between State Industrial Complexity and State Occupational Complexity Indices

In Figure 7, we show the Pearson correlation coefficient for the relationship between the SIC and SOC indices for the period 1880-2010. Throughout this period, these measures are very strongly correlated, suggesting that states that have similar occupational structures also have similar industrial structures.

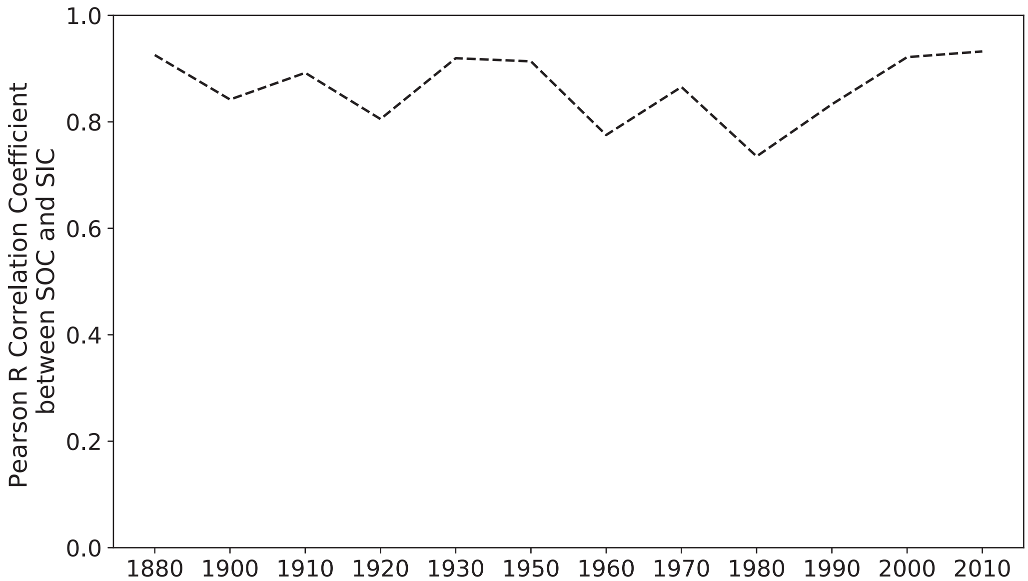

Figure 7: Pearson Correlation between State Industrial Complexity and State Occupational Complexity indices over 1880-2010.

A.3 Aggregation of 3-digit IPUMS Occupations and Industries to broad categories

Analysing the OC and IC scores for detailed occupations and industries over such a long time horizon is complicated by changes occurring in the harmonised classification scheme in 1950-1960, and 1990-2000. As 3-digit occupation and industry codes across classifications are not directly comparable over these periods, we construct broader classification groupings, and arrive at the results in Figure 3 and Figure 11 by calculating the average OC scores of 3-digit occupations falling into these broader groups over time.

For occupations, we primarily construct the broad 'Production', 'Services' and 'Professional, Technical and Finance' groupings for the OCC1950, OCC1990 and OCC2010 classifications by drawing on each 3-digit occupation's more aggregated 2-digit occupation, and then further aggregating the relevant 2-digit occupations into the relevant broad groupings. However, in some cases the 2-digit categories do not perfectly align with the broader grouping. For example in the OCC1950 classification, most occupation categories receiving the 2-digit category 'Professional, Technical' are assigned to the 'Professional, Technical and Finance' Group.

However, although the 3-digit occupations 'Religious workers' and 'Music teachers' are assigned to the 'Professional, Technical' category, we assign these occupations to the Services category. We undertake a very similar procedure for grouping IND1950, IND1990 and IND2010 3-digit occupations into the broad industry categories shown in Figure 11. Crosswalk files specifying the treatment of each 3-digit occupation and industry code are available on request.

A.4 Additional analysis and robustness tests for the OC index

In this section, we examine the relationship between the occupational OC index and occupational wage, educational requirements and total employment over time. We also test whether the relationship between the OC index and the occupational *OccY* measure is robust to a series of control variables.

The strong tendency for SOC and OC to correlate with income-related measures does lead one to question whether we could just be capturing wage-related effects associated with occupations. To test this, we draw on IPUMS occupational earnings variables and examine the relationship between an occupation's median wage and its complexity score. As we have three different occupational classifications relating to 1950, 1990 and 2010 occupational codes, we test the relationship for each classification year.[11] As shown in Figure 8, we find a fairly weak positive relationship between occupational earnings and OC scores, suggesting that our complexity measures are capturing more than general wage effects.

[11] ERSCOR50 and ERSCOR90 are IPUMS variables relating to the standardised median income earned by an occupation in 1950 and 1990. ERSCOR is not available for 2010, so we calculate the occupation's standardised median income on the basis of 2010 census data.

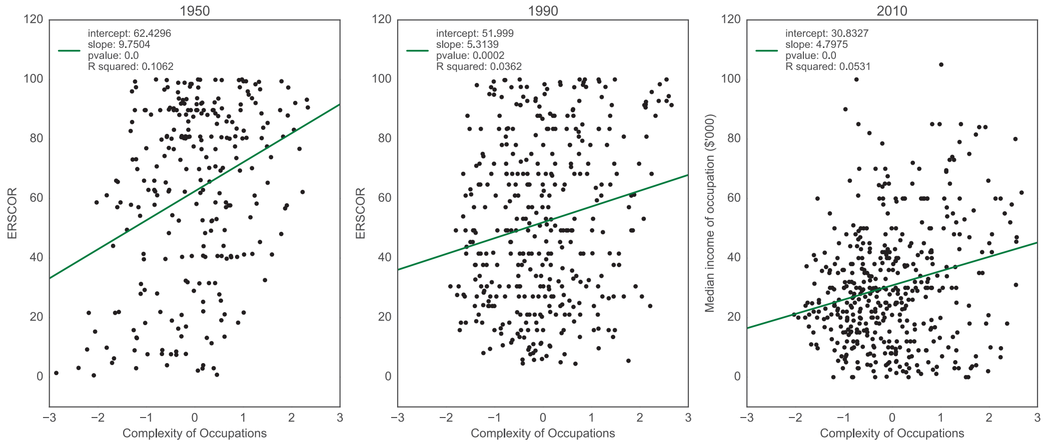

Figure 8: Complexity of occupations vs median earnings for each occupation

Since our SOC and OC measures are based on employment concentrations in each state, it could also be the case that we are capturing effects related to the sheer number of people employed in different occupations. As such, we investigate the relationship between the total number of people employed in each occupation and its complexity score in Figure 9. Again, we find limited evidence that OC scores are capturing effects associated with the quantity of people employed in each occupation category.

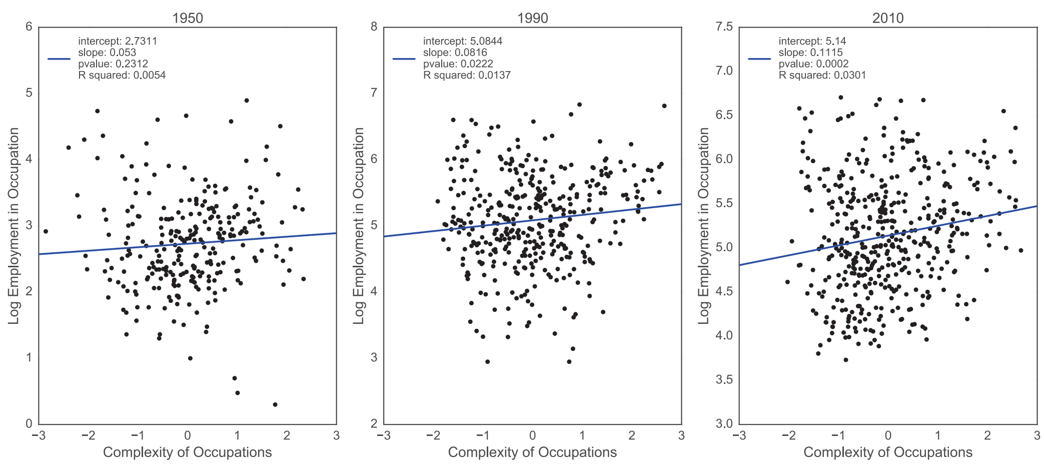

Figure 9: Complexity of occupations vs total employment in each occupation

Finally, in Figure 10 we consider the relationship between the OC index and educational requirements associated with each occupation. This time, we draw on

IPUMS education attainment variables associated with each occupation for 1950, 1990 and 2010. [12]

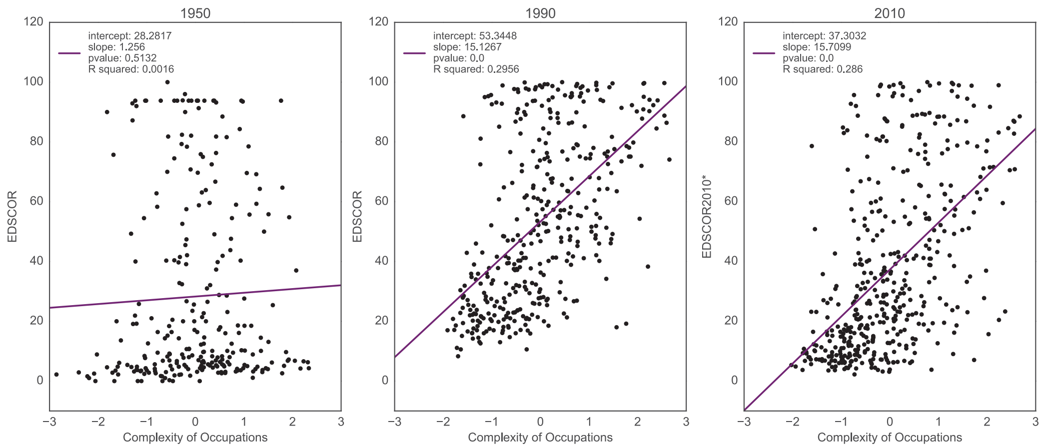

Figure 10: Complexity of occupations vs each occupation's educational requirement

Interestingly in 1950 (the first year that educational attainment data is available), we find no correlation between an occupation's OC score and its associated educational requirements. However, in more recent years we find that OC is more strongly positively correlated with occupational education requirements, suggesting that over time, OC and skill content have become more closely aligned.

As a final robustness test, we examine how the relationship between OC and OccY changes when we control for each occupation's earnings, employment and educational requirements. As the occupation classification changes in 1950, 1990 and 2010, we cannot undertake a panel regression, but instead estimate a regression model for each classification year:

$$OccY_i = \beta_1 OC_i + \beta_2 wage_i + \beta_3 edu_i + \beta_4 lemp_i + u_i \tag{14}$$

where $wage_i$ relates to the standardised median income earned by a given occupation, edu_i relates to the percentage people in each occupation category that have completed more than 1 year of college, $lemp_i$ represents the log of the number of people employed in each occupation and u_i represents the error term.

[12]EDSCOR50 and EDSCOR90 are IPUMS variables relating to the percentage of people in each occupation category that have completed at least one year of college in 1950 and 1990. EDSCOR is not available for 2010, so we use 2010 census data to calculate the percentage of people in each occupational category that have completed at least one year of college

We present the results of these regressions in Table 5. Model (1) and (2) relate to the 1950 classification scheme and compare the change in the OC variable's explanatory power when we introduce earnings, employment and education controls. Model (3) and (4) similarly make comparisons for the 1990 classification while Model (5) and (6) relate to the 2010 classification.

Table 5: Regression analysis of the relationship between the Occupational Complexity Index and Occupational Income Content ($OccY$)

	Dependent variable: OccY (Occupational Income Content)					
	1950		1990		2010	
	(1)	(2)	(3)	(4)	(5)	(6)
Occupational Complexity Index	0.080***	0.080***	0.077***	0.079***	0.083***	0.081***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)
Occupational Median Income		−0.00004		−0.00001		0.0000
		(0.0001)		(0.0001)		(0.000)
Occupational Educational Requirement		0.0003***		−0.0001		0.0001
		(0.0001)		(0.0001)		(0.0001)
Log Total Employment		0.004		0.004		0.003
		(0.004)		(0.003)		(0.004)
Constant	7.287***	7.270***	10.036***	10.027***	10.778***	10.760***
	(0.002)	(0.012)	(0.002)	(0.019)	(0.002)	(0.019)
Observations	268	268	383	383	453	453
Adjusted R^2	0.855	0.863	0.740	0.742	0.755	0.754
Note:						*p<0.1; **p<0.05; ***p<0.01

In all cases, the OC index accounts for the overriding majority of explained variance in occupational income content.

A.5 Robustness of Structural Change Pattern

In this section, we further scrutinize the observed shift in the OC scores over the 1970-1980 period.

First we test whether the distinct change in the OC scores are also present in the industrial complexity indices based on state industrial employment concentrations. Figure 11 shows the average IC scores for 3-digit industries falling into broad categories over time. Here we find a very similar pattern - with the average IC of manufacturing industries falling significantly and the average IC of information and communication industries experiencing a marked rise. The average IC scores of Professional, scientific, technical and financial industries also increase over this period, though not as sharply as information and communication industries.

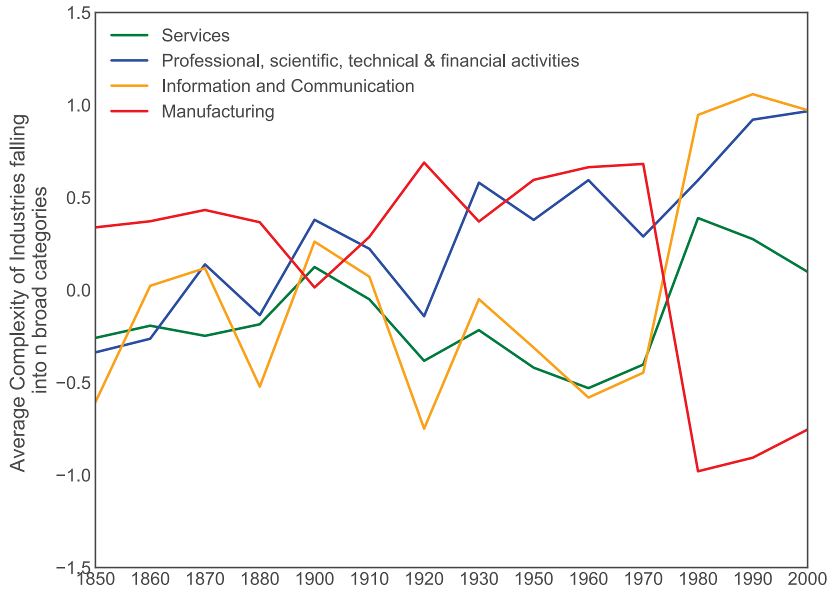

Figure 11: Average Industrial Complexity index scores of broad industry categories over 1850-2010

In Figure 12, we examine whether this change also occurs if we calculate the OC metrics for metro-level regions rather than state level. While metro-level data was only available from 1970 onwards at the time of writing, we again find a similar change occurring over the 1970-1980 period, suggesting that the dynamics are not an artefact of using state-level data.

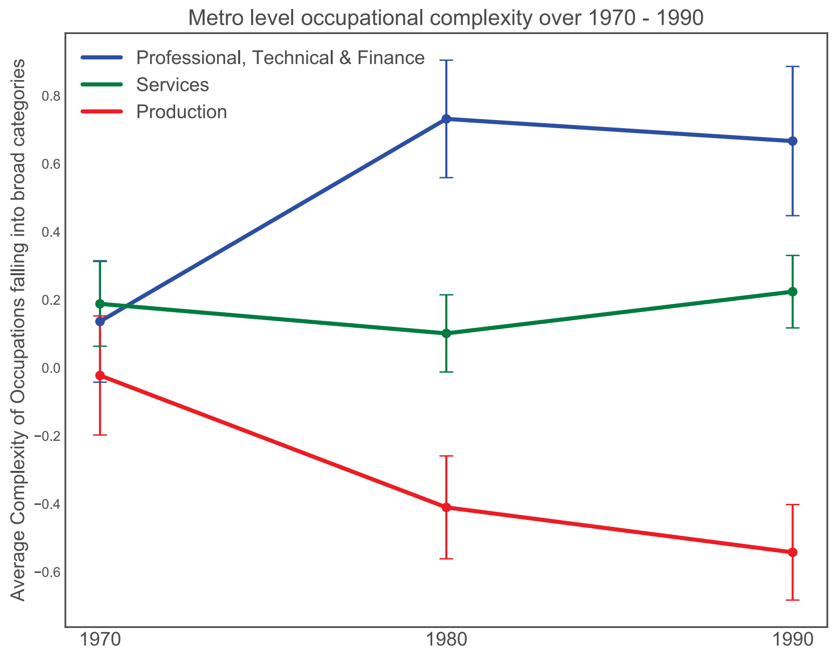

Figure 12: Average complexity of Production, Services and Professional, Technical and Finance Occupations calculated on the basis of metropolitan rather than state regions. Metropolitan data are based on the drawn from the IPUMS database (METAREA variable) (Ruggles et al., 2017), which at the time of writing was only available from 1970 onwards

Finally, we investigate whether the observed structural shift in OC scores can be accounted for by the change in employment shares of these broad categories over time. In Figure 13, Panel A shows the employment shares falling into the occupational broad categories and Panel B shows the employment shares falling into the industrial broad categories. While we do observe a decline in employment in Production and Manufacturing categories, and a rise in both services and professional, technical and finance employment, the changes occur relatively steadily over the studied time period. These smoother changes do not account for the distinct shift in OC and IC scores observed over the 1970-1980 period.

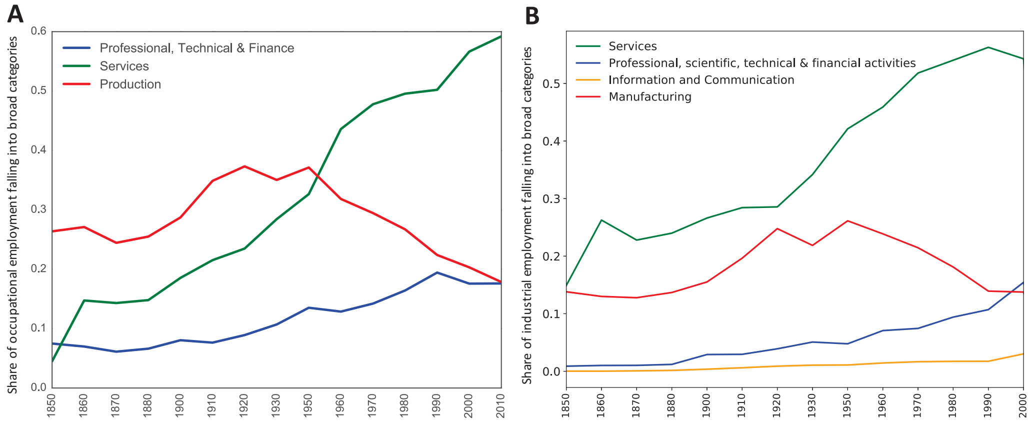

Figure 13: National employment shares over 1850-2010 Panel A shows employment shares in broad occupational categories. Panel B shows employment shares in broad industrial categories.

To summarise, these results suggest that the observed OC dynamics over the 1970-1980 period are not just an artefact of using occupation or state level data. Moreover, they cannot be well accounted for by changes in employment shares over time.

A.6 Tradable and Non-Tradable Analysis

Distinguishing between tradable and non-tradable industries or occupations is a notoriously imprecise activity. A number of approaches aim to use an activity's geographical concentration as a proxy for its 'tradability' (Porter, 2003; Jensen and Kletzer, 2010; Spence and Hlatshwayo, 2012). The key intuition is that traded activities tend to either take advantage of increasing returns to scale or particular inputs such as natural resources or particular types of skills, while non-traded activities tend to be more equally distributed across space, or at least *"follow the geographical distribution of the goods-producing population"* (Krugman, 1991, p 65).

We apply a similar approach that also draws on the work of Youn et al. (2016) who examined the tendency for the number of establishments in different industries (which tends to be closely correlated to employment) to systematically change with city size. They noted that non-tradable activities tended to scale *linearly* (i.e increase in a linear proportion) with a place's population, while tradable activities either scaled sub-linearly (such as agriculture, mining and utilities) or super-linearly (such as professional, scientific and technical and managerial services). Along these lines, a very simple approach for classifying tradable and non-tradable activities is just to measure the extent to which employment in a given activity is linearly correlated with population.

In Table 6, we show the top 20 and bottom 20 occupations ranked by the R-squared for the relationship between employment in an occupation and total population across states for the year 2010. Top ranked occupations closely align with what one would intuitively expect to be non-traded activities, while bottom ranked occupations tend to relate to more tradable manufacturing or agricultural activities.

In Figure 14 we plot the R-squared values for each occupation falling into a particular 2-digit classification and in Figure 15 we show a histogram of the entire distribution of R-squared values across occupations in 2010.

Selecting an appropriate cut-off (in this case, the appropriate R-squared value) to distinguish between a traded and non-traded activity is always relatively ad-hoc (Jensen and Kletzer, 2010). As shown in the histogram in Figure 15, a significant proportion of occupations tend to have a R-squared values greater than 0.8, with a long tail of occupations having lower R-squared values. But it is difficult to justify a threshold of 0.8 over 0.7 or 0.9.

The key purpose of this exercise is to determine whether the SOC indices perform better (in terms of more accurately distinguishing between rich and poor places) when non-traded activities are removed. As such, we experiment with three different thresholds (0.7, 0.8 and 0.9) and compare them to the original dataset, where all occupations are included. Specifically, we construct three modified W matrices of binary state occupational employment concentrations in which occupations which have an R-squared greater than the given threshold are omitted. We then calculate new SOC indices on the basis of each modified W matrix and examine the correlation between the resulting SOC index and log income per capita across states.

Figure 16 compares the Pearson R correlation coefficient for the relationship between state income per capita and the new SOC indices calculated for each threshold. We find that the correlation coefficient is highest for the SOC index calculated on the basis of *all* occupations, and declines as we remove additional occupations in accordance with the given thresholds. This indicates that the SOC index is better able to distinguish between rich and poor places when non-tradable occupations are considered.

R2	Occupation	2 digit classification
0.99	First-Line Supervisors of Sales Workers	Sales and related
0.99	Cashiers	Sales and related
0.99	First-Line Supervisors of Office and Administrative Support Workers	Office and administrative support
0.99	General and Operations Managers	Management, business, science and art
0.99	Bookkeeping, Accounting, and Auditing Clerks	Office and administrative support
0.99	Receptionists and Information Clerks	Office and administrative support
0.99	Stock Clerks and Order Fillers	Office and administrative support
0.99	Sales Representatives, Wholesale and Manufacturing	Sales and related
0.99	Food Service and Lodging Managers	Management, business, science and art
0.99	Retail Salespersons	Sales and related
0.99	Automotive Service Technicians and Mechanics	Installation, maintenance and repair
0.99	Office and administrative support workers, nec	Office and administrative support
0.98	Hairdressers, Hairstylists, and Cosmetologists	Personal care and service
0.98	Education Administrators	Management, business, science and art
0.98	Electricians	Construction and extraction
0.98	File Clerks	Office and administrative support
0.98	Managers, nec (including Postmasters)	Business operations specialist
0.98	Accountants and Auditors	Financial specialists
0.98	Transportation, Storage, and Distribution Managers	Management, business, science and art
0.98	Data Entry Keyers	Office and administrative support
…	…	…
0.17	Tool and Die Makers	Production
0.16	Tool Grinders, Filers, and Sharpeners	Production
0.16	Riggers	Installation, maintenance and repair
0.16	Derrick, rotary drill, and service unit operators, and roustabouts, oil, g	Construction and extraction
0.15	Gaming Services Workers	Personal care and service
0.15	Atmospheric and Space Scientists	Life, physical and social science
0.14	Metal Furnace Operators, Tenders, Pourers, and Casters	Production
0.14	Gaming Managers	Management, business, science and art
0.14	Economists and market researchers	Life, physical and social science
0.13	Railroad Brake, Signal, and Switch Operators	Transportation
0.13	Fishing and hunting workers	Farming, fishing and forestry
0.11	Forging Machine Setters, Operators, and Tenders, Metal and Plastic	Production
0.08	Model Makers and Patternmakers, Metal and Plastic	Production
0.08	Rolling Machine Setters, Operators, and Tenders, metal and Plastic	Production
0.07	Mining Machine Operators	Construction and extraction
0.06	Gaming Cage Workers	Office and administrative support
0.04	Textile Knitting and Weaving Machine Setters, Operators, and Tenders	Production
0.03	Tire Builders	Production
0.02	Textile Winding, Twisting, and Drawing Out Machine Setters, Operato	Production
0.01	Explosives Workers, Ordnance Handling Experts, and Blasters	Construction and extraction

Table 6: Top and bottom 20 occupations ranked in terms of the R-squared for the relationship between employment in an occupation and total population across states for the year 2010

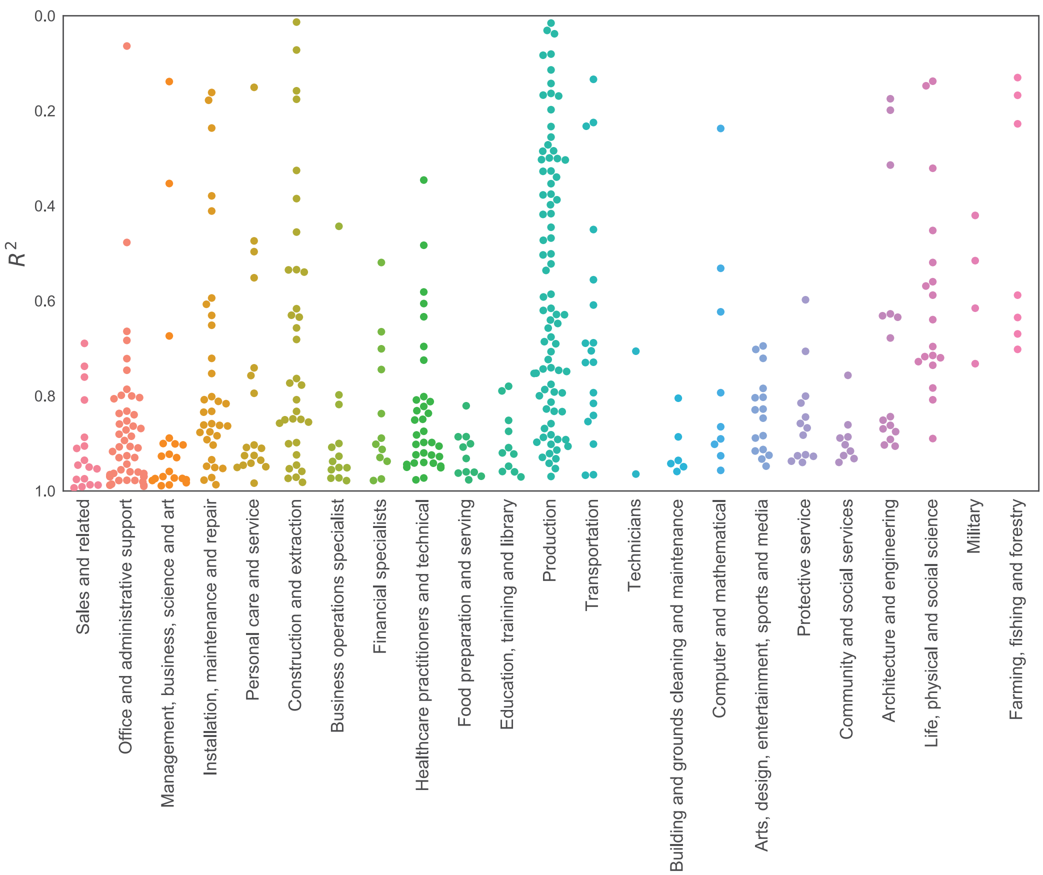

Figure 14: Distribution of R squared values for each detailed occupation within broader 2-digit classifications

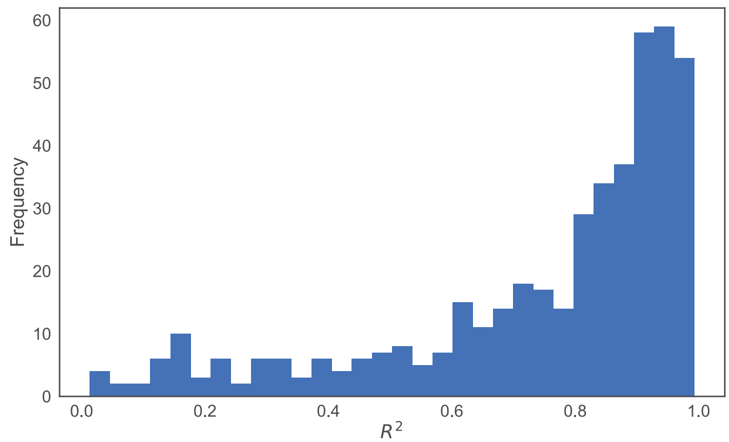

Figure 15: Histogram showing the distribution of R-squared values across occupations for 2010.

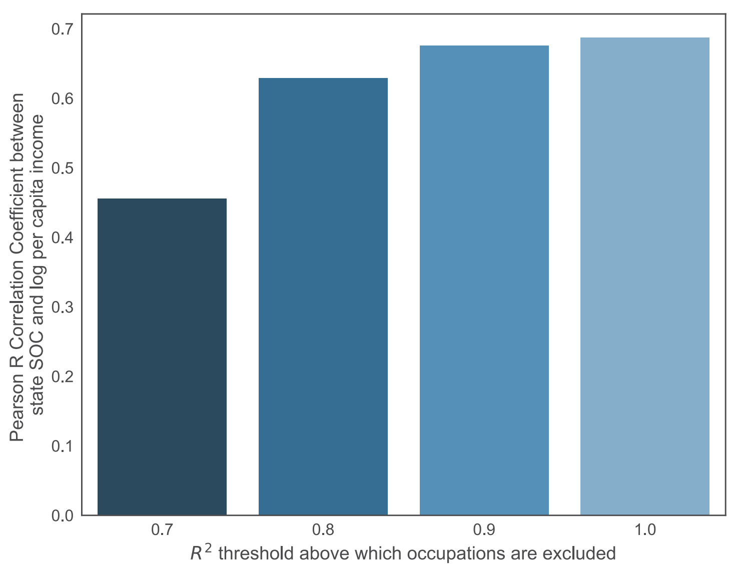

Figure 16: Comparison of the Pearson R correlation coefficient for the relationship between state income per capita and new SOC indices calculated for each threshold.

A.7 Comparisons to the Krugman Index

Some studies have utilised the Krugman Index of 'dissimilarity' (Krugman, 1991, 1993) to examine regional convergence in the US (Kim, 1998) and the UK (Martin et al., 2018). This index quantifies the difference in economic structure between one region s and another region s^* by summing up the differences in their employment shares across sectors and is given by

$$KI_{ss^*} = \sum_i |s_i - s_i^*|, \tag{15}$$

where s_i represents the employment share of region s in sector i.

While this would generate $N \times N$ pairwise comparisons for N regions, the index can also be calculated by comparing each region's employment share s_i to the national average employment share in sector i. In Figure 17, we apply both versions of the Krugman Index to state employment shares over the 1850-2010 period. The orange line shows the average of pair-wise state dissimilarities calculated for each year, while the blue line shows the average of state-dissimilarities to the national average occupational structure. With the exception of a peak in dissimilarity in 2000, the measures have shown a declining trend since 1870, suggesting that the distribution of state employment shares have becoming more similar over time.

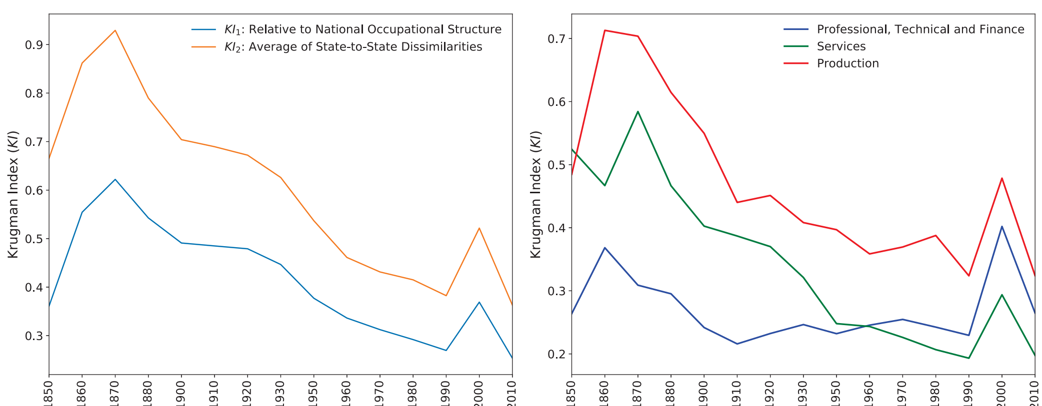

Figure 17: Panel A: Krugman Index for US state occupational employment shares. Panel B: Krugman Index for occupational broad categories.

In Panel B, we calculate the Krugman Index (based on state comparisons to the national average employment share) for the same broad occupational categories shown in Figure 3.

While the same peak occurs in 2000, the trends suggest that states are more dissimilar in relation to production-related occupations, followed by high-skill service professions followed by other services. The decreasing trend in production-related occupations has slowed over time, while the trend for high-skill service professions has been slowly increasing.

While the Krugman Index provides a useful indication of the convergence in distributions of state employment shares, we suggest that the metrics described in this paper could provide an additional complementary lens on the changing nature of regional employment concentrations.

References

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, Volume 4, pp. 1043–1171. Elsevier.

Alder, S., D. Lagakos, and L. Ohanian (2014). The decline of the us rust belt: a macroeconomic analysis.

Allcott, H. and D. Keniston (2017). Dutch disease or agglomeration? the local economic effects of natural resource booms in modern america. *The Review of Economic Studies 85*(2), 695–731.

Attaran, M. (1986). Industrial diversity and economic performance in us areas. *The Annals of Regional Science 20*(2), 44–54.

Autor, D. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives 29*(3), 3–30.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics 118*(4), 1279–1333.

Barro, R. J. and X. Sala-i-Martin (1992). Convergence. *Journal of Political Economy 100*(2), 223–251.

Berger, T. and C. B. Frey (2016). Did the computer revolution shift the fortunes of us cities? technology shocks and the geography of new jobs. *Regional Science and Urban Economics 57*, 38–45.

Berry, C. R. and E. L. Glaeser (2005). The divergence of human capital levels across cities. *Papers in Regional Science 84*(3), 407–444.

Black, D., T. McKinnish, and S. Sanders (2005). The economic impact of the coal boom and bust. *The Economic Journal 115*(503), 449–476.

Bresnahan, T. F. and V. A. Ramey (1993). Segment shifts and capacity utilization in the us automobile industry. *The American Economic Review 83*(2), 213–218.

Carrillo, J. (2004). NAFTA: the process of regional integration of motor vehicle production. In *Cars, Carriers of Regionalism?*, pp. 104–117. Springer.

Duranton, G. and D. Puga (2000). Diversity and specialisation in cities: why, where and when does it matter? *Urban Studies 37*(3), 533–555.

Duranton, G. and D. Puga (2005). From sectoral to functional urban specialisation. *Journal of Urban Economics 57*(2), 343–370.

Easterlin, R. (1960). Interregional differences in per capita income, population, and total income, 1840-1950. In *Trends in the American economy in the nineteenth century*, pp. 73–140. Princeton University Press.

Gaspar, J. and E. L. Glaeser (1998). Information technology and the future of cities. *Journal of Urban Economics 43*(1), 136–156.

Glaeser, E. L. and G. A. Ponzetto (2007). Did the death of distance hurt detroit and help new york? Technical report, National Bureau of Economic Research.

Glaeser, E. L., G. A. Ponzetto, and K. Tobio (2014). Cities, skills and regional change. *Regional Studies 48*(1), 7–43.

Glaeser, E. L. and A. Saiz (2003). The rise of the skilled city. Technical report, National Bureau of Economic Research.

Hausmann, R., C. A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, and M. A. Yildirim (2014). *The Atlas of Economic Complexity: Mapping paths to prosperity.* MIT Press.

Hausmann, R., J. Hwang, and D. Rodrik (2007). What you export matters. *Journal of Economic Growth 12*(1), 1–25.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences 106*(26), 10570–10575.

Jensen, J. B. and L. G. Kletzer (2010). Measuring tradable services and the task content of offshorable services jobs. In *Labor in the new economy*, pp. 309–335. University of Chicago Press.

Kemeny, T. and M. Storper (2015). Is specialization good for regional economic development? *Regional Studies 49*(6), 1003–1018.

Kemp-Benedict, E. (2014). An interpretation and critique of the method of reflections. Mimeo.

Kim, S. (1998). Economic integration and convergence: US regions, 1840–1987. *The Journal of Economic History 58*(3), 659–683.

Klein, A. (2013). New state-level estimates of personal income in the united states, 1880–1910. In *Research in Economic History*, pp. 191–255. Emerald Group Publishing Limited.

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

Krugman, P. R. (1993). *Lessons of Massachusetts for EMU.* in Torres F., Giavazzi

F. (eds) *Adjustment and growth in the European Monetary Union*, Cambridge University Press.

Lall, S. (1992). Technological capabilities and industrialization. *World Development 20*(2), 165–186.

Lall, S. (2000). The technological structure and performance of developing country manufactured exports, 1985-98. *Oxford development studies 28*(3), 337–369.

Lall, S., J. Weiss, and J. Zhang (2006). The "sophistication" of exports: a new trade measure. *World Development 34*(2), 222–237.

Landefeld, J. and R. Marcus (2006). Gross domestic product by state estimation methodology, survey of current business. Technical report, Bureau of Economic Analysis. US Department of Commerce.

Leamer, E. E. and M. Storper (2014). The economic geography of the internet age. In *Location of International Business Activities*, pp. 63–93. Springer.

Lee, K. and S. Ni (2002). On the dynamic effects of oil price shocks: a study using industry level data. *Journal of Monetary Economics 49*(4), 823–852.

Martin, R., P. Sunley, B. Gardiner, E. Evenhuis, and P. Tyler (2018). The city dimension of the productivity growth puzzle: the relative role of structural change and within-sector slowdown. *Journal of Economic Geography*.

Mealy, P., J. D. Farmer, and A. Teytelboym (2017). A new interpretation of the economic complexity index. *INET Oxford Working Paper*.

Michaels, G. (2008). The effect of trade on the demand for skill: Evidence from the interstate highway system. *The Review of Economics and Statistics 90*(4), 683–701.

Michaels, G., F. Rauch, and S. J. Redding (2013). Task specialization in us cities from 1880-2000. *National Bureau of Economic Research*.

Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.

Porter, M. (2003). The economic performance of regions. *Regional Studies 37*(6), 549–578.

Quigley, J. M. (1998). Urban diversity and economic growth. *Journal of Economic Perspectives 12*(2), 127–138.

Rondander, C. (1999). The classification of work: Applying 1950 census occupation and industry codes to 1920 responses. *Historical Methods 32*(3), 151–155.

Rossi-Hansberg, E., P.-D. Sarte, and R. Owens Iii (2009). Firm fragmentation and urban patterns. *International Economic Review 50*(1), 143–186.

Ruggles, S., K. Grenadek, R. Goeken, J. Grover, and M. Sobek (2017). Integrated public use microdata series: Version 7.0.

Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence 22*(8), 888–905.

Shutters, S. T., R. Muneepeerakul, and J. Lobo (2016). Constrained pathways to a creative urban economy. *Urban Studies 53*(16), 3439–3454.

Smith, A. (1776). The wealth of nations. *Oxford University Press, Oxford*.

Spence, M. and S. Hlatshwayo (2012). The evolving structure of the american economy and the employment challenge. *Comparative Economic Studies 54*(4), 703–738.

Storper, M. (2018). Separate worlds? explaining the current wave of regional economic polarization. *Journal of Economic Geography 18*(2), 247–270.

Sutton, J. and D. Trefler (2016). Capabilities, wealth, and trade. *Journal of Political Economy 124*(3), 826–878.

US Department of Commerce (1962). Personal income by states, a supplement to the survey of current business and survey of current business. Technical report, Office of Business Economics.

Youn, H., L. M. Bettencourt, J. Lobo, D. Strumsky, H. Samaniego, and G. B. West (2016). Scaling and universality in urban economic diversification. *Journal of The Royal Society Interface 13*(114), 1–7.