COMPLEXITY: A GENERALIZED APPROACH IN STATA FOR SPECIALIZATION COMPLEXITY INDICES

Documents de travail GREDEG
GREDEG Working Papers Series

CHARLIE JOYEZ

GREDEG WP No. 2025-50

https://ideas.repec.org/s/gre/wpaper.html

COMPLEXITY: A Generalized Approach in Stata for Specialization Complexity Indices

Charlie Joyez [1]
Université Côte d'Azur, CNRS, GREDEG, France
charlie.joyez@univ-cotedazur.fr

August 2025

Abstract

We introduce `complexity`, a `Stata` command available on SSC that computes generalized complexity indices for specialization matrices. Originally developed for assessing economic complexity with global trade data (Hidalgo and Hausmann, 2009), these metrics have since been extended to various domains including regional development, innovation, and labor economics. The `complexity` command implements three core methodologies: the eigenvector method (Hausmann et al., 2011a), the Method of Reflection (Hidalgo and Hausmann, 2009), and the fitness-complexity approach (Tacchella et al., 2012). It also computes relatedness metrics such as coherence, adjacency matrices of the activity space network, and the complexity outlook as a measure of complexity potential. We describe the syntax and options, review the underlying algorithms, and provide applied examples.

1 Introduction

1.1 What is "complexity"?

Economic complexity aims to measure the amount of productive knowledge embedded in a system based on its output structure. The field was pioneered by Hidalgo and Hausmann (2009), who introduced the Economic Complexity Index (ECI) and Product Complexity Index (PCI). These measures leverage the diversity and ubiquity of a country's exports to infer its underlying capabilities. This approach reframed economic development as a path-dependent process of knowledge accumulation, which could be mapped using trade data.

A major advance in this framework was the concept of the Product Space, introduced by Hidalgo et al. (2007), which represents the network of products based on their relatedness. Products are more likely to be co-exported if they require similar capabilities. This network perspective reveals that complex products cluster in a densely connected core of the Product Space, while simple products lie on the periphery. The structure of a country's exports within this space strongly predicts its opportunities for diversification and upgrading.

The Atlas of Economic Complexity (Hausmann et al., 2011a) operationalized these ideas into a widely used tool for policy analysis and academic research. Later, the Complexity Outlook

Index (Hausmann et al., 2014) was developed to assess not just current complexity but also future potential, based on proximity to more complex products.

Subsequent literature has extended complexity theory to subnational units. Balland et al. (2020) demonstrated that complex economic activities are disproportionately concentrated in large urban areas, and that regions diversify through related activities. Based on employment data, (Fritz and Manduca, 2021) show that more complex cities in the US display in average higher wages. In their study of smart specialization, Balland et al. (2019) also cautioned against increasing regional inequality, as less developed areas tend to specialize in low-complexity activities while advanced regions accumulate high-complexity ones. This highlights the importance of path dependence and local related capabilities in regional transformation. Besides economic geography and regional studies, the concepts of complexity and relatedness appear to be relevant in many other fields including labor economics, innovation studies, and development as review notably Hidalgo (2021); Hidalgo and Stojkoski (2025) or with a more critical look Pinheiro (2025)

This literature has also densified with the development of alternative metrics. Notably, (Tacchella et al., 2012) proposed the fitness-complexity approach as an alternative to symmetric eigenvector-based measures. Their iterative, non-linear method avoids numerical instabilities and reflects asymmetries between country and product capabilities. Fitness was also shown to better predict future outcomes.

1.2 Why the need for a package?

While the theoretical and empirical literature on economic complexity has expanded rapidly, its methodological diffusion has been hindered by limited access to standardized, transparent, and replicable tools. Most available implementations are embedded in web platforms or proprietary codebases, which restrict flexibility and hinder reproducibility[1]. By offering a native Stata command, `complexity` addresses these gaps, making economic complexity metrics accessible to a broad range of applied researchers. This tool is particularly valuable for economists, policy analysts, and social scientists accustomed to working within the Stata ecosystem, and who require seamless integration with survey, administrative, or microdata workflows. Although first released on SSC in 2019, `complexity` package has progressively offered easier and wider estimation of specialization complexity indices, that now deserved a proper presentation.

Moreover, by generalizing complexity measures to any bipartite specialization matrices - whether countries by products, regions by technologies, or workers by occupations - `complexity` facilitates the integration of complexity science into diverse empirical settings. It opens the door to new applications in regional planning, labor market design, innovation policy, and comparative political economy.

Building on these theoretical foundations and practical needs, the `complexity` command brings the tools of economic complexity into `Stata`. It allows researchers to:

[1]One notable exception is the release of a R package `economiccomplexity` by Mauricio Vargas Sepulveda on CRAN on December 2024

- Compute complexity indices using the eigenvector, Method of Reflection, and fitness-complexity methodologies
- Generate additional indices: Diversity, Ubiquity, Coherence, Complexity Outlook Index and Target scores.
- Generate activity and individual relatedness network.

The command accepts various forms of syntaxes, inputs and output options, that we detail in the next section.

2 Syntax and Options

2.1 Instalation and requirements

Available on SSC, installing (or upgrading) the `complexity` package can be done directly in Stata via the command:

```
ssc install complexity,replace
```

`complexity` runs on any Stata version over 11.0.

2.2 syntaxes

Two alternative syntaxes of `complexity` exist:

The first one when the specialization matrix is expressed as a rectangular block of variables :

```
complexity  , varlist()  [method() coherence outlook diversity
ubiquity rca iterations() projection() transpose xvar]
```

Where *varlist* is required, options between brackets aren't required and are detailed below.

The second syntax is set for when the specialization matrix is not in list of Stata variables but directly in a matrix form, either in Stata or Mata.

```
complexity  , matrix() matsource()  [method() coherence outlook diversity
ubiquity rca iterations() projection() transpose xvar]
```

Where `matrix()` corresponds to the matrix name, and matsource indicates whether it is stored in Stata or Mata format. Under this syntax, `matrix()` is required, and `matsource()` is assumed to be Mata if not specified.

2.3 Options

For any of those two syntaxes, a common list of options can be added; for clarity, we differentiate them into two categories

- Methods and parameters

`method()` Specifies the method followed to compute complexity. Default is *eigenvector*, alternatives are *mr* or *fitness*. See section 3 for methodological differences.

`iterations()` Sets the maximum number of iteration for MR or fitness method. Must be of even order. `projection()` Indicates which complexity index to return. *indiv* would return the individuals' complexity (e.g. countries ECI), while *activities* returns the activities' complexity (e.g. product PCI). In any case both are computed. If none are indicated, individuals' complexity is returned by default.

`transpose` Transposes initial matrix if individuals were in columns and activities in rows.

`rca` To be added if the specialization matrix is already expressed as comparative advantages (binary or not) following Balassa (1965) index contrasting the share of an activity j for an individual i to its share in the whole population $RCA_{ij} = \frac{X_{ij}/\sum_i X_{ij}}{\sum_j X_{ij}/\sum_i \sum_j X_{ij}}$. The default code transforms the specialization matrices into RCA matrices. This step is ignored if `rca` option is specified.

`xvar` Doesn't return any Stata variables (only Stata and mata matrices are stored).

- Additional Metrics

`coherence` Computes the coherence of the specialization pattern for each individual (total proximity of specialization area)

`outlook` Indicates the Complexity Outlook Index (average complexity of neighboring activities), and targets ranked by complexity gains/accessibility.

`diversity`: Returns the number of activities of specialization for each individual (with $\sum_j RCA_{ij} \geq 1$). Sometimes useful to normalize *coherence* score (average instead of total).

`ubiquity`: Returns the number of individuals with $RCA \geq 1$ for each activity

3 Methods and Implementation

The `complexity` command computes specialization complexity indices based on a binary or weighted bipartite matrix $M \in \mathbb{R}^{N\times P}$, where rows represent entities (e.g., countries, regions, individuals) and columns represent activities (e.g., products, occupations, technologies). The command implements three primary methodologies: the Method of Reflections (MR), the eigenvector method, and the fitness-complexity algorithm. Each method provides a distinct approach to quantifying the complexity inherent in the specialization patterns of entities.

3.1 Method of Reflections (MR)

Introduced by Hidalgo and Hausmann (2009), the MR is an iterative procedure that alternately computes the average properties of the nodes connected in the bipartite network. The initial step defines the diversification of an entity and the ubiquity of an activity:

$$k_i^{(0)} = \sum_j M_{ij}, \quad k_j^{(0)} = \sum_i M_{ij}$$

Subsequent iterations refine these measures:

$$k_i^{(n)} = \frac{1}{k_i^{(0)}} \sum_j M_{ij} k_j^{(n-1)}, \quad k_j^{(n)} = \frac{1}{k_j^{(0)}} \sum_i M_{ij} k_i^{(n-1)}$$

The MR captures the intuition that entities are complex if they are connected to complex activities, and activities are complex if they are connected to complex entities. The process continues for a predetermined number of iterations, typically until convergence or a set maximum number of steps. In their seminal paper, Hidalgo and Hausmann indicate that the iteration process should continue until the ranking of countries was stable, which in their studies happens at $n = 18$. However, most of the adjustment is done in the first iterations, so that a rule of thumb is that iterations higher than 20 are virtually useless. This is why we set a default number of iteration to 20, which can be modified through the `iter()` option. The number indicated in `iter()` must be of even order, as it corresponds to the complexity of individuals. For complexity of activities, we would therefore use the value in $k_j^{(n-1)}$. In any cases the process stops when the ranking of individuals is stable across two iterations. The output of `complexity` when using `method(mr)` option indicates whether the iteration process yield convergence and at which round. In the reverse case, when the maximum number of iterations is reached before convergence, the following warning displays `"note : MR's optimal iteration is of higher order than specified"`, but it doesn't prevent the computation of *Complexity_i*

3.2 Eigenvector Method

The absence of a natural convergence point (the higher the number of iterations, the more each complexity score converges toward the same value), has been pointed out as a major limitation to the Method of Reflections (MR). Indeed each iteration reflects a new, non-nested transformation of the input data, and only even iterations are interpretable for a given node set. As a result, the choice of iteration depth becomes arbitrary and affects comparability across studies. To overcome these issues, subsequent research has reformulated the MR approach using tools from spectral graph theory. Specifically, Hausmann et al. (2014) propose extracting the second eigenvector of the symmetrized co-occurrence matrix, which captures the primary axis of variation in the underlying capability space. This eigenvector-based method yields a unique, interpretable ranking of entities by complexity and has since become standard in applied work, including the Atlas of Economic Complexity. In the `complexity` command, this method is implemented through the `method(eigen)` option, providing users with a robust and theoretically grounded alternative to the iterative MR procedure.

Specifically, the eigenvector method, as utilized in the Atlas of Economic Complexity (Hausmann et al., 2011), reformulates the MR approach by focusing on the spectral properties of the matrix derived from M. Specifically, it involves constructing the matrix $\tilde{M} = D^{-1}M$, where D is a diagonal matrix of the degrees (diversification or ubiquity), and then computing the eigenvectors of the resulting matrix.

For entities, the complexity scores are obtained from the eigenvector corresponding to the second largest eigenvalue of the matrix $\tilde{M}\tilde{M}^T$, while for activities, the eigenvector is derived from $\tilde{M}^T\tilde{M}$. This method stabilizes the MR by capturing the principal components that explain the variance in the specialization patterns.

The Eigenvector method is supposed to be more efficient than the method of reflection by circumventing the iteration process. Yet, the definition of an eigenbasis is ambiguous of sign. If x is an eigenvector such that $Ax = \lambda x$, so is $-x$, as $A(-x) = -Ax = -\lambda x = \lambda(-x)$. However, here the sign of the eigenvector entries is of crucial importance as it reverses the complexity scores of each individual or activity. To identify which vector to keep between x and $-x$ we call the MR method and compute a correlation coefficient between the score of the two methods. When the correlation coefficient is negative, we take the opposite eigenvector that the one computed [2]. Also, when the $\tilde{M}$ and its transposed are large and filled with small values the mata command `eigensystemselecti` may return missing values in the eigenvector. To overcome this issue, we inflate the matrix by a common factor if missing values are detected in the second eigenvector.

3.3 Fitness-Complexity Algorithm

An alternative to both the Method of Reflections and the eigenvector approach is the fitness–complexity algorithm proposed by Tacchella et al. (2012). This method addresses what the authors identify as a key conceptual shortcoming of previous techniques: their reliance on linear averaging, which may dilute critical asymmetries in the data. The fitness method introduces a nonlinear iterative procedure that computes country "fitness" and product "complexity" based on a mutually reinforcing relationship: the fitness of a country increases with the number and difficulty of the products it exports, while a product's complexity is determined by the inverse of the fitnesses of the least-fit countries that export it. This asymmetry is intended to capture the idea that producing complex products is informative only when done by already capable (i.e., high-fitness) countries. The fitness algorithm is empirically found to be more robust to outliers, more predictive of economic growth, and better suited for datasets with heterogeneous diversification patterns, such as those with rentier or highly specialized economies. Within the `complexity` command, the `method("fitness")` option implements this non-linear algorithm, providing users with an alternative complexity metric that aligns more closely with capability-based views of development and innovation potential.

Specifically, the fitness-complexity algorithm addresses the limitations of linear methods by introducing a non-linear iterative approach. The algorithm defines the fitness F_i of an entity and the complexity Q_j of an activity through the following recursive relations:

$$F_i^{(0)} = 1 \quad \forall i$$
$$Q_j^{(n)} = \sum_i M_{ij} F_i^{(n-1)}$$

[2]This check is also done in the equilavent R package `economiccomplexity` in "how to use this package" vignette: "The eigenvalues also calls the reflections methods in order to correct the index sign in some special cases when the correlation between the output from both methods is negative."

$$F_i^{(n)} = \sum_j M_{ij} \frac{1}{Q_j^{(n-1)}}$$

After each iteration, normalization is applied to ensure numerical stability. The process continues until convergence (ranking stable over two iterations) or a specified number of iterations in option `iter()`. This algorithm captures the non-linear interplay between the capabilities of entities and the complexity of activities. One advantage of the fitness method is to yield an index whose distribution doesn't follow a normal distribution as the one of the ECI, but rather a power law distribution, more in line with the stylized facts on unequal distributions of economic performances.

3.4 Additional Outputs

In addition to complexity scores, the command can compute individuals *coherence*, or the average proximity of activities they are specialized on; and *complexity outlook index* (COI), capturing the potential of future complexity.

Both measures rely on the concept of proximity or relatedness between activities, to capture the proximity between current specializations (for coherence) or the complexity of closely related activites not yet developed (for COI). The relatedness matrix ϕ captures the empirical association between pairs of features—typically products, technologies, or occupations—based on their co-occurrence across entities such as countries or regions. This concept, central to the product space framework introduced by Hidalgo et al. (2007), provides the structural foundation for more advanced measures like complexity, density, and coherence. The rational is simple: the more often two activities are done by the same individual, the more likely they require skills. The derived "principle of relatedness" states that the probability to enter into a new activity increases with this frequency of co-occurence (Boschma et al., 2015; Hidalgo et al., 2018).

Formally, the relatedness is defined as follows. Let M_{ia} be a binary incidence matrix, where $M_{ia} = 1$ if individual i is specialized in activity a, and 0 otherwise. The relatedness between two features a and a' is defined as:

$$\phi_{aa'} = \frac{\sum_i M_{ia} \cdot M_{ia'}}{\max\left(\sum_i M_{ia}, \sum_i M_{ia'}\right)}.$$

This measure, also known as *proximity*, captures the conditional probability that a country specialized in one feature is also specialized in another. The normalization by the maximum ubiquity ensures that $\phi_{aa'}$ is symmetric and bounded between 0 and 1. A value of $\phi_{aa'} = 1$ indicates that all countries specialized in one feature are also specialized in the other.

The complete matrix $\Phi = [\phi_{aa'}]$ forms the empirical basis of the *feature space* (e.g., product space, technology space), a network where nodes represent features and links denote significant empirical associations. This network can be used to project diversification trajectories, identify coherent specialization patterns. In the `complexity` command, this matrix is computed when the `coherence` or `outlook` options are specified, and stored under a Mata matrix $Prox\_a$ which may be used directly or exported for further analysis or network visualization. Similarly, the individual-space network is stored under the adjacency matrix $Prox\_i$.

Coherence is a measure of internal relatedness of the activities an individual is already specialized in. Rather than asking how many products a region produces or how complex they are, coherence asks whether those products rely on overlapping capabilities—whether they form a consistent and integrated specialization pattern. This concept is similar to the notion of industrial and technological *cohesion* used by Neffke et al. (2011). The term *coherence* goes back to Teece et al. (1994) for corporate activities. It recently grew in popularity in contributions extending the coherence framework beyond the technological space notably to include occupations. As such, coherence has become a key indicator for understanding the capacity of regions to sustain innovation, absorb shocks, and transition into more complex activities (Henning et al., 2025; Otto et al., 2025; Pugliese et al., 2019).

Formally, given a binary incidence matrix M_{ia} indicating whether individual i is specialized in activity (or product, technology, occupation, etc.) a, and a proximity matrix $\phi_{aa'}$ capturing empirical relatedness between products p and p', the coherence of i is defined as:

$$\text{Coherence}_i = \sum_{a \in \mathcal{A}_i} \sum_{a' \in \mathcal{A}_i} \phi_{aa'},$$

where $\mathcal{A}_i = \{a : M_{ia} = 1\}$ denotes the set of activities in which i is specialized. This expression sums all pairwise relatedness values between the activities that i is currently specialized in, capturing the extent to which its productive structure is concentrated in related domains. Coherence has been used extensively in empirical studies to explain the geography of innovation, knowledge recombination, and labor market resilience (Balland et al. 2019; Neffke et al. 2017; Diodato and Weterings 2015).

Although the coherence index is often used in its raw form, it is sometimes normalized by the number of product pairs, $|\mathcal{A}_i| \cdot (|\mathcal{A}_i| - 1)$, to facilitate comparisons across entities with different levels of diversity. High coherence implies that a country's current capabilities are mutually supportive, making it more likely to adapt to shocks or transition into nearby, complex activities. In this way, coherence extends the economic complexity paradigm by emphasizing the internal structure of capabilities, beyond their sheer number or average complexity.

In turn, the *Complexity Outlook Index Complexity Outlook Index* (COI), introduced by Hausmann et al. (2014) in the second edition of the *Atlas of Economic Complexity*, combines relatedness and complexity by estimating the expected gain from transitioning into nearby features.

Formally, for country c, the COI is defined as:

$$\text{COI}_{ia'} = \sum_{a \in \mathcal{A}_i} \phi_{aa'} \cdot \text{Complexity}_{a'},$$

High COI values suggest that a country is well-positioned to diversify into sophisticated activities in the near future, making the COI particularly useful for policymaking and development strategy. Research on smart specialization and regional dynamics develop similar complexity forecast, but lack of a common index COI could provide (Balland et al., 2019; O'Clery et al., 2021).

4 Output and Stored results

4.1 Output

On window *Results*, `complexity` displays the summarize of the newly computed Complexity index and additional variables computed if any, informing the user about the name of the newly generated variables, the number of observations, and that it was computed without issues. According the the options specified, some information or error messages can be returned too when options are conflicting notably. As for the data, `complexity` returns by default the initial dataset, with a new Stata variable corresponding to the complexity score for all individuals (in rows). Yet, the variables are stored into different names for each `method()` used, in order to ease the comparison of results between them. Specifically, the (default) eigenvector method returns a variable `Complexity_i`. The method of reflection returns a variable called `MR_Complexity_i`, and the `fitness` method returns a variable *fitness_i*. `complexity` would delete any similarly named variables, anytime it needs to create a new one.

If option `projection(activities)` is specified, the same three variables can be created but referred with a $_a$ ending. Each of them will have P observations. If $P < N$ this leads to missing values in the newly created variable. If $P > N$, the existing dataset is expanded to P observations, resulting in missing values in the previous data (defined at the individual level with N observations). The activity complexity score is returned in the same order as activities are listed in the specialization matrix. To ease the correspondence between the activity and its complexity score, another variable *actlist* is created when the command is called through the *varlist()* syntax.

Similarly, *coherence_i* and *coherence_a* variables are created in addition to the complexity variable when `coherence` option is specified.

When the `outlook` option is specified, a first variable *coi* is created. This Complexity Outlook Index is always only computed for individuals and not activities, whatever the projection of the complexity, and therefore the new variable is always of dimension $Nx1$. The reason for it is that we believe it is individual that specialize in activities and not otherwise, and the interpretation of a complexity outlook for an activity would be hard to interpret. In addition to the *coi* variable, P new variables are created indicating the desirable target activities to improve individuals' complexity as a product of their proximity in terms of skills and their inherent complexity.

4.2 Stored results

In addition to the variables created in Stata, `complexity` stores some information directly as scalars, vectors or matrices. Specifically, because complexity is always computed for the two dimensions (activities and individuals) simultaneously, `complexity` saves the following in `r()` : *r(Complexity_individual)*, a column vector of Individuals' complexity values. *r(Complexity_activity)*, a column vector of activities' complexity values. *r(Diversity)*, a column vector of Individuals' diversity , and *r(Ubiquity)*, a column vector of activities' ubiquity.

The same vectors are stored in Mata, respectively under the names: *comp_i*, *comp_a*, *Diversity* and *Ubiquity*

if the *fitness* method is specified the vectors returned are *r(fitness_individual)* and *r(fitness_activity)* and respective mata vectors are *fitness_i* *fitness_a*

if the coherence or outlook option are specified two matrices are stored in mata. **Prox_a** and **Prox_i** which are adjacency matrices of the Proximity space of activities and individuals (cf the Product Space from Hidalgo et al. 2007)

5 Examples

5.1 Illustrative example

Let's start with a small sample of virtual data to show the flexibility of `complexity`.

```
clear
mata mata clear
*install complexity package
ssc install complexity,replace
* Let's generate a matrix of specialization for 7 individuals
* in 4 activities as Stata variables
    . set obs 7
    . gen a1=runiform()*10
    . gen a2=runiform()*10
    . gen a3=runiform()*10
    . gen a4=runiform()*10

*Let's store this as a bloc matrix in Mata
    . putmata M=(a*)

*Let's also store it as a Stata matrix
    . mkmat a*,matrix(S)

** Computation of Complexity scores.
* 1- Using varlist() syntax
*1.1 individual complexity (eigenvector method)
. complexity, varlist(a*)

*1.2 individual complexity (MR method) and coherence score
    . complexity, varlist(a*) method(mr) coherence diversity

    *1.3 activities complexity using fitness score and input
    *as a table of Revealed Comparative Advantages
. complexity, varlist(a*) proj(activities) method(fitness) rca ubiquity
```

A preview of the dataset in 1 shows what has been done. While the preamble created variables $a1$ to $a4$ over the 7 rows, the command 1.1 created the *Complexity_i* variable, command 1.2 the *MR_Complexity_i* variable, and both are highly correlated. The second command also created the *coherence_i*, and *diversity* variables. At last, the command 1.3 created the two four-observations variables *actlist* detailing the list of activities we have in our dataset, their *ubiquity* and *fitness*.

	a1	a2	a3	a4	Complexity_i	Diversity	MR_Complex~i	coherence_i	Ubiquity	actlist	fitness_a
1	8.069639	.3824018	2.199756	1.649701	1.686776	1	-1.020621	2.203537	7	a1	.8759755
2	2.568897	1.384412	1.788245	8.966501	-.9024724	1	-1.020621	1.4131861	6	a2	1.049532
3	6.935425	5.189824	3.6064	8.639234	-.9024724	1	-1.020621	2.3449059	6	a3	1.049532
4	2.03871	6.943715	.3673937	4.085774	-.6701295	2	.4082483	2.6743517	6	a4	1.049532
5	2.010436	5.327914	5.753793	1.949142	.1980733	2	1.837117	3.5484421	.		.
6	4.646199	7.187118	2.21964	7.840244	-.6701295	2	.4082483	2.6506715	.		.
7	8.933897	2.219616	7.979596	.70528	1.260355	2	.4082483	3.8892817	.		.

Figure 1: Data editor widow after the command 1.1 to 1.3

Let us detail a bit more of the syntax variations

```
* 2- Using Stata matrix as inputs
clear
complexity, mat(S) matsource(matrix) outlook

* 3- Using Mata matrix as inputs
    . complexity, matrix(M) method(fitness)
```

Gives us the following output (figure 2). All previous variables are deleted initially, but as the specialization matrix is store in matrices S (Stata) and M (Mata), we can directly call them. The command 2 computes the *Complexity_i* score following the eigenvector method (and thus yielding the same result as in command 1.1) from the stored matrix S. In addition the `outlook` option creates the *coi* variable, here indicating that individuals 4 and 5 which have a low complexity score, are the one with the best possible quick evolution. Variables *target* indicate that this increase will be easier by targeting a new specialization in activity 3.

The command 3 creates the fitness score for individuals, from the mata matrix M containing the bloc specialization matrix.

	Complexity_i	COI	target1	target2	target3	target4	fitness_i
1	1.686776	-.315561	0	.09060874	0	.01816879	.6533649
2	-.9024724	.4292233	0	.24704918	0	0	.795831
3	-.9024724	.4292233	0	.16208301	0	0	.795831
4	-.6701295	-.210374	0	0	0	0	1.325643
5	.1980733	-.18943718	0	0	0	.20703149	.9900673
6	-.6701295	-.210374	0	0	0	0	1.325643
7	1.260355	.28614888	0	.10863093	0	.06539364	1.113621

Figure 2: Data editor after commands 2 and 3

5.2 Real data example

Using detailed world trade flow from BACI database[3], I now show how to compute real complexity scores. Let's start with an initial trade dataset for the year 2023 with only three variables : country (*iso*3), product ($HS4d$, Harmonized System at 4 digits) and export value X_{ij}. With 195 countries and 1176 products, the total number of observation is 229,125. Here's a preview of the data:

	iso3	HS4d	Xij
1	AFG	101	.002
2	AFG	102	16.212
3	AFG	103	0
4	AFG	104	0
5	AFG	105	0
6	AFG	106	132.183
7	AFG	201	3
8	AFG	202	0
9	AFG	203	0
229116	ZMB	9615	2.184
229117	ZMB	9616	.801
229118	ZMB	9617	4.623
229119	ZMB	9618	1.247
229120	ZMB	9701	113.592
229121	ZMB	9702	19.406
229122	ZMB	9703	43.951
229123	ZMB	9704	0
229124	ZMB	9705	472.243
229125	ZMB	9706	.066

Figure 3: Preview of Trade Dataset

Then, in a two line command, we can easily obtain the complexity scores of each countries. First, let's express this as a NxP matrix (wide format).

```
reshape wide Xij, i(iso3) j(HS)
```

We know have expressed the data as a specialization matrix, with one line per countries and one column by product.

Then let's call the complexity command

```
complexity, varlist(Xij*)
```

Computes the complexity scores (in less than 4 seconds in Stata 17 MP).

A preview of the results can be seen through

```
gsort - Complexity
bro iso3 Complexity if _n<5 | _n>(_N-4)
```

	iso3	Xij101	Xij102	Xij103	Xij104
1	ABW	0	0	0	0
2	AFG	.002	16.212	0	0
3	AGO	3.744	79.235	20.187	251.338
4	ALB	.082	1.8	37.223	0
5	AND	105.354	.43	0	5.434
6	ARE	6879.097	2179.128	0	33534.61
7	ARG	21291.33	296.739	123.623	52.25
8	ARM	2.077	649.575	0	12021.97
9	ATG	0	2.222	0	9.399
186	USA	269128.1	421805.8	34821.13	4337.203
187	UZB	45.444	.132	0	87
188	VEN	0	0	0	0
189	VGB	11.343	0	0	.631
190	VNM	1.098	23603.39	111.393	0
191	VUT	0	0	0	0
192	YEM	0	0	0	0
193	ZAF	3006.604	33888.42	991.239	7410.296
194	ZMB	4.224	473.1	57.707	26.082
195	ZWE	462.579	22.156	21.905	0

Figure 4: Preview of Trade Data in wide format: Specialization Matrix

	iso3	Complexity_i
1	CHE	2.289096
2	JPN	2.273896
3	KOR	2.014745
4	DEU	1.971553
191	COG	-1.786568
192	COD	-2.067818
193	TCD	-2.64331
194	GIN	-2.710796

Figure 5: Preview of Economic Complexity Scores from Exports data.

These data and do-file are publicly available on the dedicated *Statalist* post: "Complexity package on SSC. Update and example files".

6 Potential Applications

The `complexity` command generalizes economic complexity analysis beyond the traditional country–product space, supporting analysis wherever specialization can be represented as a bipartite matrix. Below we outline key applications across research domains.

International Trade and Growth Projections

The original domain of complexity analysis remains critical. Export data at the country level can be used to compute ECI, PCI, and complexity outlooks. The tool mirrors the methodologies developed in the Atlas of Economic Complexity (Hausmann et al., 2011b, 2014). It also supports computation of the fitness index (Tacchella et al., 2012), increasingly used for robustness checks

[3] available on the CEPII website : https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37

or dynamic comparisons. New applications in international trade could also focus on complexity patterns of imports, or how the imports of complex goods can be used as a vector to develop new productive knowledge.

Regional Economic Development

Pinheiro et al. (2025) show that the diversification of regions into related activities is positive in terms of economic development, but also possibly contributes to regional inequalities. By applying `complexity` to region–industry or region–occupation matrices, researchers can analyze the uneven geography of knowledge accumulation. Balland and Boschma (2021) advocate for "smart specialization" strategies that build on regional capabilities—something the potential complexity measure can empirically support.

Innovation and Technological Diversification

Complexity scores help evaluate diversification paths in regional innovation systems. Balland and Rigby (2017) analyze patent data to show that related technological capabilities foster regional innovation. The `complexity` command can apply to region–patent class matrices to replicate such studies or support policy design in research funding and innovation promotion.

Labor Market Coherence and Resilience

Neffke et al. (2011) and Diodato et al. (2018) demonstrate that skill relatedness across occupations increases a region's capacity to absorb shocks. Alabdulkareem et al. (2018) and Joyez and Laffineur (2022) applied a network analysis and complexity thinking to labor market transitions and skill-relatedness between occupations. Using worker–occupation matrices or employment data by skill sets, the `complexity` command allows for measuring the coherence of local labor markets and estimating the complexity potential of occupational ecosystems.

These diverse applications demonstrate how `complexity` empowers researchers to explore structural transformation, industrial policy, innovation strategy, and labor market design—all from a shared foundation of economic complexity theory.

7 Concluding Remarks

The `complexity` command provides Stata users with a flexible and efficient tool for computing measures of economic complexity and related network-based metrics from any bipartite specialization matrix. It implements several widely used methods—including the Method of Reflections, the eigenvector-based approach, and the fitness–complexity algorithm—originally developed for country–product trade data but now increasingly applied to a variety of settings in regional development, innovation studies, and labor market analysis.

The command accommodates multiple input formats (Stata variables, Stata matrices, or Mata matrices) and includes options for projecting in any dimensions of the bipartite network, computing adjacency and relatedness matrices, and estimating potential complexity scores.

These features enable researchers to move beyond the canonical country–product framework and apply complexity analysis to data structures involving regions and technologies, firms and occupations, or other specialization-based relationships.

By making these techniques readily available in a native Stata environment, the `complexity` command lowers the barrier to entry for applied researchers interested in economic complexity and network-based metrics, and it supports the growing literature at the intersection of development economics, innovation, and spatial labor markets.

We hope the command will facilitate further empirical work and encourage extensions and applications across disciplines.

References

Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C. A., and Rahwan, I. (2018). Unpacking skill bias: Automation and new tasks. *Journal of the Royal Society Interface*, 15(139):20180276.

Balassa, B. (1965). Trade liberalisation and "Revealed" comparative advantage1. *The Manchester School*, 33(2):99–123. 02824.

Balland, P.-A. and Boschma, R. (2021). Complementary interregional linkages and smart specialisation: An empirical study on european regions. *Regional Studies*, 55(6):1059–1070.

Balland, P.-A., Boschma, R., Crespo, J., and Rigby, D. L. (2019). Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9):1252–1268.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P., Rigby, D. L., and Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, 4(3):248–254.

Balland, P.-A. and Rigby, D. (2017). The geography of complex knowledge. *Economic Geography*, 93(1):1–23.

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

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

Fritz, B. S. and Manduca, R. A. (2021). The economic complexity of us metropolitan areas. *Regional Studies*, 55(7):1299–1310.

Hausmann, R., Hidalgo, C., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., and Yildirim, M. (2011a). The atlas of economic complexity. *Boston. USA*. 00066.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A., and Yildirim, M. A. (2011b). *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. MIT Press.

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

Henning, M., Eriksson, R., Garefelt, P., Martin, H., and Elekes, Z. (2025). Job relatedness, local skill coherence and economic performance: a job postings approach. *Regional Studies, Regional Science*, 12(1):95–122.

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics*, pages 1–22.

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., et al. (2018). The principle of relatedness. In *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems 9*, pages 451–457. Springer.

Hidalgo, C. A. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575. 00178.

Hidalgo, C. A., Klinger, B., Barabasi, A.-L., and Hausmann, R. (2007). The product space conditions the development of nations. Paper 0708.2090, arXiv.org. 00468.

Hidalgo, C. A. and Stojkoski, V. (2025). The theory of economic complexity. *arXiv preprint arXiv:2506.18829*.

Joyez, C. and Laffineur, C. (2022). The occupation space: network structure, centrality and the potential of labor mobility in the french labor market. *Applied Network Science*, 7(1):16.

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

Otto, A., Losacker, S., and Hansmeier, H. (2025). Relatedness, complexity, and regional development paths in germany: a sequencing approach. *The Annals of Regional Science*, 74(2):50.

O'Clery, N., Yıldırım, M. A., and Hausmann, R. (2021). Productive ecosystems and the arrow of development. *Nature communications*, 12(1):1479.

Pinheiro, C. (2025). Relatedness and economic complexity as tools for industrial policy: Insights and limitations. *Structural Change and Economic Dynamics*, 72:1–10.

Pinheiro, F. L., Balland, P.-A., Boschma, R., and Hartmann, D. (2025). The dark side of the geography of innovation: relatedness, complexity and regional inequality in europe. *Regional Studies*, 59(1):2106362.

Pugliese, E., Napolitano, L., Zaccaria, A., and Pietronero, L. (2019). Coherent diversification in corporate technological portfolios. *PloS one*, 14(10):e0223403.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2(1):1–7.

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of economic behavior & organization*, 23(1):1–30.

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2025
GREDEG Working Papers Released in 2025

2025-18 Leonardo Ciambezi & Alessandro Pietropaoli
Relative Price Shocks and Inequality: Evidence from Italy

2025-19 Matteo Orlandini, Sebastiano Michele Zema, Mauro Napoletano & Giorgio Fagiolo
A Network Approach to Volatility Diffusion and Forecasting in Global Financial Markets

2025-20 Daniele Colombo & Francesco Toni
Understanding Gas Price Shocks: Elasticities, Volatility and Macroeconomic Transmission

2025-21 Simone Vannuccini
Move Fast and Integrate Things: The Making of a European Industrial Policy for Artificial Intelligence

2025-22 Patrice Bougette, Oliver Budzinski & Frédéric Marty
Revisiting Behavioral Merger Remedies in Turbulent Markets: A Framework for Dynamic Competition

2025-23 Amal Ben Khaled, Rami Haj Kacem & Nathalie Lazaric
Assessing the Role of Governance and Environmental Taxes in Driving Energy Transitions: Evidence from High-income Countries

2025-24 Sandye Gloria
Emergence. Another Look at the Mengerian Theory of Money

2025-25 Barbara Brixová, Anna Tykhonenko, Ľubica Štiblárová & Marianna Siničáková
Middle Income Convergence Trap Phenomenon in CEE Countries

2025-26 Grégory Donnat, Maxime Menuet, Alexandru Minea & Patrick Villieu
Does Public Debt Impair Total Factor Productivity?

2025-27 Nathalie Lazaric, Loubna Echajari & Dorota Leszczyńska
The Multiplicity of Paths to Sustainability, Grand Challenges and Routine Changes: The Long Road for Bordeaux Winemakers

2025-28 Olivier Brette & Nathalie Lazaric
Some Milestones for an Evolutionary-Institutional Approach to the Circular Economy Transition

2025-29 Pierre Boutros, Eliana Diodati, Michele Pezzoni & Fabiana Visentin
Does Training in AI Affect PhD Students' Careers? Evidence from France

2025-30 Luca Bargna, Davide La Torre, Rosario Maggistro & Benjamin Montmartin
Balancing Health and Sustainability: Optimizing Investments in Organic vs. Conventional Agriculture Through Pesticide Reduction

2025-31 Jean-Luc Gaffard
Equilibre général et évolution: un éclairage théorique

2025-32 Patrice Bougette, Frédéric Marty & Simone Vannuccini
Competition Law Enforcement in Dynamic Markets: Proposing a Flexible Trade-off between Fines and Behavioural Injunctions

2025-33 Taekyun Kim, Gabriela Fuentes & Simone Vannuccini
When Robots Met Language (Models): An Exploration of Science and Strategy in the vision-language-action Models Space

2025-34 Frédéric Marty & Thierry Warin
Prévenir les risques de parallélisme de comportements par développeur commun : de la constatation des limites des règles de concurrence au retour du Nouvel Instrument Concurrentiel

2025-35 Luca Fontanelli, Mauro Napoletano & Angelo Secchi
Rethinking Volatility Scaling in Firm Growth

2025-36 Charlie Joyez
Connectivity and Contagion: How Industry Networks Shape the Transmission of Shocks in Global Value Chains

2025-37 FRÉDÉRIC MARTY
L'expérience de l'Union Européenne en matière de contrôle des subventions étrangères - quels enseignements pour la Nouvelle Calédonie ?

2025-38 FRÉDÉRIC MARTY & THIERRY WARIN
Data and Computing Power: The New Frontiers of Competition in Generative AI

2025-39 JEAN-LUC GAFFARD
Finance et globalisation : les enjeux d'une mutation

2025-40 JEAN-LUC GAFFARD
Crise de la globalisation et mutations du capitalisme

2025-41 SARA AMOROSO & SIMONE VANNUCCINI
Teaming up with Large R&D Investors: Good or Bad for Knowledge Production and Diffusion?

2025-42 ILONA DIELEN & JEANNE MOUTON
Impact Assessment of the "Damages Directive": Abuse of Dominance Cases

2025-43 JEAN-LUC GAFFARD & MAURO NAPOLETANO
Vers une nouvelle industrialisation : une stratégie polycentrique

2025-44 MAURO NAPOLETANO & FRANCESCO TONI
The Tartar Steppe of Italian Growth: Strategies for Renewal in a Slowing Europe

2025-45 LAURENCE SAGLIETTO & SYLVIE GRAZIANI-INVERNON
Les relations auditeur/audité au sein de micro-espaces de qualité

2025-46 CHARLIE JOYEZ
Robustness with Adaptation. Ownership Networks of Multinationals through COVID-19

2025-47 THIERRY BLAYAC, PATRICE BOUGETTE & JULES DUBERGA
Share the Ride: The Determinants of Long-Distance Carpooling Pricing Strategies in France

2025-48 JEREMY SROUJI, DOMINIQUE TORRE & QING XU
Should BRICS Members Adopt a Wholesale Digital Payment System?

2025-49 RAFFAELE MINIACI, MICHELE PEZZONI & SOTARO SHIBAYAMA
Exploration in Research Teams: Building on the Shoulders of PhD Students

2025-50 CHARLIE JOYEZ
COMPLEXITY: A Generalized Approach in Stata for Specialization Complexity Indices