CODEBOOK
Measuring subnational trade competitiveness

Robert A. Huber Yannick Stiller Andreas Dür

Version 2.0
(March 16, 2023)

Please cite the data as follows:

Huber, Robert A., Yannick Stiller, and Andreas Dür. Measuring Subnational Trade Competitiveness. Version 2.0. DOI: https://doi.org/10.7910/DVN/BWRGUR

Contact:

robert.huber@reading.ac.uk

Contents

1 Introduction **3**

2 File names **3**

3 STC variable description **4**
3.1 Meta data . . . 4
3.1.1 country_short . . . 4
3.1.2 subnational_level . . . 4
3.1.3 subnational_code . . . 4
3.1.4 subnational_name . . . 4
3.1.5 aggregation_level . . . 4
3.1.6 year . . . 5
3.2 Labour survey specific variables . . . 5
3.2.1 ISIC_coding . . . 5
3.2.2 n_sectors . . . 5
3.2.3 estimated . . . 5
3.2.4 distance . . . 6
3.2.5 count . . . 6
3.2.6 count_* . . . 6
3.2.7 weighted_count . . . 6
3.2.8 share_* . . . 6
3.3 Subnational trade competitiveness . . . 7
3.3.1 stc_* . . . 7

4 RCA variable description **9**
4.1 Meta data . . . 9
4.1.1 country_short . . . 9
4.1.2 year . . . 9
4.1.3 ISIC_code . . . 9
4.2 Trade data . . . 9
4.2.1 A_B_C_hs . . . 9
4.2.2 C_share_A . . . 10
4.2.3 r*a_raw . . . 10
4.3 Revealed comparative advantage . . . 11
4.3.1 rca_* . . . 11

1 Introduction

This codebook accompanies the dataset capturing 'Subnational Trade Competitiveness' and provides detailed documentation of both datasets and all variables in the data. We provide two datasets: 'STC' contains information on the subnational trade competitiveness of regions. This dataset is described in section 3. 'RCA' includes information on the revealed comparative advantage of ISIC categories on a country level. This dataset is described in section 4.

The data is available through the Harvard Dataverse (https://doi.org/10.7910/DVN/BWRGUR). Additionally, data exploration is possible through the accompanying R Shiny App: https://stcapp.shinyapps.io/STCApp/.

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 724107). We also thank statistical offices in various countries for giving us access to their data.

2 File names

- *stc.rds* contains the main dataset including all variables in the data format 'RData' (rds). The data can be loaded into *R* using the *readRDS* command.
- *stc.csv* contains the main dataset including all variables in the data format 'comma separated values' (csv). The data can be loaded into *R* using the *read.csv* command or loaded into any other statistical software supporting csv.
- *rca.rds* contains the underlying RCA data used to create the STC dataset. The data can be loaded into *R* using the *readRDS* command.
- *rca.xlsx* contains the underlying RCA data used to create the STC dataset. The data can be loaded into *R* using the *readxlsx* from the readxl package. Each page in the Excel sheet corresponds to one list in the rds dataset.
- *Codebook.pdf* contains a description of all variables.

3 STC variable description

3.1 Meta data

3.1.1 country_short

This variable labels the respective country using the ISO 3166-1 alpha-3 standard.

3.1.2 subnational_level

subnational_level captures the level on which the data has been aggregated to form the subnational trade competitiveness value. While most countries offer only one level, other countries offer up to three different levels of aggregation. This decisions depends on the respective data availability and quality.

3.1.3 subnational_code

This variable provide a unique code for each subnational entity. Depending on the *subnational_level*, these codes follow the respective standards. Most generally, levels follow the ISO 3166-2 standard for labelling subnational entities. Other entities follow their own standards, such as the 2010-PUMAs in the US or the NUTS-1, NUTS-2, or NUTS-3 standard.

3.1.4 subnational_name

subnational_name provides a character label to the *subnational_code*.

3.1.5 aggregation_level

The variable *aggregation_level* is a ranking of the levels be degree of aggregation. We use the number of unique geographical entities in a subnational level to rank the different levels. In other words, if only one subnational level exists, the variable takes the value 1. If two levels exist, the one with more entities obtains the score of 2. The different levels can be nested, however, this is not necessarily the case. This variable currently ranges from 1 to 3.

3.1.6 year

year captures the year of the data-point. The dataset has a temporal coverage from 1999 to 2019. The coverage deviates by country.

3.2 Labour survey specific variables

3.2.1 ISIC_coding

This variables captures the respective coding scheme. Most estimates are based on ISIC revision 3.1. The data resolution is ISIC group. However, our data also includes labour surveys which apply ISIC revision 3 or ISIC revision 4. Other coding schemes or country specific adaptations were translated into the closest ISIC revision. The R Shiny App allows to trace the coding scheme changes by country.

3.2.2 n_sectors

The variable *n_sectors* captures the number of ISIC groups the total STC score is based upon.

3.2.3 estimated

This variable captures whether the underlying employment shares are based on real data from labour surveys or whether they are estimated by approximating between the closest data points. Before aggregating, sectors which are based on existing labour survey data receive a score of 0. If this sector existed before and after but not in this specific wave of the labour survey, we interpolate its value and the pre-aggregation variable receives a score of 1. Extrapolation receives a score of 2. Thus, when this variable has the score of 0, it would indicate that not a single ISIC group has been estimated. Values below 1 in aggregation suggest that at least some of the ISIC categories are based on real data from labour surveys. The value 1 suggests that all data points have been interpolated. Values above 1 suggest that some data points have been extrapolated. Values of 2 suggest that the employment data for each ISIC group has been extrapolated by carrying forward/backward the closest data point.

3.2.4 distance

The variable *distance* marks the distance of a year to the closest year with actual labour survey data. Thus, values of 0 indicate that this data point is based on a labour survey in the respective year, each additional point marks one year to the closest labour survey based entry.

3.2.5 count

This variable indicates the number of observations in the labour survey with valid ISIC group information and an RCA value.

3.2.6 count_*

Our dataset additionally provides counts by several industry categories. Table 1 below summarises the variables:

Table 1: Number of observation measures

Variable name	content	
count_agri	Count for agriculture	
count_mini	Count for mining	
count_manu	Count for manufacturing	
count_malt	Count for low tech manufacturing	
count_maht	Count for high tech manufacturing	
count_good	Count for goods	(Sum of count_agri, count_mini, and count_manu)
count_sr_n	Count for non-tradable services	
count_sr_t	Count for tradable services	
count_serv	Count for all services	(Sum of count_sr_n, and count_sr_t)
count_trad	Count for all tradable sectors	(Sum of all individual categories above)

3.2.7 weighted_count

The count per district after applying the respective survey weights.

3.2.8 share_*

Our dataset provides the shares of employment by several industry categories. Again, sectors without valid trade data is excluded. Table 2 below summarises the variables:

Table 2: Employment share measures

Variable name	content	
share_agri	Share for agriculture	
share_mini	Share for mining	
share_manu	Share for manufacturing	
share_malt	Share for low tech manufacturing	
share_maht	Share for high tech manufacturing	
share_good	Share for goods	(Sum of share_agri, share_mini, and share_manu)
share_sr_n	Share for non-tradable services	
share_sr_t	Share for tradable services	
share_serv	Share for all services	(Sum of share_sr_n, and share_sr_t)
share_trad	Share for all tradable sectors	(Sum of all individual categories above)

3.3 Subnational trade competitiveness

3.3.1 stc_*

The dataset provides four conceptualisations of subnational trade competitiveness per subnational unit. These measures are then provided as a total as well as per industry group. This results in a total of 20 estimates. Table 3 below provides an overview:

Table 3: Subnational trade competitiveness measures

Variable name	content
stc_sym	Symmetric STC
stc_add	Additive STC
stc_net	Net STC
stc_tba	Trade Balance STC
stc_sym_agri	Symmetric STC for agriculture
stc_add_agri	Additive STC for agriculture
stc_net_agri	Net STC for agriculture
stc_tba_agri	Trade Balance STC for agriculture
stc_sym_mini	Symmetric STC for mining
stc_add_mini	Additive STC for mining
stc_net_mini	Net STC for mining
stc_tba_mini	Trade Balance STC for mining
stc_sym_manu	Symmetric STC for manufacturing
stc_add_manu	Additive STC for manufacturing
stc_net_manu	Net STC for manufacturing
stc_tba_manu	Trade Balance STC for manufacturing
stc_sym_malt	Symmetric STC for low tech manufacturing
stc_add_malt	Additive STC for low tech manufacturing
stc_net_malt	Net STC for low tech manufacturing
stc_tba_malt	Trade Balance STC for low tech manufacturing
stc_sym_maht	Symmetric STC for high tech manufacturing
stc_add_maht	Additive STC for high tech manufacturing
stc_net_maht	Net STC for high tech manufacturing
stc_tba_maht	Trade Balance STC for high tech manufacturing
stc_sym_sr_t	Symmetric STC for tradable services
stc_add_sr_t	Additive STC for tradable services
stc_net_sr_t	Net STC for tradable services
stc_tba_sr_t	Trade Balance STC for tradable services

4 RCA variable description

The RCA dataset contains a list with nine entries. These correspond to the three ISIC revisions (3, 3.1, and 4) as well as three levels of aggregation (division, group, and class). The variable within each list is the same. The data contains RCA information for all countries captured in Comtrade and BaTiS.

4.1 Meta data

4.1.1 country_short

This variable labels the respective country using the ISO 3166-1 alpha-3 standard.

4.1.2 year

year captures the year of the data-point. The dataset has a temporal coverage from 1999 to 2019. The coverage deviates by country.

4.1.3 ISIC_code

ISIC_code captures the ISIC code of the respective ISIC division, group or class.

4.2 Trade data

4.2.1 A_B_C_hs

These variables capture different combinations of three dimensions: A) can take the value 'cntry' or 'world'. 'cntry' refers to country level estimates. 'world' variables are aggregated to the global level. B) is either 'ISIC' or 'total'. The former refers to industry level values, the latter to aggregation across all industries. C) either means exports ('ex') or imports ('im'). Thereby, these eight variables contain all information necessary to calculate all four RCA measures.

Table 4: RCA trade data

Variable name	content
cntry_ISIC_ex_hs	Exports in an ISIC category on the country level
cntry_ISIC_im_hs	Imports in an ISIC category on the country level
cntry_total_ex_hs	Total country exports
cntry_total_im_hs	Total country imports
world_ISIC_ex_hs	Global total exports in an ISIC category
world_ISIC_im_hs	Global total imports in an ISIC
world_total_ex_hs	Global total exports
world_total_im_hs	Global total imports

4.2.2 C_share_A

These four variables correspond to the shares outlined in equation 2 (RCA) in the corresponding research paper. Additionally, these values are directly used in equation 4 (RCA ($additive$)) and 6 (RCA ($trade\ balance$)).

Table 5: RCA trade data

Variable name	content
ex_share_country	Country's exports share of an ISIC category
ex_share_world	Global exports share of an ISIC category
im_share_country	Country's imports share of an ISIC category
im_share_world	Global imports share of an ISIC category

4.2.3 r*a_raw

These variables capture the RCA outlined in equation 2. the place holder refers to whether exports (rxa_raw) or imports (rma_raw) are used. This variable appears in equation 5 (RCA (net)).

4.3 Revealed comparative advantage

4.3.1 rca_*

The dataset provides four conceptualisations of the revealed comparative advantage, which is then the basis for the stc outlined above in section 3. We provide four different operationalisations, outlined in the main paper.

Table 6: RCA trade data

Variable name	content
rca_sym	RCA (*symmetric*)
rca_add	RCA (*additive*)
rca_net	RCA (*net*)
rca_tba	RCA (*trade balance*)