

11 December 2019

**Readme file for “Forced Migration and Human Capital: Evidence from Post-WWII Population Transfers” by Sascha O. Becker, Irena Grosfeld, Pauline Grosjean, Nico Voigtländer, and Ekaterina Zhuravskaya**

The data and code used to support the findings of this study have been deposited in the Open ICPSR repository for the American Economic Association Data and Code Repository under the following project reference number: **openicpsr-115202** (Project title: “Data and Code for: Forced Migration and Human Capital: Evidence from Post-WWII Population Transfers”). Please download all replication materials from that repository.

The replication folder contains several data files and do files, which can be used to i) generate all the datasets used in the analysis from the raw data and ii) replicate all results from the tables and figures in the paper and the appendix.

**In order to generate the datasets used in the analysis and then replicate all of the results in the paper one simply needs to define the path to the replication folder on line 9 of the Main do.do in the root directory of the replication folder and then run Main do.do.**

(**Main do.do** first calls the do files that transform raw data into the datasets used for the analysis and then calls the do files that replicate the results. Because relative paths were used in the coding, the path names used throughout the do files will not need to be changed.)

Below we explain the structure of the replication folder, the data, and sub-tasks executed by the **Main do.do**.

**i) Generating the datasets used in the analysis from the raw data (lines 17-47 of Main do.do):**

The files necessary to generate the datasets used in the analysis from the raw data are contained in the **Original\_source\_data** folder. This folder contains the raw data from the Ancestry survey and from the Diagnoza survey and the do files and additional data sources necessary to transform these raw data into the formats found in the original\_data subfolder of the Data folder (as described below, this folder contains the datasets used in the analysis). All of the raw data are described inside the subfolders of **Original\_source\_data** folder.

The **Original\_source\_data** folder has three subfolders:

- **Ancestry**

This folder contains the following information:

- generating\_cbos\_with\_1931\_from\_source.do: Do file that generates the following datafile: cbos\_with\_1931.dta data file in the original\_data subfolder in the Data folder from raw data contained in the source\_ancestry\_survey\_files subfolder.
- source\_ancestry\_survey\_files subfolder, which includes:
  - Ancestry\_survey\_set\_original.dta: (Becker, S., I. Grosfeld, P. Grosjean, N. Voigtländer, and E. Zhuravskaya (2016). Ancestry Survey. <http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (accessed December 11, 2019)). This is raw Ancestry survey data, the original survey conducted by the authors.
  - Ancestry survey codebook: Codebook for the Ancestry survey data. This Codebook can also be found here: <http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (accessed December 11, 2019).
  - Several data files necessary to generate cbos\_with\_1931.dta
  - Description\_of\_ancestry\_source\_files.docx: description of sources of all other raw data in the folder and codebooks for all the data files necessary to generate cbos\_with\_1931.dta in addition to the Ancestry Survey file (Ancestry\_survey\_set\_original.dta).

The following link provides more detail on how the Ancestry Survey was conducted, the sampling procedure, and the anonymization procedure:

<http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (follow a link to Ancestry Survey, then : [Detailed information on sampling and anonymization](#)) (accessed December 11, 2019).

- **Diagnoza**

This folder contains the following information:

- generating\_diagnoza\_2015\_from\_source.do: Do file that generates the following data file: diagnoza\_2015.dta data file in the original\_data subfolder of the Data folder from Diagnoza raw data included in the source\_diagnoza\_files subfolder.

- source\_diagnoza\_files subfolder that contains all raw data files from Diagnoza Survey (Council for Social Monitoring (2015). Diagnoza (Social Diagnosis). <http://diagnoza.com/index-en.html>, accessed December 11, 2019), as well as the Codebook for Diagnoza source files (excel document: Codebook\_Diagnoza\_source\_files.xlsx) and a document that describes each of the Diagnoza source file and gives the exact sources of each of these data (see word doc: Description\_of\_diagnoza\_source\_files.docx)

- **GIS\_MAPS**

This folder contains the shapefiles used for the generation of the data. Information describing each shapefile (including sources) is found in the word document: Description\_of\_GIS\_maps.docx.

ii) Replicating all results in the paper (lines 51-onward of **Main\_do.do**)

Below, we list all tables and figures in the paper and appendix, with reference to the corresponding do file and line of code.

The **DoFiles folder** contains all do files that will run as indicated in the Main\_do.do and contain all the code necessary for replication. Here, we provide a short description of each do file:

**1.Build\_datasets\_for\_analysis.do:** this do file refers to the original data sources (from the original\_data subfolder in Data) and contains all the code necessary to construct all the variables in the analysis that are transformed from the original variables. This do file generates all data files that are saved in the generated\_data subfolder in the Data folder and then used in the analysis.

All the remaining do files run on generated data and contain the code necessary for the replication of Tables and Figures in the paper. The name of each do file refers to the data sources (although the code generally combines these with additional datasets):

**2.Regressions\_Diagnoza.do:** generates Tables and Figures that use the Diagnoza data (see below for more detail on each Table and Figure);

**3.Regressions\_CBOS.do:** generates Tables and Figures that use the Ancestry survey (see below for more detail on each Table and Figure);

**4.Regressions\_Diagnoza\_AncestrySurvey.do:** combines the Diagnoza and the Ancestry survey and generates Table A.18;

**5.Fig\_App\_DataQuality\_CBOS.do** generates Table A.9 from the Ancestry Survey;

**6.Regressions\_LiTS.do:** generates Tables and Figures that use the Life in Transition Survey data (see below for more detail on each Table and Figure);

**7.Regressions\_CBOS\_community.do** generates Table A.23 from the Ancestry Survey.

The **Data** folder include two subfolders. The first one: **original\_data** contains the original data files called by the do files in the DoFiles folder. The second: **generated\_data** contains all data generated by the do files in the DoFiles folder and used in the analysis.

The original\_data folder contains the following data :

Excel files:

- **timing\_of\_arrival.xlsx:** Data on the number of settlers (of different origins) arriving in Western Territories at different dates. Source: the Document of the Ministry of Recovered Territories, No.1661 (The Central Archives of Modern Records in Warsaw).

- **pow\_1931\_center\_in\_1952\_woj.xlsx:** 1931 powiat codes spatially matched to 1952 wojewodztwo code. Authors' calculation using the map from MPIDR [Max Planck Institute for Demographic Research] and CGG [Chair for Geodesy and Geoinformatics, University of Rostock] 2011: MPIDR Population History GIS Collection – Rostock.

Codebook:

Variable: POW\_CODE\_1931: Powiaty code in 1931

Variable: WOJ\_1952: Wojewodztwo1952 name

Variable: WOJ\_CODE\_1952: Wojewodztwo1952 code

Stata files:

Surveys:

- **diagnoza\_2015.dta:** 2015 Diagnoza data. (generated by the **Main\_do.do** from raw files in the folder /Original\_source\_data/Diagnoza. The source of the raw data is (as described in Description\_of\_diagnoza\_source\_files.docx): Council for Social Monitoring. 2015. Diagnoza (Social Diagnosis) see <http://diagnoza.com/index-en.html> for questionnaire and methodological details). Data at respondent level;

- **cbos\_with\_1931.dta:** Ancestry survey. Data at respondent level. (generated by the **Main\_do.do** from raw files in the folder /Original\_source\_data/Ancestry. The source of the raw data is (as described in Description\_of\_ancestry\_source\_files.docx): Becker, S., I. Grosfeld, P. Grosjean, N. Voigtländer, and E. Zhuravskaya (2016). Ancestry Survey.

<http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (accessed December 11, 2019);

- **LiTS\_Poland\_Ukraine\_Belarus\_Lithuania\_geocoded.dta**: Data from 2016 Life in Transition Survey. Data for Poland, Ukraine, Belarus, Lithuania extracted from the full survey. (Source: European Bank for Reconstruction and Development. 2016. "Life in Transition Survey III: A decade of measuring transition." <https://www.ebrd.com/what-we-do/economic-research-and-data/data/lits.html> for link to questionnaire (with codebook) and raw data, and <http://litsonline-ebrd.com/methodology-annex/> for Methodology Appendix and sampling strategy). Data at respondent level.

Additional historical, geo, and census data:

- **1897\_russianscensus.dta**: 1897 Russian Census for the Russian partition of the Polish-Lithuanian Commonwealth. Data on literacy. Data at 1931 powiat level. (Source: Troynitsky, N. (1899). *The First General Census of Russian Empire in 89 Volumes*. The Central Statistical Bureau of the Ministry of Internal Affairs of the Russian Empire, St. Petersburg, 1899-1905).

Codebook:

Variable: ID\_powiaty\_code\_1931 - Label (and meaning): "Powiaty code 1931"

Variable: litrusrate\_p - Label (and meaning): "literacy of polish native speakers in russian language 1897"

Variable: lithotherrate\_mp - Label (and meaning): "literacy of polish native speakers in non-russian language 1897, male poles"

Variable: lithotherrate\_fp - Label (and meaning): "literacy of polish native speakers in non-russian language 1897, female poles"

Variable: lithotherrate\_p - Label (and meaning): "literacy of polish native speakers in non-russian language 1897"

Variable: educate\_p - Label (and meaning): "rate of education for polish native speakers in 1897"

- **1921\_census\_pow\_for\_kresy\_survey.dta**: 1921 Polish Census (Second Polish Republic). Data on total population (including urban versus rural areas) and literacy of Roman Catholics. Data at 1931 powiat level. (Source: Pierwszy Powszechny Spis Rzeczypospolitej Polskiej z dnia 30 wrzesnia 1921 roku. 1926, 1927, 1928. Warszawa: Glowny Urzad Statystyczny).

Codebook: See Description\_of\_ancestry\_source\_files.docx in Original\_source\_data folder, Ancestry subfolder, source\_ancestry\_survey\_files subfolder.

- **1931\_census\_percentages\_with\_kresy\_dummy**: 1931 Polish Census (Second Polish Republic). Data on literacy, ethnicity, religion, rural vs urban. Data at 1931 powiat level. (Source: Drugi Powszechny Spis Ludnosci z dn. 9.XII.1931 R. 1938. Warszawa: Glowny Urzad Statystyczny)

### Codebook:

Variable: powiaty\_code\_1931- Label (and meaning): "Powiaty code 1931"

Variable: r\_roman\_catholic\_1931pow- Label (and meaning): "Share Roman Catholics in powiat 1931"

Variable: r\_greek\_catholic\_1931pow - Label (and meaning): "Share Greek Catholics in powiat 1931"

Variable: r\_orthodox\_1931pow- Label (and meaning): "Share Orthodox in powiat 1931"

Variable: r\_protestant\_lutherian\_1931pow- Label (and meaning): "Share Lutheran in powiat 1931"

Variable: r\_protestant\_reformed\_1931pow- Label (and meaning): "Share Reformed Protestants in powiat 1931"

Variable: r\_non\_identified\_1931pow - Label (and meaning): "Share religion n/a in powiat 1931"

Variable: r\_christian\_other\_1931pow- Label (and meaning): "Share Christian other in powiat 1931"

Variable: r\_non\_christian\_other\_1931pow- Label (and meaning): "Share non-Christian other in powiat 1931"

Variable: r\_noconfession\_1931pow- Label (and meaning): "Share no confession other in powiat 1931"

Variable: r\_missing\_1931pow- Label (and meaning): "Share religion missing other in powiat 1931"

Variable: r\_jewish\_1931pow- Label (and meaning): "Share Jewish in powiat 1931"

Variable: r\_roman\_catholic\_1931rural- Label (and meaning): "Share Roman Catholics in powiat 1931 - rural areas"

Variable: r\_greek\_catholic\_1931rural- Label (and meaning): "Share Greek Catholics in powiat 1931 - rural areas"

Variable: r\_orthodox\_1931rural- Label (and meaning): "Share Orthodox in powiat 1931 - rural areas"

Variable: r\_protestant\_lutherian\_1931rural- Label (and meaning): "Share Lutheran in powiat 1931 - rural areas"

Variable: r\_protestant\_reformed\_1931rural- Label (and meaning): "Share Reformed Protestants in powiat 1931 - rural areas"

Variable: r\_non\_identified\_1931rural - Label (and meaning): "Share religion n/a in powiat 1931 - rural areas"

Variable: r\_christian\_other\_1931rural- Label (and meaning): "Share Christian other in powiat 1931 - rural areas"

Variable: r\_non\_christian\_other\_1931rural- Label (and meaning): "Share non-Christian other in powiat 1931 - rural areas"

Variable: r\_noconfession\_1931rural- Label (and meaning): "Share no confession other in powiat 1931 - rural areas"

Variable: r\_missing\_1931rural- Label (and meaning): "Share religion missing other in powiat 1931 - rural areas"

Variable: r\_jewish\_1931rural- Label (and meaning): "Share Jewish in powiat 1931 - rural areas"

Variable: r\_roman\_catholic\_1931urban- Label (and meaning): "Share Roman Catholics in powiat 1931 - urban areas"

Variable: r\_greek\_catholic\_1931urban- Label (and meaning): "Share Greek Catholics in powiat 1931 - urban areas"

Variable: r\_orthodox\_1931urban - Label (and meaning): "Share Orthodox in powiat 1931 - urban areas"

Variable: r\_protestant\_lutherian\_1931urban- Label (and meaning): "Share Lutheran in powiat 1931 - urban areas"

Variable: r\_protestant\_reformed\_1931urban- Label (and meaning): "Share Reformed Protestants in powiat 1931 - urban areas"

Variable: r\_non\_identified\_1931urban - Label (and meaning): "Share religion n/a in powiat 1931 - urban areas"

Variable: r\_christian\_other\_1931urban- Label (and meaning): "Share Christian other in powiat 1931 - urban areas"

Variable: r\_non\_christian\_other\_1931urban- Label (and meaning): "Share non-Christian other in powiat 1931 - urban areas"

Variable: r\_noconfession\_1931urban- Label (and meaning): "Share no confession other in powiat 1931 - urban areas"

Variable: r\_missing\_1931urban- Label (and meaning): "Share religion missing other in powiat 1931 - urban areas"

Variable: r\_jewish\_1931urban - Label (and meaning): "Share Jewish in powiat 1931 - urban areas"

Variable: r\_protestant\_1931pow - Label (and meaning): "Share All Protestants in powiat 1931"

Variable: r\_protestant\_1931rural- Label (and meaning): "Share All Protestants in powiat 1931- rural areas"

Variable: r\_protestant\_1931urban- Label (and meaning): "Share All Protestants in powiat 1931- urban areas"

Variable: Lang\_polish\_1931pow - Label (and meaning): "Share speaks Polish in powiat 1931"

Variable: Lang\_ukranian\_1931pow- Label (and meaning): "Share speaks Ukrainian in powiat 1931"

Variable: Lang\_ruthenian\_1931pow- Label (and meaning): "Share speaks Ruthenian in powiat 1931"

Variable: Lang\_belorussian\_1931pow- Label (and meaning): "Share speaks Belarusian in powiat 1931"

Variable: Lang\_russian\_1931pow- Label (and meaning): "Share speaks Russian in powiat 1931"

Variable: Lang\_czech\_1931pow - Label (and meaning): "Share speaks Czech in powiat 1931"

Variable: Lang\_other\_1931pow - Label (and meaning): "Share speaks other language in powiat 1931"

Variable: Lang\_lithuanian\_1931pow- Label (and meaning): "Share speaks Lithuanian in powiat 1931"

Variable: Lang\_german\_1931pow - Label (and meaning): "Share speaks German powiat 1931"

Variable: Lang\_polish\_1931rural - Label (and meaning): "Share speaks Polish in powiat 1931 - rural areas"

Variable: Lang\_ukranian\_1931rural- Label (and meaning): "Share speaks Ukrainian in powiat 1931 - rural areas"

Variable: Lang\_ruthenian\_1931rural- Label (and meaning): "Share speaks Ruthenian in powiat 1931 - rural areas"

Variable: Lang\_belorussian\_1931rural- Label (and meaning): "Share speaks Belarusian in powiat 1931 - rural areas"

Variable: Lang\_russian\_1931rural- Label (and meaning): "Share speaks Russian in powiat 1931 - rural areas"

Variable: Lang\_czech\_1931rural- Label (and meaning): "Share speaks Czech in powiat 1931 - rural areas"

Variable: Lang\_other\_1931rural- Label (and meaning): "Share speaks other language in powiat 1931 - rural areas"

Variable: Lang\_lithuanian\_1931rural - Label (and meaning): "Share speaks Lithuanian in powiat 1931 - rural areas"

Variable: Lang\_german\_1931rural- Label (and meaning): "Share speaks German powiat 1931 - rural areas"

Variable: Lang\_polish\_1931urban- Label (and meaning): "Share speaks Polish in powiat 1931- urban areas"

Variable: Lang\_ukranian\_1931urban - Label (and meaning): "Share speaks Ukrainian in powiat 1931 - urban areas"

Variable: Lang\_ruthenian\_1931urban- Label (and meaning): "Share speaks Ruthenian in powiat 1931 - urban areas"

Variable: Lang\_belorussian\_1931urban- Label (and meaning): "Share speaks Belarusian in powiat 1931 - urban areas"

Variable: Lang\_russian\_1931urban - Label (and meaning): "Share speaks Russian in powiat 1931 - urban areas"

Variable: Lang\_czech\_1931urban - Label (and meaning): "Share speaks Czech in powiat 1931 - urban areas"

Variable: Lang\_other\_1931urban"Share speaks other language in powiat 1931 - urban areas"

Variable: Lang\_lithuanian\_1931urban - Label (and meaning): "Share speaks Lithuanian in powiat 1931 - urban areas"

Variable: Lang\_german\_1931urban- Label (and meaning): "Share speaks German powiat 1931 - urban areas"

Variable: Lang\_jews\_1931\_pow - Label (and meaning): "Share speaks Yiddish powiat 1931- Label (and meaning): "

Variable: Lang\_jews\_1931\_rural- Label (and meaning): "Share speaks Yiddish powiat 1931 - rural areas"

Variable: Lang\_jews\_1931\_urban- Label (and meaning): "Share speaks Yiddish powiat 1931 - urban areas"

Variable: urbanization\_1931- Label (and meaning): "Urbanisation powiat 1931"

Variable: literate\_1931pow- Label (and meaning): "Share literate powiat 1931"

Variable: readonly\_1931pow - Label (and meaning): "Share can read only powiat 1931"

Variable: notliterate\_1931pow- Label (and meaning): "Share illiterate powiat 1931"

Variable: literate\_1931rural- Label (and meaning): "Share literate powiat 1931 - rural areas"

Variable: readonly\_1931rural- Label (and meaning): "Share can read only powiat 1931 - rural areas"

Variable: notliterate\_1931rural - Label (and meaning): "Share literate powiat 1931 - rural areas"

Variable: literate\_1931urban- Label (and meaning): "Share literate powiat 1931 - urban areas"

Variable: readonly\_1931urban- Label (and meaning): "Share can read only powiat 1931- Label (and meaning): "

Variable: notliterate\_1931urban - Label (and meaning): "Share literate powiat 1931 - urban areas"

Variable: Name\_powiat\_31- Label (and meaning): "Name of the powiat in 1931"

Variable: in\_kresy- Label: "In Kresy": Dummy if powiat is in Kresy ( 1 = Yes)

- **1931\_pow\_curzon\_line\_distance.dta**: distance from centroid of each 1931 powiat to the Curzon line (in km). Data at 1931 powiat level. (Distance computed in ArcGIS using shapefile: Poland\_1931\_powiaty.shp. The Curzon line delimitation is in the Border\_Curzon\_line.shp shapefile [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources])

#### Codebook:

Variable: powiaty\_code\_1931- Label (and meaning): "Powiaty code 1931"

Variable: disCursLineKm - Label: "dist powiat 1931 - curzon line, km": Distance between centeroid of the powiat and the Curzon line, in km.

- **anc\_id\_curzon\_line\_distance.dta**: distance of location of each ancestor in Ancestry survey to the Curzon line (in km). Data at ancestor level, for each ancestor for whom we could identify and geo-code the place of residence. (Source: Becker, S., I. Grosfeld, P. Grosjean, N. Voigtländer, and E. Zhuravskaya (2016). Ancestry Survey. <http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (accessed December 11, 2019)) The Curzon line delimitation is in the Border\_Curzon\_line.shp shapefile [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources].



### Codebook

Variable: id\_ancestor – Label: “id\_ancestor”: ID number of the ancestor in the Ancestry Survey

Variable: disCursLineKm – Label: “disCursLineKm”: Distance from ancestor’s origin location to the Curzon line, in km.

- **anc\_ids\_modern\_country\_origin.dta**: country of origin of each ancestor in Ancestry survey. Data at ancestor level, for each ancestor for whom we could identify and geo-code the place of residence. (Source: Becker, S., I. Grosfeld, P. Grosjean, N. Voigtländer, and E. Zhuravskaya (2016). Ancestry Survey. <http://www.parisschoolofeconomics.com/zhuravskaya-ekaterina/erc-economics-of-prejudice/> (accessed December 11, 2019))

### Codebook:

Variable: id\_ancestor – Label: “id\_ancestor”: ID number of the ancestor in the Ancestry Survey

Variable: anc\_origin\_country\_gmi – Label: "Ancestor's origin country - abbreviated": Ancestor country of origin: abbreviated country name

label var anc\_origin\_country "Ancestor's origin country": Ancestor country of origin

- **centroid\_of\_pow\_1931\_in\_powiaty\_today.dta**: correspondence between 1931 powiaty and present-day powiaty. Data at 1931 powiat level. (Shapefiles: powiaty.shp for present-day powiaty and Polen\_1931.shp and Poland\_1931\_powiaty.shp for historical powiaty [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources])

### Codebook

Variable: powiaty\_code\_str – Label: “powiaty code”: Present-day powiat code

Variable: pow\_code\_1931 – Label: “powiaty code, 1931”: 1931 powiat code

- **centroid\_pow\_1931.dta**: coordinates of centroid of 1931 powiaty. Data at 1931 powiat level. (Shapefiles: powiaty.shp for present-day powiaty and Polen\_1931.shp and Poland\_1931\_powiaty.shp for historical powiaty [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources])

### Codebook:

Variable: powiaty\_code\_1931 – Label (and meaning): "Powiaty code 1931"

Var:lat\_1931 – Label (and meaning): "Latitude of centroid of 1931 powiaty"

Var:lon\_1931 – Label (and meaning): "Longitude of centroid of 1931 powiaty"

- **current\_powiaty\_curzon\_line\_distance.dta**: distance from centroid of each present-day powiat to the Curzon line (in km). Data at present-day powiat level. Distance computed from following shapefile: powiaty.shp [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources].

Codebook:

Variable: powiaty\_code – Label (and meaning) "Powiaty code"

Variable: disCursLineKm - Label: “dist current powiat - curzon line, km”: Distance between present-day powiat and Curzon line, in km.

- **emigrants\_census2011.dta**: Data from the Polish 2011 Census. Data on population and answers to the question: “*How many members of your household have emigrated?*”. Data at wojewodztwa level. (Source: Statistics Poland. 2011. National Census of Population and Housing 2011. <https://stat.gov.pl/en/national-census/national-census-of-population-and-housing-2011/>, accessed on December 11, 2019)

Codebook:

Variable: woj – Label (and meaning) "województwa"

Variable: population\_2011 – Label (and meaning) "Population, 2011 Census"

Variable: number\_emigrants – Label (and meaning) "Number of emigrants in population, 2011 Census"

- **industrial\_production\_1954\_ready.dta**: Data on industrial production per capita in 1954. Data from the 1954 statistical yearbook. Data at 1952 region (województwa) level. (Source: Poland, Ministry of Information, Statistical Yearbook of Poland 1954)

Codebook

Variable: name\_województwa1952 – Label: “name woj. 1952”: name of 1952 wojewodztwa

Variable: ind\_prod\_1954\_per\_population – Label (and meaning) “log industrial production 1954 per capita in woj”

Variable: ind\_prod\_1954\_per\_urban\_pop – Label (and meaning): “log industrial production 1954 per urban resident in woj”

- **pow\_1931\_geo.dta**: 1931 powiaty matched to FAO data. Data at 1931 powiat level. (Spatial matching done using: Poland\_1931\_powiaty.shp shapefile [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources]). (Sources: FAO and IIASA. 2012. GAEZ: Global Agro-Ecological Zones. FAO GAEZ Data Portal version 3.0

(release May 25, 2012) <http://gaez.fao.org/Main.html> ; Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (<http://srtm.csi.cgiar.org>)

Codebook:

Variable: zone\_code – Label: “Unique pow identifier in 1931”: Powiaty code in 1931

Variable: ann\_prpc\_cv – Label (and meaning): “Coefficient of variation of annual precipitation (mm)”

Variable: ann\_prpc\_pet – Label (and meaning): “Mean annual precipitation potential evatranspiration ratio”

Variable: ann\_prpc\_sd – Label (and meaning): “Standard deviation of annual precipitation (mm)”

Variable: ann\_prpc – Label (and meaning): “Mean annual precipitation (mm)”

Variable: ann\_temp – Label (and meaning): “Mean annual temperature (C)”

Variable: barley – Label (and meaning): “Barley suitability index (1-100)”

Variable: elevation – Label (and meaning): “Altitude (m)”

Variable: grow\_per – Label (and meaning): “Mean growing period length”

Variable: potato – Label (and meaning): “Potato suitability index (1-100)”

Variable: q1\_prpc\_pet – Label (and meaning): “Mean Q1 precipitation potential evatranspiration ratio”

Variable: q2\_prpc\_pet – Label (and meaning): “Mean Q2 precipitation potential evatranspiration ratio”

Variable: q3\_prpc\_pet – Label (and meaning): “Mean Q3 precipitation potential evatranspiration ratio”

Variable: q4\_prpc\_pet – Label (and meaning): “Mean Q4 precipitation potential evatranspiration ratio”

Variable: rice – Label (and meaning): “Rice suitability index (1-100)”

Variable: ruggedness – Label (and meaning): “Ruggedness index”

Variable: sunflower – Label (and meaning): “Sunflower suitability index (1-100)”

Variable: wheat – Label (and meaning): “Wheat suitability index (1-100)”

Variable: powiaty\_code\_1931 – Label (and meaning): “powiaty code 1931 at origin”

Variable: powiaty\_code\_1931\_str – Label (and meaning): “powiaty code 1931 at origin”

- **powiaty\_area.dta**: powiat area, in hectares (generated from ArcGIS from powiaty.shp shapefile [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources]). Data at present-day powiat level.

Codebook:

Variable: pow\_area\_ha – Label (and meaning): “powiaty area, hectares”

Variable: powiaty\_code – Label (and meaning): “powiaty code”

- **powiaty\_code\_today\_1931.dta**: correspondence between present-day powiaty and 1931 powiaty. Constructed from two shapefiles: powiaty.shp and Poland\_1931\_powiaty.shp [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and information on sources]

### Codebook:

Variable: powiaty\_code\_1931 – Label (and meaning): "Powiaty code 1931"

Variable: powiaty\_code – Label (and meaning): "Powiaty code"

- **powiaty\_history\_geography.dta**: historical and geographic information on present-day powiaty. Data at present-day powiat level. (Sources: *Narodowy Spis Powszechny z 3 grudnia 1950 r. 1955*. Warszawa: Główny Urząd Statystyczny; Ministry of Recovered Territories, Document No. 1661. The Central Archives of Modern Records in Warsaw; Kaiserliches Statistisches Amt (Editor): *Statistik des Deutschen Reichs, Band 150: Die Volkszählung im Deutschen Reich am 1. Dezember 1900*. Berlin; Puttkammer & Mühlbrecht 1903)

### Codebook:

Variable: powiat – Label (and meaning): "name powiat, polish characters 2005"

Variable: siedziba – Label (and meaning): "capital of powiat, polish characters 2005"

Variable: wojewodztw – Label (and meaning): "województwo, polish character 2005"

Variable: pow\_ha – Label (and meaning): "surface of powiat, ha"

Variable: ludn\_1998 – Label (and meaning): "population 1998"

Variable: ludn\_1999 – Label (and meaning): "population 1999"

Variable: ludn\_2000 – Label (and meaning): "population 2000"

Variable: woj – Label (and meaning): "województwo code"

Variable: pgz – Label (and meaning): "type of powiat 1 - city; 2 -normal powiat". Values: 1: city; 2: normal powiat

Variable: longitude – Label (and meaning): "longitude centroid pow"

Variable: latitude – Label (and meaning): "latitude centroid pow"

Variable: powiaty\_code – Label (and meaning): "powiaty\_code"

Variable: name\_powiat – Label (and meaning): "name powiat, latin characters, 2005"

Variable: name\_województwa – Label (and meaning): "name woj, latin characters, 2005"

Variable: powiat\_1952 – Label (and meaning): "name powiat\_1952"

Variable: name\_województwa1952 – Label (and meaning): "name woj. 1952"

Variable: previous\_powiaty – Label (and meaning): "admin changes of powiaty" (see labels for changes)

Variable: WT – Label : "dummy for western territory": Dummy =Yes if powiat in Western Territories

Variable: Austria – Label (and meaning): "austrian partition": Powiat was part of the Austrian partition (Values: 0: No; 1: Yes)

Variable: Russia – Label (and meaning): "russian partition": Powiat was part of the Russian partition (Values: 0: No; 1: Yes)

Variable: Prussia – Label (and meaning): "prussian partition": Powiat was part of the Prussian partition (Values: 0: No; 1: Yes)

Variable: border – Label (and meaning): "pow. on a historical border": Powiat is on former border between former Empires (Values: 0: No; 1: Yes)

Variable: prussia\_notwt – Label (and meaning): "prussian, non WT": Powiat was part of the Prussian partition but not Western Territories (Values: 0: No; 1: Yes)

Variable: border\_russia\_prussia – Label (and meaning): "border russia-prussia": Powiat is on former

border between Russian Empire and Prussia (Values: 0: No; 1: Yes)

Variable: border\_wt\_prussianotwt – Label (and meaning): “border WT\_prussianotWT” Powiat is on former border between Prussia Western Territories and Prussia – not Western Territories (Values: 0: No; 1: Yes)

Variable: border\_russia\_austria – Label (and meaning): “border russia-austria”: Powiat is on former border between Russian Empire and Austria-Hungary (Values: 0: No; 1: Yes)

Variable: border\_russia\_prussia\_notwt – Label (and meaning): “border russia-prussia\_notwt”: Powiat is on former border between Russian Empire and Prussia – Not Western Territories (Values: 0: No; 1: Yes)

Var borderregion – Label (and meaning): “woj. on the border”: Wojewodztwo is on former border between former Empires (Values: 0: No; 1: Yes)

Variable: empire\_code – Label: “empire code” code of former Empire: Values: 1: Russian Empire; 2: Austria Hungary; 3: Prussia without Western Territories; 4: Western Territories ; 5: On the border of former Empires.

Variable: city – Label (and meaning): “dummy city”: dummy for city with a status of powiat (Values: 0: No; 1: Yes)

Variable: all\_lived\_1939 – Label (and meaning): “WOJ.L:among ALL censused in 1950 how many lived in this woj in 1939” [WOJ or woj stands everywhere for wojewodztwo]

Variable: population1950 – Label (and meaning): “WOJ.L:total population of woj in 1950”

Variable: lived\_current\_territory\_of\_poland – Label (and meaning): “WOJ.L:lived in the current territory of poland in 1939”

Variable: not\_known\_location\_in\_1939 – Label (and meaning): “WOJ.L:not known location in 1939”

Variable: abroad\_1939 – Label (and meaning): “WOJ.L:inflow from abroad in 1939 (including USSR, France, Germany, Other)”

Variable: ussr\_1939 – Label (and meaning): “WOJ.L:inflow from USSR in 1939”

Variable: france\_1939 – Label (and meaning): “WOJ.L:inflow from France in 1939”

Variable: germany\_1939 – Label (and meaning): “WOJ.L:inflow from Germany in 1939”

Variable: other\_abroad1939 – Label (and meaning): “WOJ.L:inflow from other abroad in 1939 (other than USSR, France, Germany)”

Variable: lived\_other\_woj\_in\_old\_territory – Label (and meaning): “WOJ.L:inflow from other woj in central poland in 1939

Variable: inflow\_from\_wt – Label (and meaning): “WOJ.L:inflow from WT” [WT stands everywhere for Western Territories]

Variable: inflow\_from\_old\_territories – Label (and meaning): “WOJ.L:inflow from central poland in 1939”

Variable: lived\_this\_woj\_in1939\_wt\_in1950 – Label (and meaning): “WOJ.L:outflow from this woj. to WT between 1939 and 1950”

Variable: lived\_this\_woj\_in1939\_oldterr\_in – Label (and meaning): “WOJ.L:outflow from this woj. to central poland between 1939 and 1950”

Variable: lived\_same\_woj1939 – Label (and meaning): “WOJ.L: lived in the same woj. in 1939”

Variables: inflow\_warszawa1939 to inflow\_rzeszowskie1939: indicate inflows from different wojewodztwa. The name of the wojewodztwo is indicated in the variable name and in the variable label. WT stands everywhere for Western Territories, and old. ter. stands for Central Poland.

Variable: migr\_ussr\_1955\_1959 – Label (and meaning): “WOJ.L:inflow from USSR b/w 1955 and 1959”

Variable: inflow\_ussr\_all – Label (and meaning): “WOJ.L:inflow from USSR b/w 1939 and 1959”

Variable: perc\_ussr\_all – Label (and meaning): “WOJ.L:share of population coming from USSR b/w 1939 and 1959 in 1950”

The following variables have information at the Powiat level [POW stands everywhere for Powiat]:

Variable: population\_powiat\_1950\_t – Label (and meaning): “POW.L:population in 1950”

Variable: autochtons\_t – Label (and meaning): “POW.L:lived in WT in 1939”

incoming\_population\_total\_t – Label (and meaning): “POW.L:inflow into pow. after war”

Variables: *from\_Warsaw\_t* to: *unknown\_origin\_t* indicate inflows into the powiat from other areas, as indicated in the variable names and/or labels.

Variables *population\_powiat\_1950\_r* to *unknown\_origin\_r* have similar information for rural areas in powiaty. All variables in this list have self-explanatory labels.

Variables *population\_powiat\_1950\_u* to *unknown\_origin\_u* have similar information for urban areas in powiaty. All variables in this list have self-explanatory labels.

All other variables have self-explanatory labels.

- **rail\_lines\_stations\_1946\_by\_powiat.dta**: railway density at powiat level. Data from historical map of the Polish railway system in 1946. Data on number of railway stations per powiat, and on number of rail lines passing through powiat. Data at present-day powiaty level. Generated and digitized by the authors from the *Map Archive of Wojskowy Instytut Geograficzny (WIG), 1919-1939* using as source the image available here:

[http://maps.mapywig.org/m/Polish\\_maps/various/Small\\_scale\\_maps/MAPA\\_SIECI\\_KOLEJOWEJ\\_RP\\_1\\_M\\_1946.jpg](http://maps.mapywig.org/m/Polish_maps/various/Small_scale_maps/MAPA_SIECI_KOLEJOWEJ_RP_1_M_1946.jpg) (accessed on July 4, 2019) [see GIS\_MAPS source subfolder in Original\_source\_data folder for shapefile and details on sources].

#### Codebook:

Variable: powiaty\_code – Label (and meaning): “powiaty\_code”

Variable: nb\_railway\_stations\_1946 – Label (and meaning): “number of railway stations in 1946 in powiat”

Variable: nb\_rail\_lines\_1946 – Label (and meaning): “number of rail lines passing through powiat in 1946”

- **ukrainians\_forced\_to\_move\_to\_USSR.dta**: Data on percentage of Ukrainians at powiat level forced to move to USSR in 1945 and 1946. Data on the number of Ukrainians who left in 1945 and 1946 (for relevant powiaty) is from Gawryszewski, A. (2005). *Ludnosc Polski w XX wieku*.

Warszawa: IGIPZ PAN, p.448. Powiat population data are from the 1946 Polish Census (Source: *Powszechny Sumaryczny Spis Ludnosci z dn. 14.II.1946 R. 1947*. Warszawa: Główny Urząd Statystyczny). Data at present-day powiaty level.

Codebook:

Variable: pc\_evac\_ukrainians – Label (and meaning): “% Ukrainians moved to USSR in pop exchange, 1946”

Variable: powiaty\_code – Label (and meaning): “powiaty\_code”

- **war\_destruction\_final.dta**: administrative data on the extent of war-related destruction for rural and urban areas. For rural areas the variable reported by the authorities is the percent of rural buildings affected or destroyed (out of rural buildings available in 1939), and for urban areas, the variable is the percent of volume (in cubic meters) of real estate destroyed in WWII out of all available in 1939. Data at present-day powiaty level. (Source: *Zniszczenia wojenne w zabudowie miast i wsi wg stanu w dniu 1 V 1945*. 1967. Warszawa: Główny Urząd Statystyczny.

Codebook:

Variable: powiaty\_code – Label (and meaning): “powiaty\_code”

Variable: nb\_rural\_destroyed – Label (and meaning): “number of rural buildings destroyed”

Variable: value\_rural\_destruction – Label (and meaning): “value of rural buildings destroyed”

Variable: perc\_destroyed\_rural – Label (and meaning): “percent of rural buildings affected by war”

Variable: perc\_destroyed\_rural\_scale – Label (and meaning): “percent of rural buildings affected by war x % of destruction by building”

Variable: nb\_urban\_destroyed – Label (and meaning): “number of pre-war urban buildings destroyed”

variable\_urban\_destroyed – Label (and meaning): “volume of urban buildings destroyed, in cube meters”

value\_urban\_destruction – Label (and meaning): “value of urban buildings destroyed”

perc\_vol\_urban\_destroyed – Label (and meaning): “percent of total pre-war volume of urban buildings destroyed”

**The Results folder** has two sub-folders: Tables and Figures. These include numbered Tables (in tex or smcl format) and Figures (in pdf and eps format) corresponding to the paper and appendix and that are generated from the code.



### **TABLES IN THE PAPER:**

Table 1: Overview: Polish Population in 1950 (in thousands). This table is taken from official, publicly available Census document, for 1950. We reproduce below the cover of the source document:

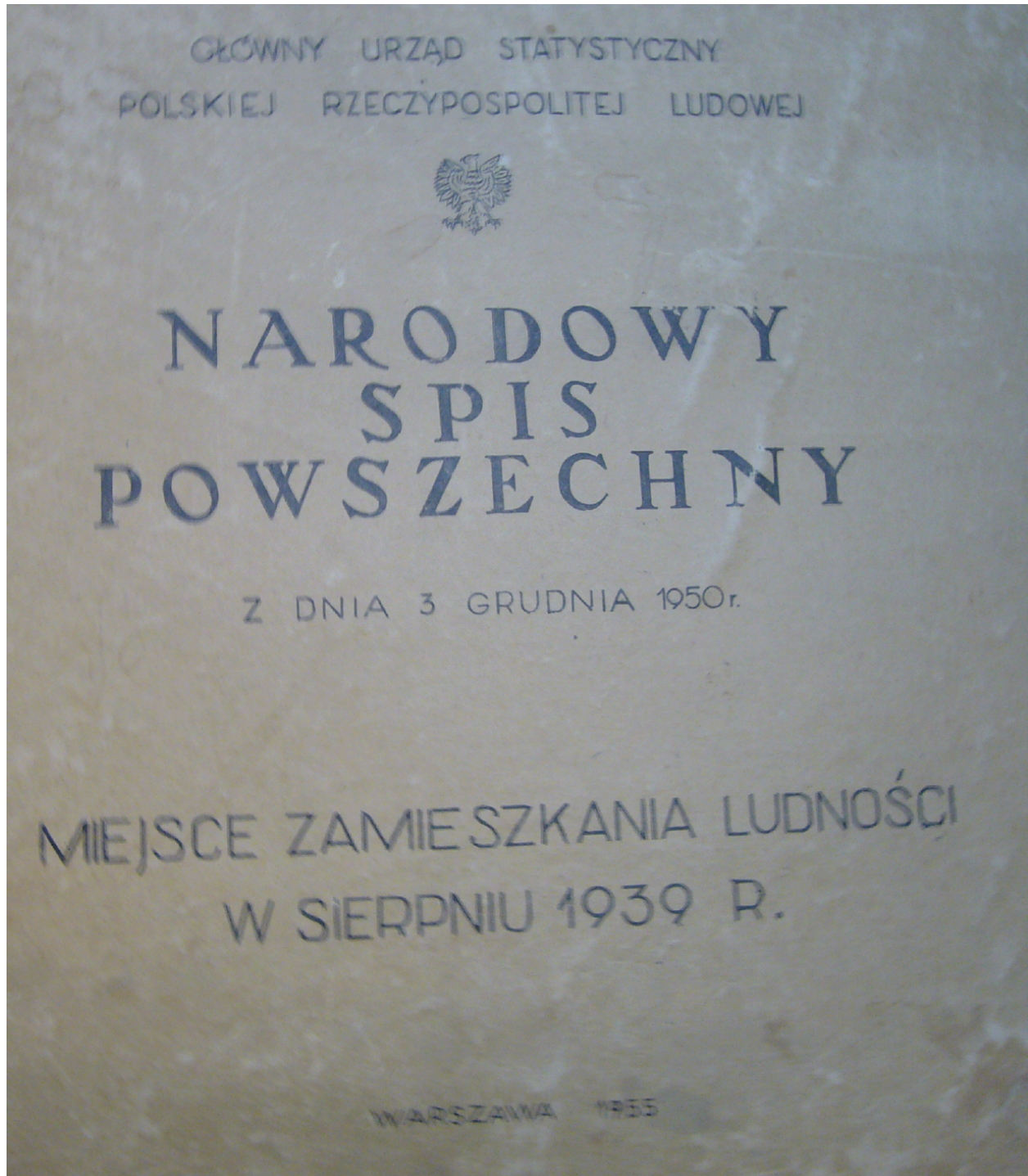




Table 2: Forced Migration from Kresy and Education: Diagnoza Survey Results.  
2.Regressions\_Diagnoza.do: lines 511 to 558.

Table 3: Forced Migration from Kresy and Education in Western Territories: Ancestry Survey.

Panel A: 3.Regressions\_CBOS.do: lines 29 to 50.

Panel B: 3.Regressions\_CBOS.do: lines 199 to 221.

Table 4: Main Results for Kresy Migrants from Rural vs. Urban Areas, and from Ukraine Only.  
3.Regressions\_CBOS.do: lines 227 to 259.

Table 5: Attitudes towards Education and Material Possessions. 2.Regressions\_Diagnoza.do:  
lines 565 to 583.

Table 6: Other Potential Channels: Congestion, Returns to Schooling, Out-Migration,  
Differential Fertility. 2.Regressions\_Diagnoza.do: lines 598 to 642.

## **FIGURES IN THE PAPER:**

Figure 2: Historical and Contemporaneous Patterns in Education. 2.Regressions\_Diagnoza.do:  
lines 30 to 113.

Figure 3: The Kresy Effect on Years of Education, by Birth Cohort. 2.Regressions\_Diagnoza.do:  
lines 124 to 205 (also Figure A.10)

Figure 4: Kresy Border Sample: 1921 Census and 2015 Diagnoza Survey.  
2.Regressions\_Diagnoza.do: lines 213 to 306.

Figure 5: Kresy Border Sample: Ancestry Survey. 3.Regressions\_CBOS.do: lines 714 to 756.

## **TABLES IN APPENDIX:**

Table A1. Summary Statistics for Education Variables.

Panel A: 2.Regressions\_Diagnoza.do: lines 654 to 657.

Panel B: 3.Regressions\_CBOS.do: lines 62 to 68.

Table A2. Summary Statistics for Variables Describing the Origin of Ancestors.

Panels A, B, C: 2.Regressions\_Diagnoza.do: lines 663 to 674.

Panel D: 3.Regressions\_CBOS.do: lines 74 to 80.

Panel E: 3.Regressions\_CBOS.do: lines 186 to 191.

Table A3: Kresy Ancestors and Education – Across Cohorts. 2.Regressions\_Diagnoza.do: lines 680 to 700.

Table A4: Labor Market Outcomes. 2.Regressions\_Diagnoza.do: lines 707 to 725.

Table A5: Ancestry Survey Results (Respondent Level): Weighted. 3.Regressions\_CBOS.do: lines 97 to 119.

Table A6: (Potential) Role of Majority of Kresy Ancestors: Ancestry Survey Results. 3.Regressions\_CBOS.do: lines 125 to 145.

Table A.7: Ancestry Survey Results (Respondent Level): By Generation of Ancestors. 3.Regressions\_CBOS.do: lines 152 to 166.

Table A8: Ancestry Survey Results: Control Group are Respondents with ‘Uniform’ Ancestry. 3.Regressions\_CBOS.do: lines 270 to 284.

Table A9. Confirming the main results in LiTS. 6.Regressions\_LiTS.do: lines 21 to 37.

Table A10: Border Sample from the Diagnoza Survey. 2.Regressions\_Diagnoza.do: lines 731 to 748.

Table A11: Education in the Western Territories: Ancestors Originating Near Kresy Border. 3.Regressions\_CBOS.do: lines 293 to 328.

Table A12: Subsample of Ancestors from Contested Kresy Border Areas. 3.Regressions\_CBOS.do: lines 334 to 374.

Table A13: No Heterogeneous Effects with Respect to Ancestors’ Origin Characteristics. 3.Regressions\_CBOS.do: lines 387 to 465.

Table A14: No Heterogeneous Effects w.r.t. Geographic Features at Ancestors’ Origin. 3.Regressions\_CBOS.do: lines 476 to 526.

Table A15: Robustness of Education Results in LiTS and WWII Victimization. 6.Regressions\_LiTS.do: lines 43 to 67.

Table A16: Education of Kresy Migrants in the Western Territories and Central Poland. 2.Regressions\_Diagnoza.do: lines 754 to 777.

Table A17: Education Today and Historically in Counties of Origin of Ancestors. 3.Regressions\_CBOS.do: lines 533 to 608.

Table A18: Education Difference Between Destination and Origin of Migrants from CP to WT.  
4.Regressions\_Diagnoza\_AnccestorSurvey.do (entire dofile).

Table A19. Household Savings and Individual-Level Insurance in Diagnoza.  
2.Regressions\_Diagnoza.do: lines 783 to 798.

Table A20: Education and Risk-Aversion in the 2016 Life in Transition Survey (LiTS).  
6.Regressions\_LiTS.do: lines 74 to 102.

Table A.21: Education and Smoking (as a Proxy for Discount Rates) in Diagnoza.  
2.Regressions\_Diagnoza.do: lines 803 to 822.

Table A.22: Economic Development at Destination Locations. 2.Regressions\_Diagnoza.do: lines 829 to 892.

Table A.23: Size of Ancestor Communities in each Municipality: Ancestor-Level Data.  
7.Regressions\_CBOS\_community.do (entire dofile).

Table A.24: Further population movements: Diagnoza Data. 2.Regressions\_Diagnoza.do: lines 898 to 932.

Table A.25: Accounting for Missing Ancestor Information in the Ancestry Survey.  
3.Regressions\_CBOS.do: lines 617 to 681.

Table A.26: Accounting for Missing Ancestor Information in the Ancestry Survey - Using Sampling Weights. 3.Regressions\_CBOS.do: lines 687 to 703.

## **FIGURES IN APPENDIX**

Figure A6: The Timing of Arrival of Migrants to the Western Territories. See excel file: timing\_of\_arrival.xlsx (Source: Ministry of Recovered Territories, Document No. 1661, The Central Archives of Modern Records in Warsaw)

Figure A8: Data Quality Check of Diagnoza Survey. 2.Regressions\_Diagnoza.do: lines 319 to 382.

Figure A9: Data Quality Check of our Ancestry Survey – WT Only.  
5.Fig\_App\_DataQuality\_CBOS.do (entire dofile).

Figure A10: Ancestors from Kresy and Education, by Birth Cohort. 2.Regressions\_Diagnoza.do: lines 124 to 205 (also Figure 3).

Figure A11 and A12. (Kresy Border Sample: Geo-climatic Characteristics; and Kresy Border Sample: Crop Suitability): 2.Regressions\_Diagnoza.do: lines 394 to 464.

Figure A.15: Two Alternative Measures of the Share of Autochthons across WT Counties. 2.Regressions\_Diagnoza.do: lines 471 to 475.

Figure A.16: Stated Intent to Emigrate vs. Emigration Rates. 2.Regressions\_Diagnoza.do: lines 484 to 497.

**Footnotes:**

Footnote 19: 3.Regressions\_CBOS.do: line 91