# Online Appendix: "Trading Hard Hats for Combat Helmets"

Yuri M. Zhukov
University of Michigan

August 7, 2015

## Contents

# 1 Eastern Ukraine Violent Events Data

The main article employs a new dataset of violent incidents and territorial control during the armed conflict in Ukraine's Donbas region. The **cross-sectional** version of the dataset includes observations in $N = 3037$ municipalities (naseleni punkty) in two provinces (obasti) of Ukraine: Donetsk and Luhansk. The **panel** version of the dataset includes 53 weekly observations for the same 3,037 municipalities ($N = 3037 \times 53 = 160961$).

The date range (28 February 2014 and 22 February 2015) includes the early protest phase of the conflict immediately following the ouster of former president Viktor Yanukovych (March 2014), the initial violent uprising and seizure of government buildings (April 2014) and the full-scale combat operations that followed the independence referendum of 11 May 2014. The date range ends in the days immediately following the second Minsk ceasefire agreement of 15 February 2015.

The sample of municipalities is universal, encompassing all populated places within these regions, as listed in the National Geospatial-Intelligence Agency's GEOnet Names Server (GNS). I used fuzzy string matching to geocode these violent events to the municipalities in sample, accounting for alternate spellings in Russian, Ukrainian and English. The dataset includes micro-level information on the dates, geographic coordinates, participants, and other details of episodes of political violence and patterns of territorial control distributed across these geographical units. The following appendix provides additional information on the data collection strategy, coding rules, aggregation and summary statistics.

## 1.1 Overview

The violent event data are based on human-assisted machine coding of news reports, press releases and blog posts from Ukrainian, Russian, rebel and international sources. Ukrainian sources include Channel 5, Espresso.tv, Information Resistance, and the newswire services Interfax-Ukraine and Ukrinform. Russian sources include the non-government news websites Gazeta.ru, Lenta.ru and BFM.ru, and the Interfax newswire service. Pro-rebel sources include Rusvesna.su. Also included are the Russian-language edition of Wikipedia, and daily briefings from the OSCE Special Monitoring Mission to Ukraine. For each data source, I created a separate electronic text corpus that contained all incident reports published on the Donbas since February 2014 ($N = 53,754$).

To determine the geographic locations of events mentioned in the reports, I ran an automated geocoding script that identified populated place names referenced in the text, and matched them against the U.S. National Geospatial Intelligence Agency's GeoNames database. I used a one-to-many mapping algorithm, to account for multiple events mentioned in the same report. To identify and correct geocoding errors and double-counts, each list of geocoded locations was referenced against a lookup table of regular errors (e.g. to ensure that 'Donetsk oblast' isn't mis-coded as 'Donetsk city,' and that references to the 'Shakhtar battalion' are not mis-coded as 'Shakhtarsk city'). I also performed manual inspection. Figure 1 shows the resulting spatial distribution of events, by source.

To determine the content of the incident reports, I used a supervised learning algorithm – Support Vector Machine – to classify each event into a series of pre-defined categories. These categories include event type, initiator, target, tactic, and casualties.

The event of primary interest is a *rebel attack*. For a report to be classified as a rebel attack, it must involve a specific act of organized violence initiated by any anti-Kyiv armed group. A specific act of violence is a reference to a single ongoing or recent military operation, act of terrorism, targeted killing, detention, other violent event. Not included in this category are general summaries of war statistics or press statements. Anti-Kyiv groups include any forces explicitly labeled as 'insurgents,' 'rebels,' 'terrorists,' as well as specific formations like the Novorossiya Armed Forces, Donetsk People's Republic (DNR), Lugansk People's Republic (LNR), Vostok Battalion, Oplot, Kal'mus battalion, Bezler band, Zarya battalion, Russian Orthodox Army (RPA), People's Militia of Donbass (NOD), Prizrak battalion, Army of the South East, Don Cossacks, Russian National Unity, Eurasian Youth Union, Yovan Sevic. References to actions by the Russian Armed Forces (mostly in Ukrainian media) are also labeled 'rebel.'

### 1.2 Training set

For each dataset shown in Figure 1, I and a team of research assistants read a randomly-selected training set of 130-600 reports (depending on the size of the corpus), in Russian, Ukrainian and/or English. The author and one research assistant read all training set documents in their original languages. Research assistants not fluent in Russian or Ukrainian read training sets containing the same reports, machine-translated into English.

All research personnel received codebook with instructions for event classification. The sections of this codebook relevant to the analysis in the paper is pasted below.

## Training Set Codebook

For this assignment, you will read media reports from Ukrainian and Russian press (translated into English), and classify them by location, actors and tactics. It will require downloading and installing an open-source statistical software package (R), and running a simple program that displays the text of a media report and asks you a series of questions about its content.

Below is a set of instructions on how to open and analyze these data.

1. ... [technical instructions on downloading and installing R]

2. Open the R application.

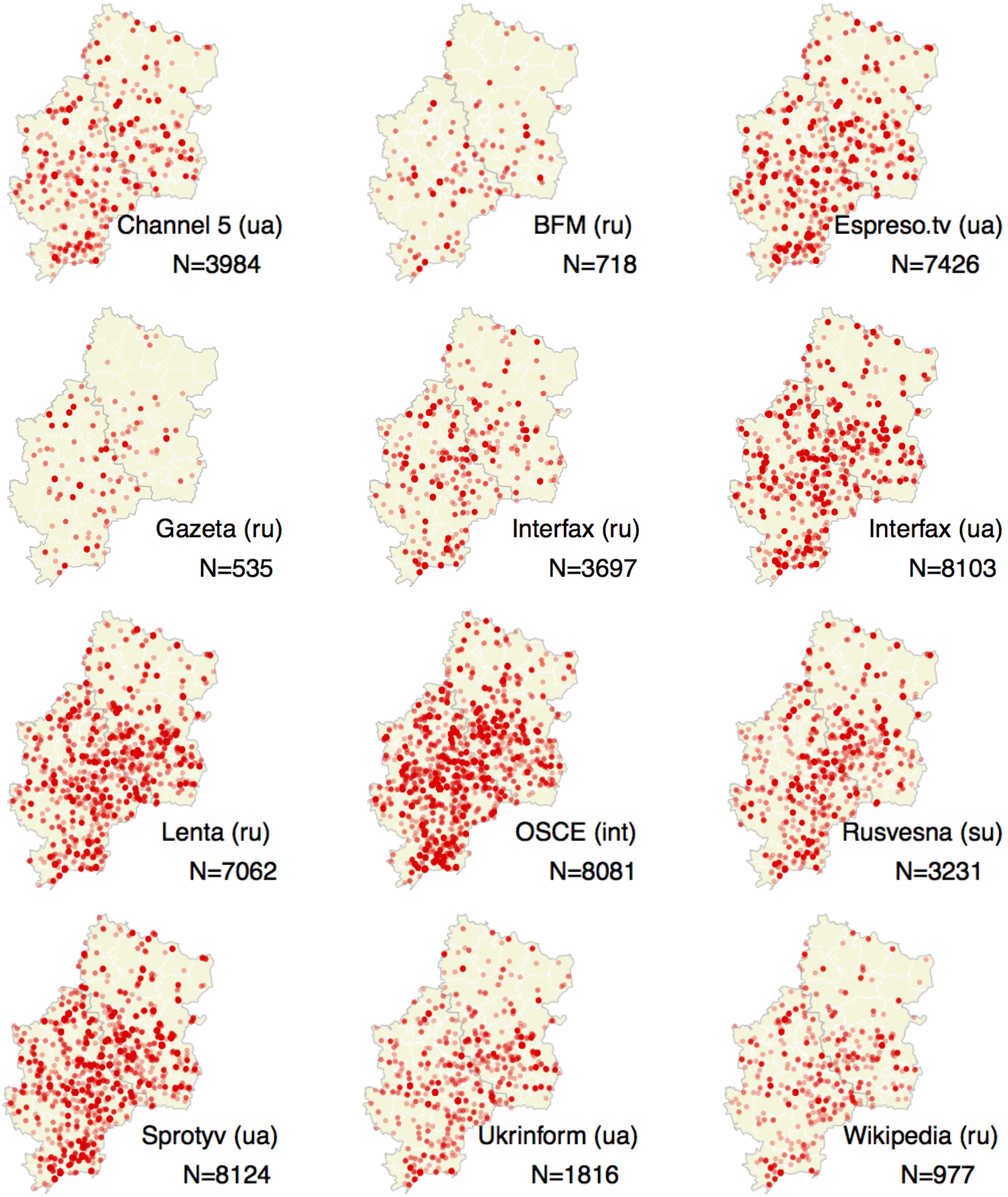
3. In R console, enter the following lines of code at the '>' prompt:

```
setwd("My Directory")
source("databoom.R")
```

   Be sure to change "My Directory" to your actual working directory from (3).

4. You will then be shown a media report on the screen (in Ukrainian, Russian, or translated, sometimes poorly, into English), and will be asked several questions about it's content and tone. Here is an example:

Figure 1: Event locations, by data source.



```
> setwd("My Directory")
> source("databoom.R")

[1] "Summary of the United Army of the South-East (at 9.44 MSK). According to her,
early in the morning Ukrainian gunships attacked Belenky near Kramatorsk.
In Donetsk, the Ukrainian military militia broke through roadblocks in Karlivka,
fights go in Netaylovo. In the Lugansk region are fighting in the village
Zheltoye, in the village Metalist (a suburb of Lugansk) and Stanitsa-Luganskaya,
as there are fights in the Krasnyi Partizan (next to Russian Gukovo). Shelling also
began in Snezhnoye and Saur-Mogila."

1 Violent event? (Y/N)
1 Gibberish / Incomprehensible / Missing text / Foreign Language? (Y/N)
1 INITIATOR: Government/rebel/unknown/civilian/other (G/R/U/C/O)
1 INITIATOR: name of unit? (see list)
1 TARGET: Government/rebel/unknown/civilian/other (G/R/U/C/O)
1 TARGET: name of unit? (see list)
1 TYPE OF ACTION: tactic or weapons system used (see list)
1 CASUALTIES: civilians killed
1 CASUALTIES: civilians wounded
1 CASUALTIES: rebels killed
1 CASUALTIES: rebels wounded
1 CASUALTIES: government killed
1 CASUALTIES: government wounded
1 Comments? (optional)
1 Do you want to re-enter your responses? (Y/N)
1 Additional events in record? (Y/N)
```

Some notes:

- You can enter UPPER CASE or lower case responses (‘Y’ or ‘y’)
- If you feel **you have made an error**, you can re-code the event. Toward the end of the questionnaire, you will be asked if you’d like to re-enter your responses. Answer ‘Y’ (or ‘y’ or ‘yes’ or ‘1’) to that question.
- If there are **multiple records per event** (e.g. ‘attacks happened in villages A, B and C’), enter each record separately. At the end of each questionnaire, you will be asked if there are additional events in the report. Answer ‘Y’ (or ‘y’ or ‘yes’ or ‘1’) to that question.
- Press ‘ESC’ to interrupt the program at any time. To continue, type source(“databoom.R”) in the command window again.
- The program automatically saves your place in the training set. So, if you close R and then restart it, you should start where you left off.

Here is some background on how to respond to these questions.

- `Violent event? (Y/N)`

  Answer ‘Y’ (or ‘y’ or ‘yes’ or ‘1’) if the report describes a specific military operation, rebel attack, or other incident of violence. Answer ‘N’ (or ‘n’ or ‘no’ or ‘0’) if the report describes something else, like a general summary of war statistics (‘in the two months since April, XXX have been killed’), or a press statement (except for statements that describes specific events).

- `Gibberish / Incomprehensible / Missing text / Foreign Language? (Y/N)`

  Self explanatory.

- `INITIATOR: Government/rebel/unknown/civilian/other (G/R/U/C/O)`

  Enter G for government, R for rebel, U for unknown, C for civilian, O for other (e.g. Russian armed forces).

  Note that Ukrainian and Russian sources use different terms for rebels. Russian sources may call them ‘militia’ or ‘guerilla’ or ‘insurgents.’ Ukrainian sources will call them ‘terrorists’ or ‘occupiers.’

- `INITIATOR: name of unit? (see list)`

  If the report specifies the army service, unit, rebel group or volunteer ‘battalion’ carrying out the attack, enter it here.

  The main units on the Ukrainian **government** side are:

  - `ARMY`: Army
    - Air defense
    - Airmobile
    - Armored
    - Infantry
    - Rocket Forces
  - `AF`: Air Force
  - `AIRBORNE`: Airborne/Paratroopers
  - `MARINE`: Marines
  - `MVD`: Interior Ministry
  - `NG`: National Guard
  - `BG`: Border Guard
  - `SBU`: State Security Services

- `VOLUNTEER`: Volunteer battalions

  Note that many of the volunteer battalions are named after cities and regions, so double-check to make sure it's really a battalion. The more prominent battalions are marked with an asterisk (*).

  - Aidar*
  - Artemivsk
  - Azov*
  - Batkivshchyna*
  - Bogdan
  - Chernihiv
  - Dnipro/Dnepr*
  - Donbas*
  - Donetsk-1
  - Donetsk-2
  - Ivano-Frankivsk
  - Kharkiv-1
  - Kharkiv-2
  - Kherson
  - Kirovohrad
  - Kremenchuk
  - Kryvbas*
  - Kyiv-1
  - Kyiv-2
  - Kyivshchyna
  - Kyivska Rus*
  - Luhansk-1
  - Lviv
  - Maidan
  - Mariupol
  - Myrnyi
  - Myrotvorets
  - Poltava
  - Prykarpattya*
  - Rukh Oporu*
  - Shakhtar*
  - Shakhtarsk
  - Shtorm
  - Sich

- ∗ Sicheslav
- ∗ Skif
- ∗ Slobozhanshchyna
- ∗ Sumy
- ∗ Svityaz
- ∗ Svyatyi Mykolai
- ∗ Ternopil
- ∗ Ukraine*
- ∗ Vinnytsia
- ∗ Volyn*
- ∗ Volya*
- ∗ Zaporizhia
- ∗ Zoloti Vorota

The main units on the **rebel** side are:

- `NOVROS`: Novorossiya Armed Forces
- `DNR`: Donetsk People's Republic (DNR)
- `LNR`: Lugansk People's Republic (LNR)
- `VOSTOK`: Vostok Battalion
- `OPLOT`: Oplot
- `KALMUS`: Kal'mus battalion
- `BEZLER`: Bezler band
- `ZARYA`: Zarya battalion
- `RPA`: Russian Orthodox Army (RPA)
- `NOD`: People's Militia of Donbass (NOD)
- `PRIZRAK`: Prizrak battalion
- `AUV`: Army of the South East
- `COSSACK`: Don Cossacks
- `RNE`: Russian National Unity
- `ESM`: Eurasian Youth Union
- `YS`: Yovan Sevic
- `RUSSIA`: Russian Armed Forces

- `TARGET: Government/rebel/unknown/civilian/other (G/R/U/C/O)`

  Enter G for government, R for rebel, U for unknown, C for civilian, O for other (e.g. Russian armed forces).

- `TARGET: name of unit? (see list)`

  If the report specifies the army service, unit, rebel group or volunteer 'battalion' carrying out the attack, enter it here.

- `TYPE OF ACTION: tactic or weapons system used`

  Below is a list of common categories:

  - `AAD`: anti-air defense, Buk, shoulder-fired missiles (Igla, Strela)
  - `AMBUSH`: surprise attack
  - `AIRSTRIKE`: air strike, strategic bombing, helicopter strike
  - `ARMOR`: tank battle or assault
  - `ARREST`: arrest/detention
  - `ARTILLERY`: shelling by field artillery, howitzer, mortar ('mine-thrower')
  - `CONTROL`: establishment/claim of territorial control over population center
  - `KILLING`: assassination, execution, extrajudicial killing, other targeted killing
  - `KILLING_A`: attempted (unsuccessful) assassination, execution, extrajudicial killing, other targeted killing
  - `FIREFIGHT`: any exchange of gunfire with handguns, semi-automatic rifles, automatic rifles, machine guns, rocket-propelled grenades (RPGs)
  - `IED`: improvised explosive device, roadside bomb, landmine, car bomb
  - `PROPERTY`: property descruction
  - `PROTEST`: non-violent protest
  - `PROTEST_V`: violent protest
  - `RAID`: assault/attack, followed by a retreat
  - `RIOT`: violent public disturbance against property or people
  - `ROBBERY`: robbery, burglary, theft
  - `ROCKET`: shelling by artillery rockets like Grad/BM-21, Uragan/BM-27, other Multiple Launch Rocket System (MRLS)
  - `OCCUPY`: occupation of territory or building
  - `STORM`: storming of a building or base
  - `UNKNOWN`

- `CASUALTIES: civilians killed`

  Number of reported civilian deaths.

- CASUALTIES: civilians wounded

  Number of reported civilian wounded or injured.

- CASUALTIES: rebels killed

  Number of reported rebel deaths.

- CASUALTIES: rebels wounded

  Number of reported rebels wounded.

- CASUALTIES: government killed

  Number of reported government/military deaths.

- CASUALTIES: government wounded

  Number of reported government/military wounded.

- Comments? (optional)

  Write any information you feel is relevant, but not captured by the questionnaire.

- Do you want to re-enter your responses? (Y/N)

  Answer ‘Y’ (or ‘y’ or ‘yes’ or ‘1’) if you made a mistake, or omitted something.

- Additional events in record? (Y/N)

  Answer ‘Y’ (or ‘y’ or ‘yes’ or ‘1’) if the report contains multiple events. You will then have a chance to enter info for additional events in the same report.

### 1.3 Intercoder reliability

To account for potential disagreement between coders, at least two sets of eyes read each training set document, including the author and another member of the research team. Inter-coder reliability statistics, reported below, indicate a high and statistically significant level of agreement between coders on the relevant categories, including where coders read the same documents in different languages.

Table 1: Intercoder reliability statistics: Rebel violence.

| | Agree | Fleiss' Kappa | Kendall's W | Krippendorff's Alpha | N |
|---|---|---|---|---|---|
| **5.ua** | 2 coders: both Ukrainian | | | | |
| Violent event: 'Yes' | 82.54 | 0.65*** | 0.84*** | 0.65 (0.51,0.78) | 401 |
| INITIATOR: 'Rebel' | 84.79 | 0.6*** | 0.8*** | 0.59 (0.42,0.77) | 401 |
| **BFM (ru)** | 3 coders: 1 Russian, 2 English | | | | |
| Violent event: 'Yes' | 80.69 | 0.45*** | 0.75*** | 0.46 (0.21,0.65) | 606 |
| INITIATOR: 'Rebel' | 83.17 | 0.41*** | 0.72*** | 0.4 (0.17,0.62) | 606 |
| **Espreso.tv (ua)** | 2 coders: both Ukrainian | | | | |
| Violent event: 'Yes' | 89.14 | 0.78*** | 0.9*** | 0.79 (0.66,0.9) | 313 |
| INITIATOR: 'Rebel' | 85.94 | 0.62*** | 0.81*** | 0.6 (0.44,0.76) | 313 |
| **Gazeta.ru** | 2 coders: 1 Russian, 1 English | | | | |
| Violent event: 'Yes' | 96.2 | 0.75*** | 0.87*** | 0.73 (0.45,0.94) | 500 |
| INITIATOR: 'Rebel' | 87.34 | 0.61*** | 0.8*** | 0.61 (0.42,0.79) | 500 |
| **Interfax (ru)** | 2 coders: 1 Russian, 1 English | | | | |
| Violent event: 'Yes' | 97.33 | 0.65*** | 0.83*** | 0.59 (0.22,0.96) | 500 |
| INITIATOR: 'Rebel' | 88.67 | 0.39*** | 0.7*** | 0.4 (0.14,0.67) | 500 |
| **Interfax (ua)** | 2 coders: 1 Russian, 1 English | | | | |
| Violent event: 'Yes' | 96.45 | 0.73*** | 0.87*** | 0.73 (0.38,0.93) | 301 |
| INITIATOR: 'Rebel' | 88.07 | 0.58*** | 0.8*** | 0.57 (0.37,0.77) | 301 |
| **Lenta.ru** | 3 coders: 1 Russian, 2 English | | | | |
| Violent event: 'Yes' | 94.94 | 0.69*** | 0.84*** | 0.63 (0.26,0.88) | 500 |
| INITIATOR: 'Rebel' | 82.28 | 0.53*** | 0.77*** | 0.51 (0.34,0.69) | 500 |
| **OSCE (int)** | 2 coders: both English | | | | |
| Violent event: 'Yes' | 93.08 | 0.77*** | 0.89*** | 0.77 (0.58,0.9) | 300 |
| INITIATOR: 'Rebel' | 89.23 | 0.63*** | 0.82*** | 0.62 (0.43,0.81) | 300 |
| **Rusvesna.su** | 2 coders: both Russian | | | | |
| Violent event: 'Yes' | 82.92 | 0.55*** | 0.79*** | 0.55 (0.38,0.74) | 281 |
| INITIATOR: 'Rebel' | 80 | 0.51*** | 0.76*** | 0.5 (0.3,0.68) | 281 |
| **Sprotyv (ua)** | 3 coders: 1 Russian, 2 English | | | | |
| Violent event: 'Yes' | 92.01 | 0.59*** | 0.81*** | 0.58 (0.33,0.83) | 511 |
| INITIATOR: 'Rebel' | 96.49 | 0.65*** | 0.83*** | 0.65 (0.26,1) | 511 |
| **Ukrinform** | 3 coders: 1 Russian, 2 English | | | | |
| Violent event: 'Yes' | 76.65 | 0.53*** | 0.77*** | 0.53 (0.38,0.68) | 394 |
| INITIATOR: 'Rebel' | 86.67 | 0.56*** | 0.78*** | 0.56 (0.31,0.77) | 394 |
| **Wikipedia (ru)** | 2 coders: both Russian | | | | |
| Violent event: 'Yes' | 91.54 | 0.64*** | 0.83*** | 0.63 (0.43,0.8) | 130 |
| INITIATOR: 'Rebel' | 78.5 | 0.52*** | 0.76*** | 0.52 (0.32,0.7) | 130 |

*$p < .05$, **$p < .01$, ***$p < .001$

## 1.4 Support vector machine

I used the randomly-selected reference texts in each training set to train a Support Vector Machine (SVM) classifier to predict the categories for all previously unseen corpus texts. The SVM classifies documents by fitting a maximally-separating hyperplane to a feature space, examining combinations of features that best yield separable categories. Formally, the SVM separates data points from each other according to their labels ($y_{it} \in \{-1, 1\}$), and finds maximum marginal distance $\Delta$ between the points labeled $y_{it} = 1$ and $y_{it} = -1$, solving the optimization problem

$$\arg\max_{\Delta,\alpha,\phi} \Delta \text{ s.t. } y_{it}(\alpha + \phi(X_{it})) > \Delta$$

where $y_{it}(\alpha + \phi(X_{it})\beta)$ is a functional margin, $\phi()$ is a function that maps the training data $X$ to a high-dimensional space, and $\mathbf{K}(x_i, x_j) = \phi(x_i)'\phi(x_j)$ is a kernel function. The advantage of the SVM is that it is well-suited to sparse, high-dimensional data, is highly robust, and can handle a low training-to-test data ratio.

I created a separate document-term matrix for each corpus, and ran the SVM classifier separately for each. In the document-term matrix, the rows are documents $d \in \{1, \ldots, D\}$, columns are terms $t \in \{1, \ldots, T\}$, cell entries are weighted term frequencies, and each row vector $\mathbf{y}_d \in \mathbb{R}^T$ represents document $d$ in a $T$-dimensional feature space. Features were weighted by term frequency - inverse document frequency,

$$tf.idf_{dt} = tf_{dt} \log\left(\frac{D}{df_t}\right)$$

where $tf_{dt}$ is term frequency (number of times term appears in $d$), and $df_t$ is document frequency (# documents with term $t$). A high $tf.idf_{dt}$ weight indicates that a term appears a lot in document $d$, but rarely in the corpus.

In the preprocessing stage, I removed HTML tags, control characters, non-alphanumeric characters, capitalization, punctuation and stopwords for all corpora, but ran a stemming algorithm only on English-language texts, so as to preserve inflections in Ukrainian and Russian (i.e. tense, voice, aspect, person, number, gender and case) – which contain important information for differentiating between initiators and targets.

Table 2 shows wordcloud examples for documents classified by SVM as rebel attacks, from Ukrainian and Russian media. The first column shows classification results from training and test sets based on original-language corpora (i.e. Ukrainian or Russian). The second column shows results based on machine-translated corpora in English – wordclouds based on different models, rather than direct translation of Russian or Ukrainian wordcloud text. Font size is proportional to a word's average term frequency-inverse document frequency in each class.

We can make several observations from the information in Table 2. First, there is a high level of similarity between original and English-language results. The same sets of words appear in large type, with slightly larger differences in weights in the Ukrainian source. Second, Ukrainian and Russian sources use very different terms to describe rebels (i.e. terrorist vs. militia), underscoring the utility of separate training sets and classifiers for each corpus.

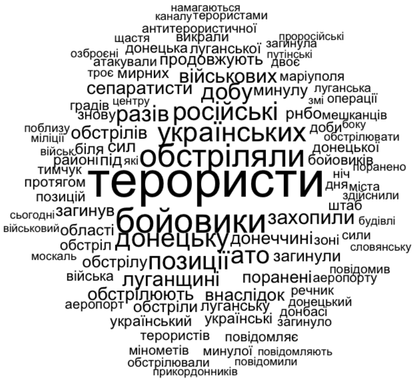
Finally, there is significant variation in the types of rebel attacks reported in Ukrainian and Russian media. Ukrainian media – in this case Channel 5, a television channel owned by Ukraine's President Petro Proshenko – report mainly acts of indiscriminate violence like artillery shelling (*obstrilyaly*, translated as 'fire'). Russian media – in this case BFM – report more selective forms of

Table 2: Wordclouds of SVM classifications by source.

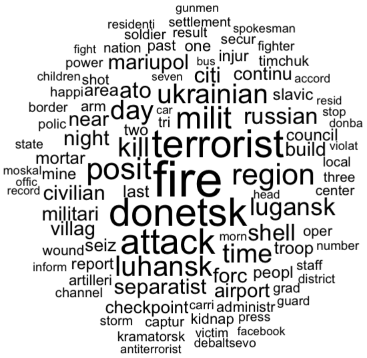
### Rebel-initiated attack

Ukrainian source (Channel 5)

(original)

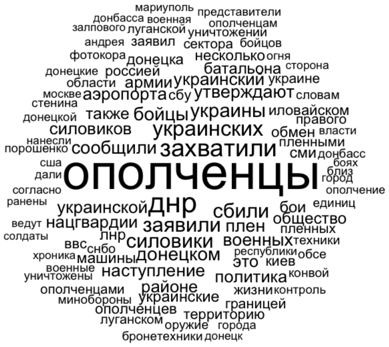


(English)

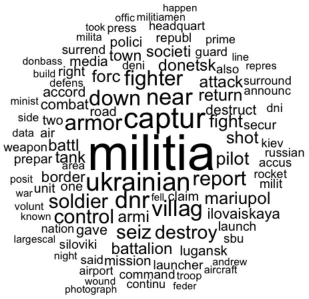


Russian source (BFM)

(original)



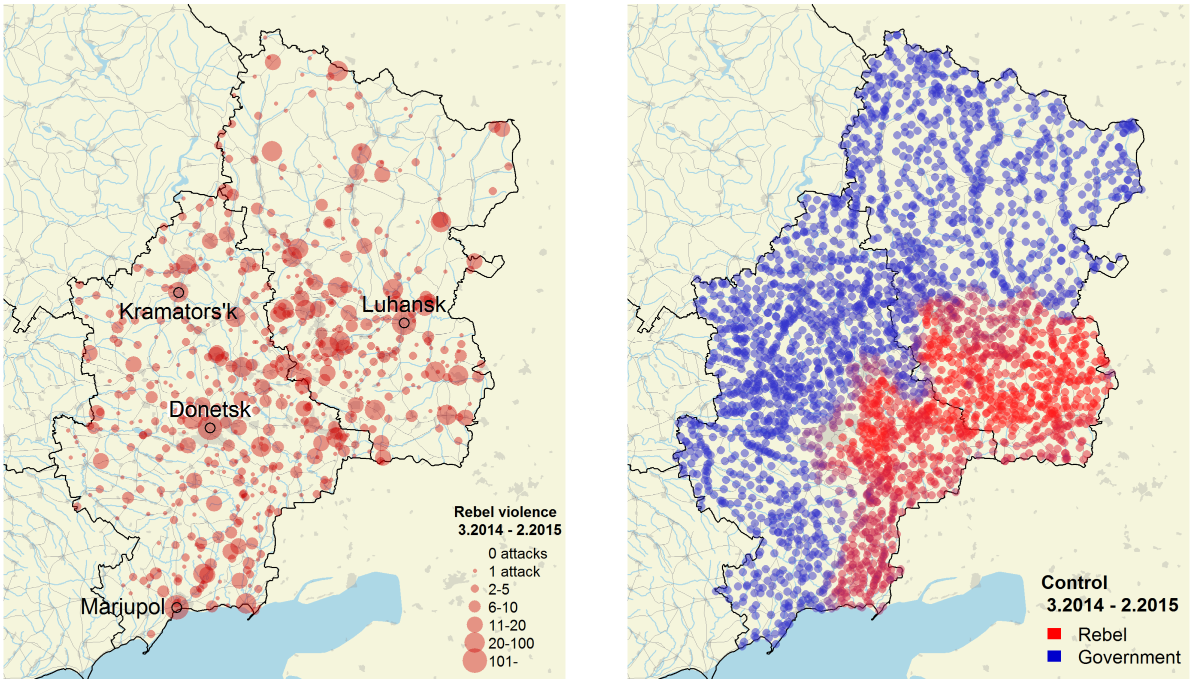
(English)



rebel violence, like the capture of Ukrainian troops. Such reporting biases are common in media coverage of armed conflict. They also highlight the importance of multiple source data collection, in offsetting systematic under-reporting or over-reporting in any one source.

To pool the data across the sources shown in Figure 1, I used a one-a-day filter for each municipality-day. For each of 3,037 unique populated places in Donetsk, Luhansk oblasts, on each day between February 28, 2014 and February 22, 2015, I coded a rebel attack as occurring if at least one of the twelve SVM-classified datasets reported it as occurring. This technique, common in event data research, is designed to eliminate double-counts.

This filter produced 10,567 unique violent events. Because the one-a-day filter implies a maximum of one event per municipality on a given day, I aggregated the atomic-level events to the level of municipality-week and municipality-year. This temporal aggregation permits a measure of intensity, ranging from 0 to 7 on the weekly level, and 0 to 361 on the annual level. Figure 2a visualizes the resulting overall intensity of violence across the 3,037 populated places of the Donbas.



(a) Rebel attacks (SVM-classified)

(b) Rebel control

Figure 2: Outcome variables.

## 1.5 Territorial control

The second outcome variable of interest is territorial control, particularly whether a populated place was under rebel or government control on a given day. For this information, I draw on three sets of sources. First are official daily situation maps publicly released by Ukraine's National Security and Defense Council (RNBO). Second are daily maps assembled by the pro-rebel bloggers 'dragon_first_1' and 'kot_ivanov.' Third are Facebook posts on rebel checkpoint locations prior to June 18, 2014 – the earliest date for RNBO and pro-rebel maps. For RNBO and pro-rebel blog maps – released to the public as high-resolution image files – my research team georeferenced and vectorized each map into spatial polygons. To construct polygons from user-reported checkpoint locations, I used the geographic convex hull of the coordinates of observed checkpoints on each day.

I coded a municipality as being under *rebel control* if, on a given day, it fell inside the rebel control polygons from one of the map collections. I created separate daily indicators for each of the three collections, and two combined indicators – rebel control according to at least one map collection, and control according to both RNBO and the bloggers. Prior to June 18, 2014, the two measure are equal, as there is only one source (Facebook rebel checkpoints.)

Figure 2b displays the distribution of rebel control over the full period of observation. The points are colored according to the proportion of time each municipality spent under rebel control since March 2014, with bright red indicating that a municipality spent almost the full period under rebel control, and blue indicating that a municipality was under government control for most of the period. Purple shades indicate that a municipality spent a significant duration of time under the control of each actor.

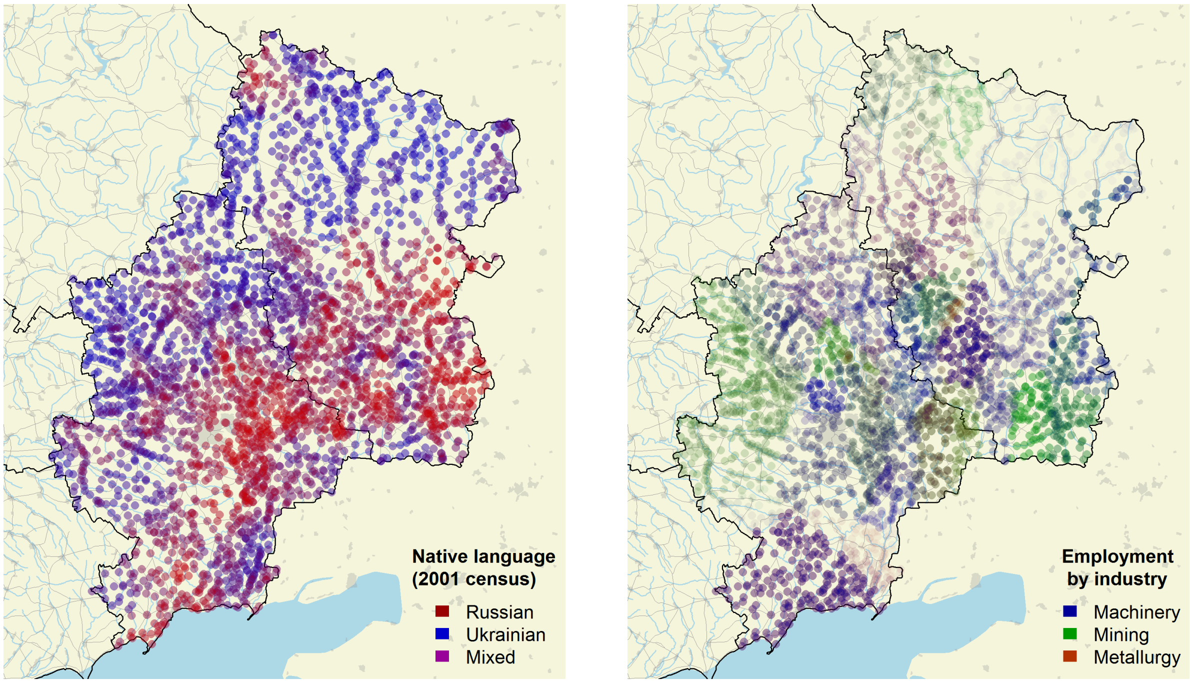
### 1.6 Explanatory variables

The primary source for data on local languages is the 2001 Ukrainian Census (State Committee on Statistics of Ukraine, 2001). For each municipality, I created a measure of *proportion Russian-speaking* (Figure 3a). The census uses respondents' self-reported 'native language' to measure this variable.

For information on the local employment mix, I used Bureau van Dijk's Orbis database (Bureau van Dijk Electronic Publishing, 2015). The databse includes records for 445,399 private and publicly-owned firms in Donetsk and Luhansk provinces, with names, addresses, industry and employment information. For each municipality, I calculated the *proportion of the local labor force employed in machine-building*, *mining* and *metals industries* (Figure 3b). 'Local labor force' is defined as total employment at the 1 percent of companies geographically closest to a given municipality. Industry designations of individual firms (machine-building, mining and metals) are based on European industrial activity classification (NACE) codes and descriptions.

In addition to these covariates of primary theoretical interest, I included a series of other predictors common in subnational conflict research. These include characteristics of the local military geography, like population density (CIESIN and Columbia University, 2005), elevation (U.S. Geological Survey, 1996), forest cover (Loveland et al., 2000), distance to the nearest road (Defense Mapping Agency, 1992), and distance to the Russian border (Global Administrative Areas, 2012). The first of these is a proxy for the number of potential military targets in a municipality. The next three capture the mobility of armed forces, as well as – in the case of elevation and roads – the strategic value of a municipality for artillery firing positions and logistics. The fifth – distance from the Russian border – is a proxy for the availability of Russian military support for the rebels. If military geography drives variation in fighting, one should expect rebel violence and control to be greatest in areas of high population density, elevation and forest cover, and proximate to roads and the Russian border.

I also include data on the prewar political loyalties of the local population. For each municipality, I recorded the percent of the popular vote received by Viktor Yanukovych in the 2010 presidential election (Central Election Commission of Ukraine, 2010). If prewar political loyalties are a driving force in the fighting, one should expect violence and control to be greatest where Yanukovych had



(a) Russian language

(b) Local employment

Figure 3: Explanatory variables.

received more support.

## 2 Variable descriptions for aggregated data

**Black** variable names indicate presence in cross-sectional data. Red variable names indicate presence in panel data only.

### 2.1 Geographic locations and dates

**Case ID (rayon-week)** (`GWID`) Unique identifier for municipality-week observation. Use for sorting data, creation of time lags.

**Time ID (week)** (`WID`) Unique identifier for each week.

**Year** (`YEAR`) Year of observation.

**Month** (`MONTH`) Month of observation.

**Municipality ID** (`geonameid`) Unique identifier for municipality, from GeoNames.

**Municipality Name** (`name`) Name of municipality, from GeoNames.

**Municipality Name (alt)** (`asciiname`) Name of municipality, from GeoNames (alternate).

**Latitude** (`latitude`) Use UTM 37N for projected coordinate system, WGS84 for geographic coordinate system.

**Longitude** (`longitude`) Use UTM 27N for projected coordinate system, WGS84 for geographic coordinate system.

**Province ID** (`GADM_ID_1`) Unique identifier for region.

**Province Name** (`GADM_NAME_1`) Name of province/oblast).

**District ID** (`GADM_ID_2`) Unique identifier for rayon.

**District Name** (`GADM_NAME_2`) Name of district/rayon.

## 2.2 Violence and territorial control

**Rebel violence (count)** (`REB_ALL`) total number of episodes of rebel violence of any type, observed in municipality $i$

**Rebel violence (binary)** (`REB_ALL_b`) $\begin{cases} 1 & \text{if at least one episode of rebel violence} \\ & \text{was observed in municipality } i \\ 0 & \text{otherwise} \end{cases}$

**Rebel violence (weekly count)** (`REB_ALL`) total number of episodes of rebel violence of any type, observed in municipality $i$ during week $t$

**Rebel violence (weekly binary)** (`REB_ALL_b`) $\begin{cases} 1 & \text{if at least one episode of rebel violence} \\ & \text{was observed in municipality } i \text{ during week } t \\ 0 & \text{otherwise} \end{cases}$

**Duration until rebel control** (`SURV_CTR`) number of days until municipality fell to rebel control.

**Censoring dummy** (`SURV_CTR_C`)

**Duration until loss of rebel control** (`SURV_LIB2`) number of days from May 11, 2014 referendum until municipality fell to government forces.

**Censoring dummy** (`SURV_LIB2_C`)

## 2.3 Explanatory variables

**Machine-building** (`IndustryMachinery_Emp_PROP`) proportion of local population employed in machine-building industry.

**Mining** (`IndustryMining_Emp_PROP`) proportion of local population employed in mining industry.

**Metals** (`IndustryMetals_Emp_PROP`) proportion of local population employed in metals industry.

**Percent Russian speaking** (`RUS_CENSUS2001`) percent of census respondent that claim Russian as their 'native language'.

**Russian-speaking majority** (`RUS_MAJ`) $\begin{cases} 1 & \text{if } \texttt{RUS\_CENSUS2001} > 50 \\ 0 & \text{otherwise} \end{cases}$

**Elevation** (dem) elevation of municipality, in meters. Sea level = 0.

**Forest** (FOREST) $\begin{cases} 1 & \text{if land cover classification is forested} \\ 0 & \text{otherwise} \end{cases}$

**Distance to nearest major road** (DIST2ROAD) Euclidean distance (in kilometers) from municipality to the closest primary or secondary road.

**Population density** (POP) population per square kilometer.

**Distance to Russian border** (DIST2RUS) Euclidean distance (in kilometers) from municipality to the closest point on Ukrainian-Russian border.

**Percent Yanukovych vote** (V2010A_YANUKOVYCH) percent of popular vote received by Yanukovych in the 2010 presidential elections.

## 3 Summary statistics

Figure 4: Correlation matrix, municipality-level data.

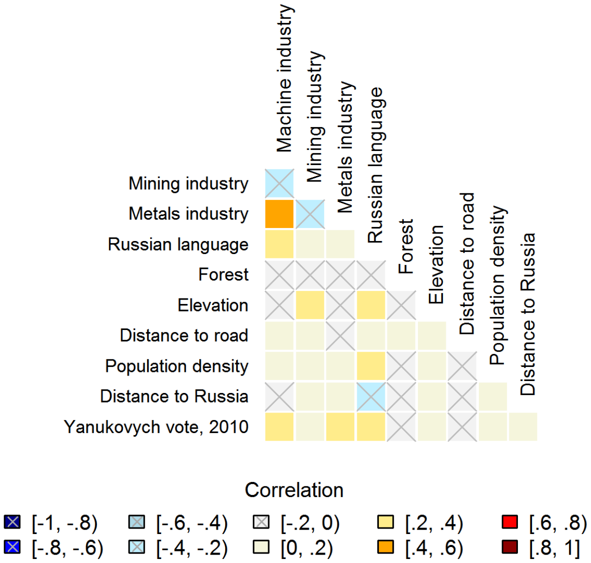


Table 3: SUMMARY STATISTICS. Municipality-level data, unless otherwise noted. Lags excluded. Covariates re-scaled to $x \in [0, 1]$ for all analyses.

| | Obs | Mean | Median | SD | Min | Max |
|---|---|---|---|---|---|---|
| *Outcome variables* | | | | | | |
| Any rebel violence | 3037 | 0.20 | 0.00 | 0.40 | 0.00 | 1.00 |
| Duration until rebel control | 3037 | 676.48 | 113.00 | 748.77 | 40.00 | 1650.00 |
| Duration until rebel control (censoring dummy) | 3037 | 0.63 | 1.00 | 0.48 | 0.00 | 1.00 |
| Duration until loss of rebel control | 3037 | 197.03 | 0.00 | 495.48 | 0.00 | 1576.00 |
| Duration until loss of rebel control (censoring dummy) | 3037 | 0.89 | 1.00 | 0.32 | 0.00 | 1.00 |
| Intensity of rebel violence (municipality-week) | 160,961 | 0.02 | 0.00 | 0.26 | 0.00 | 7.00 |
| *Covariates* | | | | | | |
| Machine industry | 3037 | 0.15 | 0.15 | 0.11 | 0.02 | 0.45 |
| Mining industry | 3037 | 0.20 | 0.17 | 0.20 | 0.00 | 0.78 |
| Metals industry | 3037 | 0.11 | 0.07 | 0.11 | 0.00 | 0.41 |
| Russian language | 3037 | 39.82 | 38.00 | 22.41 | 2.38 | 95.41 |
| Forest cover | 3037 | 0.89 | 1.00 | 0.31 | 0.00 | 1.00 |
| Elevation | 3037 | 143.77 | 143.00 | 64.00 | -2.00 | 343.00 |
| Distance to road | 3037 | 3.34 | 2.66 | 2.87 | 0.00 | 17.23 |
| Population density | 3037 | 188.01 | 36.53 | 362.44 | 12.28 | 2407.10 |
| Distance to Russia | 3037 | 69.51 | 60.64 | 48.85 | 0.02 | 197.05 |
| Yanukovych vote, 2010 | 3037 | 0.75 | 0.76 | 0.05 | 0.61 | 0.84 |

# 4 Bayesian Model Averaging results

Table 4: Bayesian Model Averaging results: Any rebel violence. Posterior inclusion probabilities and model-weighted posterior distributions of coefficients reported in Figure 4a. Posterior inclusion probabilities and model-weighted posterior distributions of coefficients. The core model specification is $y_i = g^{-1}(\alpha + X_i^{\prime}\beta + \epsilon)$, where $g^{-1}(\cdot)$ is the inverse logit link. The level of analysis is municipality-year.

| | $P(\beta \neq 0 \mid y, X)$ | $P(\beta \mid \beta \neq 0, y, X)$ | $sd(\beta \mid \beta \neq 0, y, X)$ |
|---:|---:|---:|---:|
| Machine industry | 97.06 | 1.65 | 0.51 |
| Mining industry | 2.57 | -0.01 | 0.07 |
| Metals industry | 1.73 | 0.00 | 0.08 |
| Russian language | 4.02 | 0.01 | 0.03 |
| Forest | 4.39 | -0.01 | 0.05 |
| Elevation | 1.41 | 0.00 | 0.03 |
| Distance to road | 97.38 | -1.09 | 0.34 |
| Population density | 98.96 | 1.13 | 0.29 |
| Distance to Russia | 1.78 | -0.00 | 0.03 |
| Yanukovych vote, 2010 | 5.12 | 0.02 | 0.09 |
| Machine x Russian | 0.00 | 0.00 | 0.00 |
| Mining x Russian | 0.00 | 0.00 | 0.00 |
| Metals x Russian | 0.00 | 0.00 | 0.00 |

Table 5: Bayesian Model Averaging results: Intensity of rebel violence (weekly). Posterior inclusion probabilities and model-weighted posterior distributions of coefficients reported in Figure 4b. The core model specification is $y_{it} = g^{-1}(y_{i,t-1}\gamma + \mathbf{W}y_{i,t-1} + X_i'\beta + \epsilon)$, where $g^{-1}(\cdot)$ is an inverse quasi-Poisson link, $y_{i,t-1}$ is a one-week time lag of the outcome, $\mathbf{W}y_{i,t-1}$ is a spatial lag of the time lag, and $X_i$ is a matrix of time-invariant covariates.

| | $P(\beta \neq 0 \mid y, X)$ | $P(\beta \mid \beta \neq 0, y, X)$ | $sd(\beta \mid \beta \neq 0, y, X)$ |
|---:|---:|---:|---:|
| Machine industry | 100.00 | 1.72 | 0.18 |
| Mining industry | 2.02 | 0.00 | 0.04 |
| Metals industry | 0.63 | 0.00 | 0.03 |
| Russian language | 0.29 | -0.00 | 0.00 |
| Forest | 0.29 | 0.00 | 0.00 |
| Elevation | 1.14 | 0.00 | 0.02 |
| Distance to road | 100.00 | -1.84 | 0.15 |
| Population density | 100.00 | 0.82 | 0.09 |
| Distance to Russia | 86.57 | -0.29 | 0.14 |
| Yanukovych vote, 2010 | 1.16 | 0.00 | 0.02 |
| Machine x Russian | 0.00 | 0.00 | 0.00 |
| Mining x Russian | 0.00 | 0.00 | 0.00 |
| Metals x Russian | 0.00 | 0.00 | 0.00 |
| $\mathbf{W}$ Rebel Violence$_{t-1}$ | 99.57 | 0.60 | 0.12 |
| Rebel Violence$_{t-1}$ | 100.00 | 0.80 | 0.01 |

Table 6: Bayesian Model Averaging results: Establishment of rebel control. Posterior inclusion probabilities and model-weighted posterior distributions of coefficients reported in Figure 5a. Cox proportional hazard model results reported, $h(t|X_i) = h_0(t)\exp(X_i'\beta)$.

| | $P(\beta \neq 0 \mid y, X)$ | $P(\beta \mid \beta \neq 0, y, X)$ | $sd(\beta \mid \beta \neq 0, y, X)$ |
|---:|---:|---:|---:|
| Machine industry | 100.00 | 5.14 | 0.44 |
| Mining industry | 100.00 | 1.82 | 0.24 |
| Metals industry | 100.00 | -0.36 | 0.34 |
| Russian language | 100.00 | 1.07 | 0.26 |
| Forest | 19.52 | 0.03 | 0.07 |
| Elevation | 93.69 | 0.46 | 0.18 |
| Distance to road | 1.59 | 0.00 | 0.02 |
| Population density | 100.00 | 1.76 | 0.14 |
| Distance to Russia | 100.00 | -1.00 | 0.12 |
| Yanukovych vote, 2010 | 100.00 | 2.26 | 0.15 |
| Machine x Russian | 91.00 | -2.07 | 0.92 |
| Mining x Russian | 73.39 | -0.67 | 0.47 |
| Metals x Russian | 100.00 | -1.95 | 0.47 |

Table 7: BAYESIAN MODEL AVERAGING RESULTS: LOSS OF REBEL CONTROL. Posterior inclusion probabilities and model-weighted posterior distributions of coefficients reported in Figure 5b. Cox proportional hazard model results reported, $h(t|X_i) = h_0(t)\exp(X_i'\beta)$.

| | $P(\beta \neq 0 \mid y, X)$ | $P(\beta \mid \beta \neq 0, y, X)$ | $sd(\beta \mid \beta \neq 0, y, X)$ |
|---:|---:|---:|---:|
| Machine industry | 100.00 | -2.07 | 0.39 |
| Mining industry | 100.00 | -0.41 | 0.17 |
| Metals industry | 100.00 | 1.48 | 0.28 |
| Russian language | 100.00 | 0.02 | 0.09 |
| Forest | 2.74 | -0.00 | 0.02 |
| Elevation | 100.00 | -2.15 | 0.15 |
| Distance to road | 29.50 | -0.11 | 0.19 |
| Population density | 100.00 | -0.90 | 0.17 |
| Distance to Russia | 100.00 | 0.71 | 0.14 |
| Yanukovych vote, 2010 | 21.78 | 0.08 | 0.18 |
| Machine x Russian | 2.62 | 0.02 | 0.17 |
| Mining x Russian | 100.00 | -1.48 | 0.28 |
| Metals x Russian | 0.00 | 0.00 | 0.00 |

# References

Bureau van Dijk Electronic Publishing. 2015. “Orbis database.”.

Central Election Commission of Ukraine. 2010. “Vybory Prezydenta 2010 [Presidential elections 2010].” http://www.cvk.gov.ua/vp_2010/.

CIESIN and Columbia University. 2005. “Gridded Population of the World, Version 3 (GPWv3) Data Collection.” Center for International Earth Science Information Network (CIESIN), Centro Internacional de Agricultura Tropical (CIAT).

Defense Mapping Agency. 1992. “Development of the Digital Chart of the World.” U.S. Government Printing Office.

Global Administrative Areas. 2012. “GADM Database of Global Administrative Areas.”.

Loveland, TR, BC Reed, JF Brown, DO Ohlen, Z Zhu, LWMJ Yang and JW Merchant. 2000. “Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data.” *International Journal of Remote Sensing* 21(6-7):1303–1330.

State Committee on Statistics of Ukraine. 2001. *Ukrainian census 2001.* State Committee on Statistics of Ukraine.

U.S. Geological Survey. 1996. “Global 30-Arc-Second Elevation Data Set.”. Dni.ru.