Codebook and Technical Appendix: "Roads and the Diffusion of Insurgent Violence: The Logistics of Conflict Russia's North Caucasus"

Yuri M. Zhukov
Department of Government, Harvard University

12 November 2011

Contents

1 North Caucasus Violent Events Dataset **2**
1.1 Automated event coding . . . 2
1.2 Event coding rules . . . 3
1.3 Reliability of automated event coding . . . 4
1.4 Road network data . . . 6

2 Coding Rules for Aggregated Data **7**
2.1 Geographic Locations and Dates . . . 7
2.2 Conflict Dynamics . . . 7
2.3 Control Variables . . . 8
2.4 Interactions (pre-coded for transitional model) . . . 8
2.5 Additional variables used in instrumental variable regressions . . . 9

3 Markov Transition Model with Spatial Spline **17**

4 Republic-Level Fixed Effects and Residual Diagnostics **19**

1 North Caucasus Violent Events Dataset

The article employs a new dataset of violent incidents in the Russian North Caucasus. The panel dataset is based on monthly observations across 4,033 municipalities in the seven autonomous republics of the North Caucasus.[1] The sample of villages and towns is universal, encompassing all populated places within these regions, as listed in the National Geospatial-Intelligence Agency's GEOnet Names Server (GNS). For each month between July 2000 and December 2008, the incidence and number of violent events in each village were measured through automated text mining of the independent Memorial Group's "Hronika nasiliya [Chronicle of Violence]" event summaries (Memorial, 2009). Fuzzy string matching was used to geocode these violent events to the municipalities in sample, so as to account for alternate spellings in Russian and a host of local languages. The dataset includes micro-level information on the dates, geographic coordinates, participants, and casualties of episodes of political violence and other forms of unrest distributed across these villages and towns. To capture the connective topology of the study region, a dynamic network dataset was created, with individual villages as the units (or nodes, in network analysis terms), and road distances as the connections (or edges) between them. The following appendix provides a description of the data collection strategy, coding rules, road network estimation, aggregation and summary statistics.

1.1 Automated event coding

Since the original Memorial data are in raw text format, automated text analysis was used to mine the Memorial timeline for dates, locations, actors involved, casualty tolls, and types of incidents. The data extraction strategy I employed differs from traditional automated approaches in several ways. First, dictionary-based event coding algorithms typically use parsing techniques or pattern recognition to code incidents in a "who-does-what-to-whom" format, of which category typologies like VRA and TABARI are prime examples (Schrodt and Gerner, 1994; Schrodt, 2001; Gerner et al., 2002; King and Lowe, 2003; Shellman, 2008). I opted for a somewhat simpler approach based on Boolean association rules and indexing algorithms (Han and Kamber 2001, 230-236; Kim et al. 2001). While not appropriate for all applications, this approach is far more efficient for data-mining highly structured event summaries of the sort that comprise the Memorial timeline – where all entries are of approximately the same length (1-2 sentences) and content (date, location, what happened, who was involved). Second, while various studies have shown that reliance on a single news source in event data analysis can mask important differences in media reporting, most previous uses of events data have relied on only one news source (Reeves et al., 2006; Davenport and Stam, 2006; Davenport and Ball, 2002). The advantage of Memorial's event summaries is that they compile daily reports from international news wires, Russian state and local newspapers, news websites, radio and television broadcasts, and independent reporters, permitting a diverse approach to corpus building which reduces the risk of reporting bias.[2]

From these raw data, the Text Mining (`tm`) package in the R statistical language was used

[1] In alphabetical order, the republics are Adygea, Chechnya, Dagestan, Ingushetia, Kabardino-Balkaria, Karachaevo-Cherkessiya, and North Ossetia. The dataset includes 4,033 villages $\times$102 months = 411,366 village-month observations.

[2] A natural concern with this, like all disaggregated events datasets, is that media are more likely to report incidents located in accessible areas (Raleigh and Hegre, 2009, 234). This problem is addressed somewhat by Memorial's reliance on reports from human rights observers and local independent sources – who benefit from greater access to isolated areas than mass media organization with relatively few local ties.

to assemble a corpus of over 38,000 text documents, perform natural language processing (removing word order and Russian stop words) and create a document-term matrix (Feinerer, 2008; Feinerer et al., 2008). Two custom dictionaries were used to (1) classify events and (2) automatically georeference them against a universal sample of 7,583 cities, towns and villages listed in the U.S. National Geospatial Intelligence Agency's GNS database of populated places in the seven North Caucasus Republics (Dagestan, Chechnya, Ingushetia, North Ossetia, Kabardino-Balkaria, Karachaevo-Cherkesiya, Adygea) and two adjacent majority Russian regions (Stavropol'skiy Kray and Krasnodarskiy Kray).[3] In all, 29,806 unique events were recorded between January 2000 and September 2009, representing as close to a comprehensive sample of state and nonstate violence in Russia as open sources currently permit.[4] The sample was then truncated to the 28,102 events in the seven ethnic republics (4,033 villages), and the time window narrowed to include only the insurgency phase of the conflict (July 2000 - December 2008).

1.2 Event coding rules

Insurgent violence: Event must involve at least one of the following *actors*: nonstate armed groups (NVF), defined by Russian law as any armed group, militia, guerilla or terrorist organization, formed outside the frameworks of existing laws and operating outside the command and control structure of the Russian state; **and** at least one of the following *actions*: terrorist attack, hostage-taking, firefight, bombing, ambush, hit and run attack. Definition does not include events initiated by government forces and non-political acts of violence – such as those resulting from unambiguously criminal activity like burglary and armed robbery.

> *Example:* В ночь на 29 июня в с. Елистанжи Веденского района Чеченской Республики вошел отряд боевиков до 70 человек. Они обстреляли место дислокации роты батальона Юг, а так же место дислокации ПОМ поселкового отдела милиции, который состоит из сотрудников милиции, прикомандированных из других регионов РФ. Боевики убили водителя главы администрации Веденского района, местного жителя. Его вывели из дома и застрелили на улице. Также была обстреляна машина с сотрудниками батальона Юг, которые ехали из с. Агишбатой в с. Елистанжи. В результате погиб сотрудник батальона. К утру боевики ушли из села.
>
> *Translation*: On the night of 29 July a detachment of up to 70 insurgents entered the village of Elistanzhi, Venedo district, Chechen Republic. They opened fire on the positions of a company of the "Yug" Battalion, as well as the positions of the municipal police department, which consists of police officers dispatched from other regions of the Russian Federation. The insurgents killed the driver of the head of

[3]The GNS list was trimmed to remove several types of duplicates: cross-language double-counts (e.g. Grozny and Грозный), within-language double-counts (e.g. Groznyi and Grozny), historical name double counts (Ordzhonikidze and Vladikavkaz), and categorical double counts (e.g. villages listed as both PPL and PPLA). These were removed in several phases, by (a) transliterating all place names into Latin, (b) fuzzy-matching on place names and locations, (c) matching by locations alone for the multiply-named cases, down to finest decimal place included in GNS, and (d) manual inspection for remaining double-counts. In the end, the GNS list contained 4,033 unique villages in the seven republics and 7,584 if one also includes Stavropol'skiy and Krasnodarskiy Kray.

[4]This statistic can be compared with 925 Russian events for the post-Soviet period in the Global Terrorism Database (LaFree and Dugan, 2007) and 14,177 events in the North Caucasus dataset collected by O'Loughlin and Witmer (2011, 183)

Vedeno District, a local resident. He was taken from his home and shot on the street. A car with "Yug" Battalion personnel also came under fire, as it was driving from Agishbatoy village to Elistanzhi. As a result one serviceman was killed. By morning the insurgents had left the village. [Event ID: 34117; Date: 20080629]

Mop-up operations: Event must involve at least one of the following *actors*: Russian Armed Forces, Federal Security Services, Special Forces, Ministry of Internal Affairs, local police, local administration, federal administration; **and** at least one of the following *actions*: cordon-and-sweep operation, or any of the following if simultaneously accompanied by efforts to block or disrupt lines of communication to a village: search and destroy missions, artillery strikes, air strikes, raids, any incidents of government violence that took place as part of a "counterterrorist operation" (KTO), defined in Russian law as a "combination of special-purpose combat operations and other measures involving military hardware, weapons and special means to prevent terrorist acts, neutralize terrorists, provide physical security to persons and facilities, as well as to minimize the consequences of terrorist actions."

Example: Многочисленная группировка силовых структур, включая военнослужащих Министерства обороны, блокировала ст. Вознесеновская Малгобекского района Республики Ингушетия. В станице началась зачистка.

Translation: A large grouping of security forces, including personnel from the Ministry of Defense, has blocked the village of Voznesenovskaya, Malgobek district, Republic of Ingushetia. A sweep operation has begun in the village. [Event ID: 23472; Date: 20070422]

1.3 Reliability of automated event coding

The reliability of content analysis as a data collection method can be separated into three components: (1) consistency, (2) replicability, and (3) accuracy (Weber, 1990, 17). While previous events datasets for the North Caucasus have relied on hand-coding of newspaper articles and incident reports (Lyall, 2009, 2010), there are several advantages to the automated approach employed here. Foremost among these advantages are consistency and replicability – both of which will be critical if the epidemic model is to be meaningfully extended to other cases. Hand-coded event data collection is extremely labor-intensive, involving months of tedious and painstaking work by large teams of undergraduate research assistants (King and Lowe, 2003, 618). Even with experienced coders following well-defined tasks and classification rules, inter-coder reliability can be notoriously low (Mikhaylov and Benoit, 2008). Humans have limited working memories and tend to rely on heuristics, resulting in informal, subjective and ad hoc decisions, not to mention broader risks associated with fatigue, inattention and prior knowledge of hypotheses (Grimmer and King, 2009, 4-5).

Automated coding is no panacea; it also requires a deep working knowledge of the subject matter in the construction of coding rules, and a considerable – though nowhere near as onerous – time investment in data collection, pre-processing and programming. Once these coding rules are established, however, the consistency of machine coding becomes 100% since the program is executing a fixed algorithm (Schrodt and Gerner, 1994). The replicability of the codings across two or more machines – given the same set of rules, actor/action dictionary and corpus of texts – is similarly high. Further, automated coding is not subject to errors induced by the context of an

event, political or cultural biases, fatigue or boredom.

Automated coding methods have been shown to produce results at least as accurate as hand coding but with complete consistency, replicability and more randomness in the errors (Schrodt and Gerner, 1994; King and Lowe, 2003). Whereas bias in the errors can create bias in the results, randomness in errors will tend to attenuate the results, not improve them. The Boolean matching approach uses in this paper capitalizes on the highly structured form of the coded texts – short, two-three sentence incident reports, which have a limited vocabulary and narrow substantive focus. Methods like TABARI and VRA Reader assume little to no structure in the text, thereby opening themselves to additional sources of error. If the assumptions about the nature of the texts are correct, the Boolean matching approach is likely not only to match the coding accuracy of TABARI and VRA Reader but actually exceed it.

The most common types of inaccurate codings in automated events extraction (i.e.: incorrect dates, geocodings or event types) usually occur due to unusually-structured sentences, unrecognized terms not included in the dictionary, or references to historical events (Schrodt, 2001). The first of these was addressed in part by selecting the highly-structured Memorial event summaries as the text corpus (see examples above). The second problem, usually induced through the use of off-the-shelf coding dictionaries, was addressed in the dictionary design phase. Rather than use a pre-existing list of terms that may or may not be in the text, I adopted an *ex-post* dictionary construction technique, in which the system generated a list of most-frequent terms (and permutations thereof) included in the Memorial summaries, and the dictionary lists of relevant political actors, actions, targets and place names were constructed based on this list.[5] This approach enables the fine-tuning of coding rules to the substantive domain of the texts, informed by prior knowledge of what sorts of events can be coded accurately.

While the approach taken here was designed to avoid many of the systematic sources of bias and error common to human coding and certain categories of automated coding, I performed a series of checks to assess the accuracy of the automated event codings and matchings to geographic place names and dates. The first of these was to examine the face validity of the data: does the spatio-temporal distribution of the coded events align with narrative accounts of the evolution of the Caucasus conflict during the period in question (2000-2008). Most analysts of the region – Russian and Western, qualitative and quantitative – have described an increasingly diffuse pattern of violence. A conflict which, until the consolidation of power in Chechnya by the Kadyrov family in 2004-2005, was largely limited to Chechnya, has in recent years spread to neighboring regions, particularly Dagestan, Ingushetia and Kabardino-Balkaria (Malashenko and Trenin, 2002; Kramer, 2004, 2005; Sagramoso, 2007; Souleimanov, 2007; Vendina et al., 2007; **?**; Kuchins et al., 2011). As shown in Figures 1-4, my data largely support these narratives. In 2000-2002, fighting was mostly confined to the Chechen Republic, with occasional rebel incursions into neighboring republics. Following a spike in violence in 2004-2005 (after the assassination of Akhmat Kadyrov), violent attacks became less frequent, but covered a broader swath of territory. Attacks in Ingushetia and Dagestan became more common, while Chechnya became more calm.

An equally important issue was whether some individual events may be mis-coded due to references to historical events, odd phrasings or other problems that could be more easily detected and avoided by a human coder with subject matter expertise. While, due to the many sources of error

[5]Due the complexities of Russian grammar, I did not use stemming as part of natural language processing. This enabled us to distinguish between various grammatical permutations of location and actor names in the construction of the dictionary.

described above, we should be wary of treating any human codings as a "gold standard," a basic comparison of the two types of measures can serve as a useful "sanity check." With this reasoning, I performed the following procedure multiple times: a set of 50 event summaries were randomly selected from the corpus, and hand-coded according to their location, date, and event type. The human event coding rules used were the same as the machine rules outlined in section 1.1. The human codings were then compared against the automated codings, and the level of agreement was calculated as the proportion of event summaries where the two sets of codings were identical. If the level of agreement fell below .9 (more than five disagreements out of 50), the set of events was then manually inspected to determine the source of disagreement.

If the source of disagreement was determined to be systematic, I modified the coding procedure to flag such potential problems for manual inspection with a dummy variable called "`INSPECT`." For instance, in the case of miscodings of paramilitary units' home bases as locations of events – as in "Novgorodskiy OMON" – I set `INSPECT=1` if a location name was followed or preceded by a term representing a political actor in an event summary.[6] To address historical references directly, I set `INSPECT=1` if more than one date, month or year was mentioned in a summary, or if more than one location was mentioned in a summary. This procedure also helped us distinguish between cases where event summaries included references to multiple simultaneous events (e.g. "air strikes were carried out on March 13 in villages A, B and C"), as opposed to event summaries that made references to a single current event and one or more historical events (e.g. "an air strike was carried out on May 15 in village A. This operation marks the first series of air strikes in the area since March 13.") The goal here was to minimize the risk of double-counts and false positives, while avoiding false negatives that would result from mistaking multiple events for historical references.

I then performed a manual inspection of all cases where `INSPECT=1` (originally, 24% of the events), and corrected the codings by hand where deemed necessary. We then selected another 50 event summaries at random, and repeated the entire procedure (a total of 7 times) until the level of agreement exceeded .9 for three consecutive sets of 50. Only after I became convinced that the accuracy of individual event codings approached those of a human subject matter expert ($>$.9), did I aggregate the events to the level of village-month as described in detail below.

1.4 Road network data

To model the spread of insurgent violence as a network process and construct spatially-lagged variables, I measured the accessibility between populated places with an origin-destination (OD) matrix $\mathbf{D}$, in which entries d_{ij} are shortest-path distances (km) between places i and j along the local network of roads.[7] OD matrices have been the subject of a vast literature in urban planning and transportation engineering,[8] but have not – to my knowledge – been widely used in political geography, despite the many advantages of network relative to geodesic distance. Although the calculation of road network distances is far more computationally intensive than their planar or spherical counterparts, OD matrices can be estimated with Python scripts, Java programs or ArcGIS extensions (Steenberghen et al., 2009). For my data, I used a geoprocessing script that relies on

[6] This procedure was performed through string operations on the original text, rather than the "bag of words" representation of the text following the removal of stop words and the discarding of word order.

[7] Geospatial data on the road network in the Caucasus, as well as other spatial data of interest (population density, elevation, land cover), were taken from the U.S. Geological Survey's Global GIS Database (Hearn et al., 2005)

[8] See Cherkassky et al. (1996); Zhan and Noon (1998)

ArcMap's Network Analyst engine.[9] The result is a dense $7,583 \times 7,583$ matrix, with 57,517,056 shortest-path road distances between villages. Used in the preceding analysis is a $4,033 \times 4,033$ submatrix, which covers only the seven autonomous republics.

Valued network data are often dichotomized for ease of interpretation (by distinguishing between neighbors and non-neighbors) and computational efficiency (the valued matrix is over 3GB in size). However, dichotomization also risks the loss of potentially important information (Thomas and Blitzstein, 2009). Because the epidemiological model assumes continuous measures of network distance, I avoided the use of dichotomizing cutpoints and preserved the continuous distance data. A visual representation of the road network structure is provided in Figure 5.

2 Coding Rules for Aggregated Data

2.1 Geographic Locations and Dates

Case ID (municipality-month) (`TSID`) Unique identifier for municipality-month observation. Use for sorting data, creation of time lags.

Case ID (month-municipality) (`TSID2`) Unique identifier for month-municipality observation. Use for sorting data, creation of spatial lags.

Date (`YRMO`) Date of observation, in format YYYYMM.

Place ID (`CID`) Unique identifier for city, town, village or populated place.

Place Name (`NAME`) Name of city, town, village or populated place, from GeoNames (2009).

Region ID (`OBLAST_ID`) Unique identifier for region (*oblast*).

Region Name (`OBLAST_NAME`) Name of region (*oblast*).

District ID (`RID`) Unique identifier for district (*rayon*).

District Name (`RAYONS_NAME`) Name of district (*rayon*).

Latitude (`LAT`) Use UTM 38N or UTM 39N for projected coordinate system, WGS84 for geographic coordinate system.

Longitude (`LONG`) Use UTM 38N or UTM 39N for projected coordinate system, WGS84 for geographic coordinate system.

2.2 Conflict Dynamics

Insurgent Violence

Insurgent attack (count) (`REBEL`) number of episodes of insurgent violence, as defined above, observed in municipality i during month t.

[9] A 5km buffer was used to determine which villages were connected to the road network. For municipalities further off the grid (17%), the script calculated the geodesic distance to the closest on-road village, and used the latter's distance values, penalized by the additional travel-to-road distance.

Insurgent attack (binary) (`REBEL.b`) $\begin{cases} 1 & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t \\ 0 & \text{otherwise} \end{cases}$

Insurgent attack (count, time lagged) (`REBEL.t1`) number of episodes of insurgent violence, as defined above, observed in municipality i during month $t-1$.

Insurgent attack (binary, time lagged) (`REBEL.b.t1`) $\begin{cases} 1 & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$

Distance to nearest recent insurgent attack (geodesic network) (`D.REBEL.GEO.t1`) measured as $\min(w_i \text{Insurgent Violence}_{j \neq i, t-1})$, where w_i is a vector of geodesic distances between village i and all other villages j.

Distance to nearest recent insurgent attack (road network) (`D.REBEL.ROAD.t1`) measured as $\min(w_i \text{Insurgent Violence}_{j \neq i, t-1})$, where w_i is a vector of road distances between village i and all other villages j.

Government Actions

Mop-up operations (count) (`GOV_MOP`) number of government-initiated mop-up operations, as defined above, observed in municipality i during month t.

Mop-up operations (binary) (`GOV_MOP.b`) $\begin{cases} 1 & \text{if at least one mop-up operation was observed in village } i \text{ during month } t \\ 0 & \text{otherwise} \end{cases}$

Mop-up operations (count, time lagged) (`GOV_MOP.t1`) number of government-initiated mop-up operations, as defined above, observed in municipality i during month $t-1$.

Mop-up operations (binary, time lagged) (`GOV_MOP.b.t1`) $\begin{cases} 1 & \text{if at least one mop-up operation was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$

2.3 Control Variables

Accessibility by road (`ROAD_5KM`) $\begin{cases} 1 & \text{if village } i \text{ is located within 5 km of a major road} \\ 0 & \text{otherwise} \end{cases}$

Population density (`POP`) Population per square kilometer.

Elevation (`ELEVATION`) In meters. Sea level = 0.

2.4 Interactions (pre-coded for transitional model)

Insurgent Attack × Dist. to Attack (`R_D.REBEL.GEO.t1`) $\begin{cases} \texttt{D.REBEL.GEO.t1} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$

Insurgent Attack × Dist. to Attack (R_D.REBEL.ROAD.t1)
$$\begin{cases} \texttt{D.REBEL.ROAD.t1} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$$

Insurgent Attack × Population Density (R_ROAD_5KM)
$$\begin{cases} \texttt{ROAD_5KM} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$$

Insurgent Attack × Mop-Up (R_GOV_MOP.b)
$$\begin{cases} \texttt{GOV_MOP.b} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$$

Insurgent Attack × Population Density (R_POP)
$$\begin{cases} \texttt{POP} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$$

Insurgent Attack × Elevation (R_ELEVATION)
$$\begin{cases} \texttt{ELEVATION} & \text{if at least one episode of insurgent violence was observed in village } i \text{ during month } t-1 \\ 0 & \text{otherwise} \end{cases}$$

2.5 Additional variables used in instrumental variable regressions

Elections (ELECTIONS)
$$\begin{cases} 1 & \text{if month } t \text{ overlaps with federal election cycles for the State Duma (quadrennial, Oct-Dec) or Presidency (quadrennial, Jan-Mar)} \\ 0 & \text{otherwise} \end{cases}$$

Distance to military base (DIST_KM) measured as d_{ik}, the road distance (in kilometers) between village i and military facility k

Table 1: **Summary statistics for aggregated data (village-month level) and list of sources**

Variable Description	Variable Name	Min	Median	Mean	Max	NAs	Source
Insurgent Attack (binary)	REBEL.b	0	0	0.004	1	0	GeoNames (2009); Memorial (2009)
Insurgent Attack (binary, time lagged)	REBEL.b.t1	0	0	0.004	1	4033	GeoNames (2009); Memorial (2009)
Distance to Nearest Attack (geodesic)	D.REBEL.GEO.t1	0	70.87	142.64	869.77	4033	GeoNames (2009); Memorial (2009)
Distance to Nearest Attack (road)	D.REBEL.ROAD.t1	0	106.87	178.32	1084.2	4033	GeoNames (2009); Memorial (2009)
Mop-Up (binary, time lagged)	GOV_MOP.b.t1	0	0	0.003	1	4033	GeoNames (2009); Memorial (2009)
Accessibility by road (5km)	ROAD_5KM	0	1	0.83	1	0	GeoNames (2009); Memorial (2009)
Population Density	POP	0	17	179.1	11576	612	GeoNames (2009); Goskomstat (2009)
Elevation	ELEVATION	-31	678	827.4	2818	0	GeoNames (2009); Hearn et al. (2005)
Insurgent Attack × Dist. to Nearest Attack (geo)	R_D.REBEL.GEO.t1	0	0	0.098	532	4033	GeoNames (2009); Memorial (2009)
Insurgent Attack × Dist. to Nearest Attack (road)	R_D.REBEL.ROAD.t1	0	0	0.206	610.23	4033	GeoNames (2009); Memorial (2009)
Insurgent Attack × Mop-Up	R_GOV_MOP.b.t1	0	0	0.001	1	4033	GeoNames (2009); Memorial (2009)
Insurgent Attack × Road Accessibility	R_ROAD_5KM	0	0	0.004	1	4033	GeoNames (2009); Memorial (2009)
Insurgent Attack × Population Density	R_POP	0	0	6.254	11576	4639	GeoNames (2009); Goskomstat (2009)
Insurgent Attack × Elevation	R_ELEVATION	-27	0	1.463	2146	4033	GeoNames (2009); Hearn et al. (2005)
Elections	ELECTIONS	0	0	0.137	1	0	GeoNames (2009); Memorial (2009)
Distance to Nearest Military Base	DIST_KM	0.009	63.54	68.14	248.6	1020	GeoNames (2009); Janko (2009)
Mop-Up (instrumented)	IV_MOPUP	0	0.002	0.003	0.884	5649	GeoNames (2009); Memorial (2009)
Insurgent Attack × Mop-Up (instrumented)	R_IV_MOPUP	0	0	0.001	0.884	5649	GeoNames (2009); Memorial (2009)
Mop-Up (instrumented, geo)	IV_MOPUP.g	0	0.001	0.003	0.888	5649	GeoNames (2009); Memorial (2009)
Insurgent Attack × Mop-Up (instrumented, geo)	R_IV_MOPUP.g	0	0	0.001	0.888	5649	GeoNames (2009); Memorial (2009)
Mop-Up (instrumented, road)	IV_MOPUP.r	0	0.001	0.003	0.899	5649	GeoNames (2009); Memorial (2009)
Insurgent Attack × Mop-Up (instrumented, road)	R_IV_MOPUP.r	0	0	0.001	0.899	5649	GeoNames (2009); Memorial (2009)

Table 2: **Correlation Matrix**

	REBEL.b	REBEL.b.t1	D.REBEL.GEO.t1	D.REBEL.ROAD.t1	ROAD_5KM	POP	ELEVATION	GOV_MOP.b.t1	R_D.REBEL.GEO.t1	R_D.REBEL.ROAD.t1	R_ROAD_5KM	R_POP	R_ELEVATION	R_GOV_MOP.b.t1	ELECTIONS	DIST_KM	IV_MOPUP	R_IV_MOPUP	IV_MOPUP.g	R_IV_MOPUP.g	IV_MOPUP.r	R_IV_MOPUP.r
REBEL.b	1.0	0.3	-0.0	-0.0	0.0	0.1	-0.0	0.2	0.1	0.2	0.3	0.2	0.1	0.2	0.0	-0.0	0.3	0.3	0.3	0.3	0.3	0.3
REBEL.b.t1	0.3	1.0	-0.0	-0.0	0.0	0.1	-0.0	0.3	0.5	0.6	1.0	0.5	0.7	0.5	0.0	-0.0	0.7	0.8	0.7	0.8	0.7	0.8
D.REBEL.GEO.t1	-0.0	-0.0	1.0	1.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.0
D.REBEL.ROAD.t1	-0.0	-0.0	1.0	1.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.0
ROAD_5KM	0.0	0.0	0.0	-0.0	1.0	0.1	-0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.1	0.0	0.1	0.0	0.1	0.0
POP	0.1	0.1	-0.0	-0.0	0.1	1.0	-0.1	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.0	-0.1	0.4	0.2	0.4	0.2	0.4	0.2
ELEVATION	-0.0	-0.0	-0.0	0.0	-0.2	-0.1	1.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.1	-0.1	-0.0	-0.1	-0.0	-0.1	-0.0
GOV_MOP.b.t1	0.2	0.3	-0.0	-0.0	0.0	0.1	-0.0	1.0	0.1	0.2	0.3	0.2	0.2	0.6	-0.0	-0.0	0.3	0.3	0.3	0.3	0.3	0.3
R_D.REBEL.GEO.t1	0.1	0.5	-0.0	-0.0	0.0	0.1	-0.0	0.1	1.0	0.6	0.5	0.3	0.4	0.2	-0.0	-0.0	0.4	0.4	0.3	0.3	0.3	0.4
R_D.REBEL.ROAD.t1	0.2	0.6	-0.0	-0.0	0.0	0.1	-0.0	0.2	0.6	1.0	0.5	0.5	0.4	0.4	-0.0	-0.0	0.5	0.6	0.5	0.5	0.4	0.5
R_ROAD_5KM	0.3	1.0	-0.0	-0.0	0.0	0.1	-0.0	0.3	0.5	0.5	1.0	0.5	0.7	0.5	0.0	-0.0	0.7	0.8	0.7	0.8	0.7	0.8
R_POP	0.2	0.5	-0.0	-0.0	0.0	0.3	-0.0	0.2	0.3	0.5	0.5	1.0	0.2	0.4	0.0	-0.0	0.9	0.9	0.9	0.9	0.9	0.9
R_ELEVATION	0.1	0.7	-0.0	-0.0	0.0	0.1	-0.0	0.2	0.4	0.4	0.7	0.2	1.0	0.3	-0.0	-0.0	0.4	0.4	0.4	0.4	0.4	0.4
R_GOV_MOP.b.t1	0.2	0.5	-0.0	-0.0	0.0	0.1	-0.0	0.6	0.2	0.4	0.5	0.4	0.3	1.0	-0.0	-0.0	0.5	0.5	0.5	0.5	0.5	0.5
ELECTIONS	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	1.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
DIST_KM	-0.0	-0.0	-0.0	0.0	-0.0	-0.1	0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	1.0	-0.1	-0.0	-0.1	-0.0	-0.1	-0.0
IV_MOPUP	0.3	0.7	-0.0	-0.0	0.1	0.4	-0.1	0.3	0.4	0.5	0.7	0.9	0.4	0.5	-0.0	-0.1	1.0	1.0	1.0	1.0	1.0	1.0
R_IV_MOPUP	0.3	0.8	-0.0	-0.0	0.0	0.2	-0.0	0.3	0.4	0.6	0.8	0.9	0.4	0.5	-0.0	-0.0	1.0	1.0	1.0	1.0	0.9	1.0
IV_MOPUP.g	0.3	0.7	-0.1	-0.1	0.1	0.4	-0.1	0.3	0.3	0.5	0.7	0.9	0.4	0.5	-0.0	-0.1	1.0	1.0	1.0	1.0	1.0	0.9
R_IV_MOPUP.g	0.3	0.8	-0.0	-0.0	0.0	0.2	-0.0	0.3	0.3	0.5	0.8	0.9	0.4	0.5	-0.0	-0.0	1.0	1.0	1.0	1.0	0.9	1.0
IV_MOPUP.r	0.3	0.7	-0.1	-0.1	0.1	0.4	-0.1	0.3	0.3	0.4	0.7	0.9	0.4	0.5	-0.0	-0.1	1.0	0.9	1.0	0.9	1.0	1.0
R_IV_MOPUP.r	0.3	0.8	-0.0	-0.0	0.0	0.2	-0.0	0.3	0.4	0.5	0.8	0.9	0.4	0.5	-0.0	-0.0	1.0	1.0	0.9	1.0	1.0	1.0

Figure 1: **Spatio-temporal distribution of insurgent attacks**, July 2000 - Dec 2002

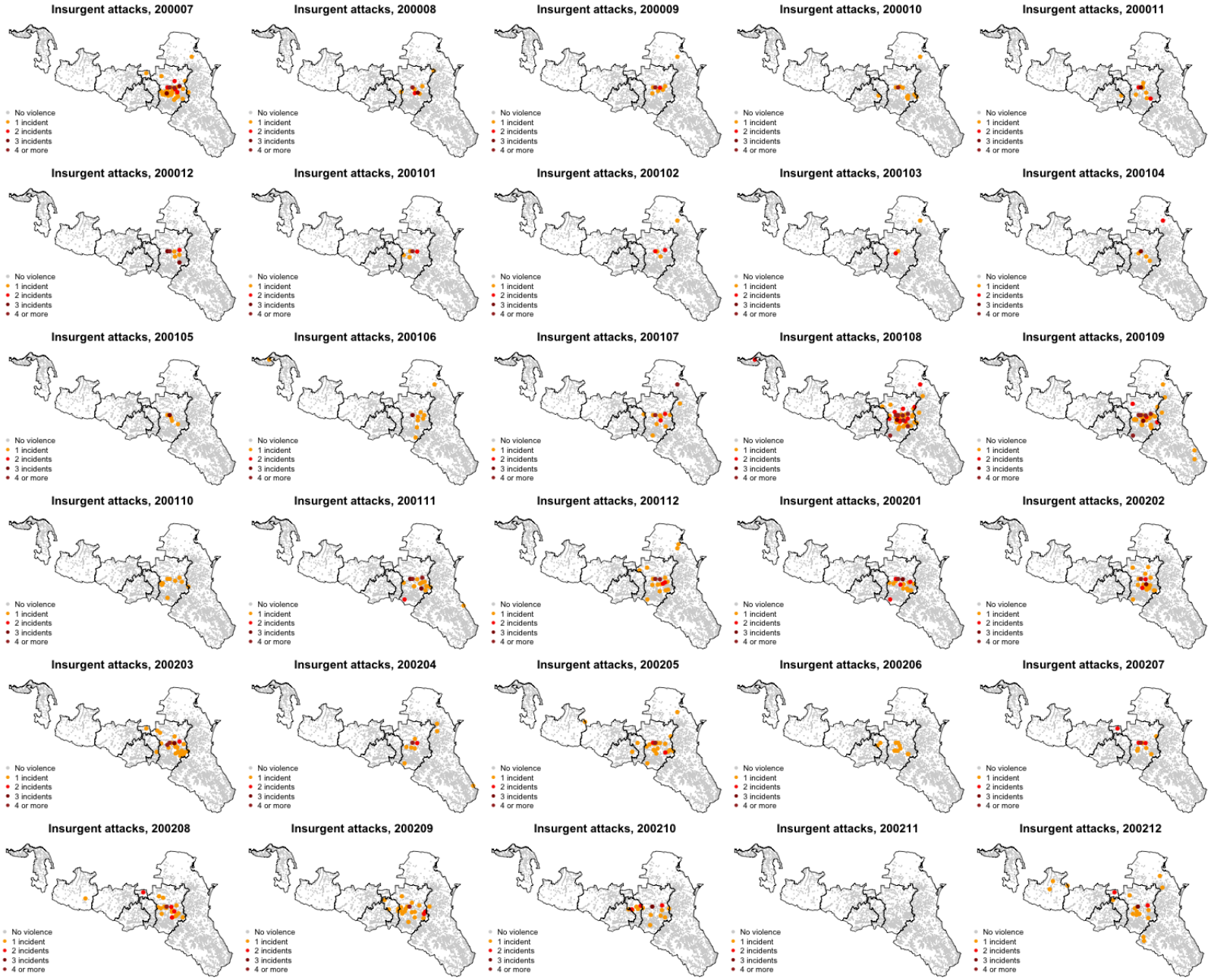


Figure 2: **Spatio-temporal distribution of insurgent attacks**, Jan 2003 - Jun 2005

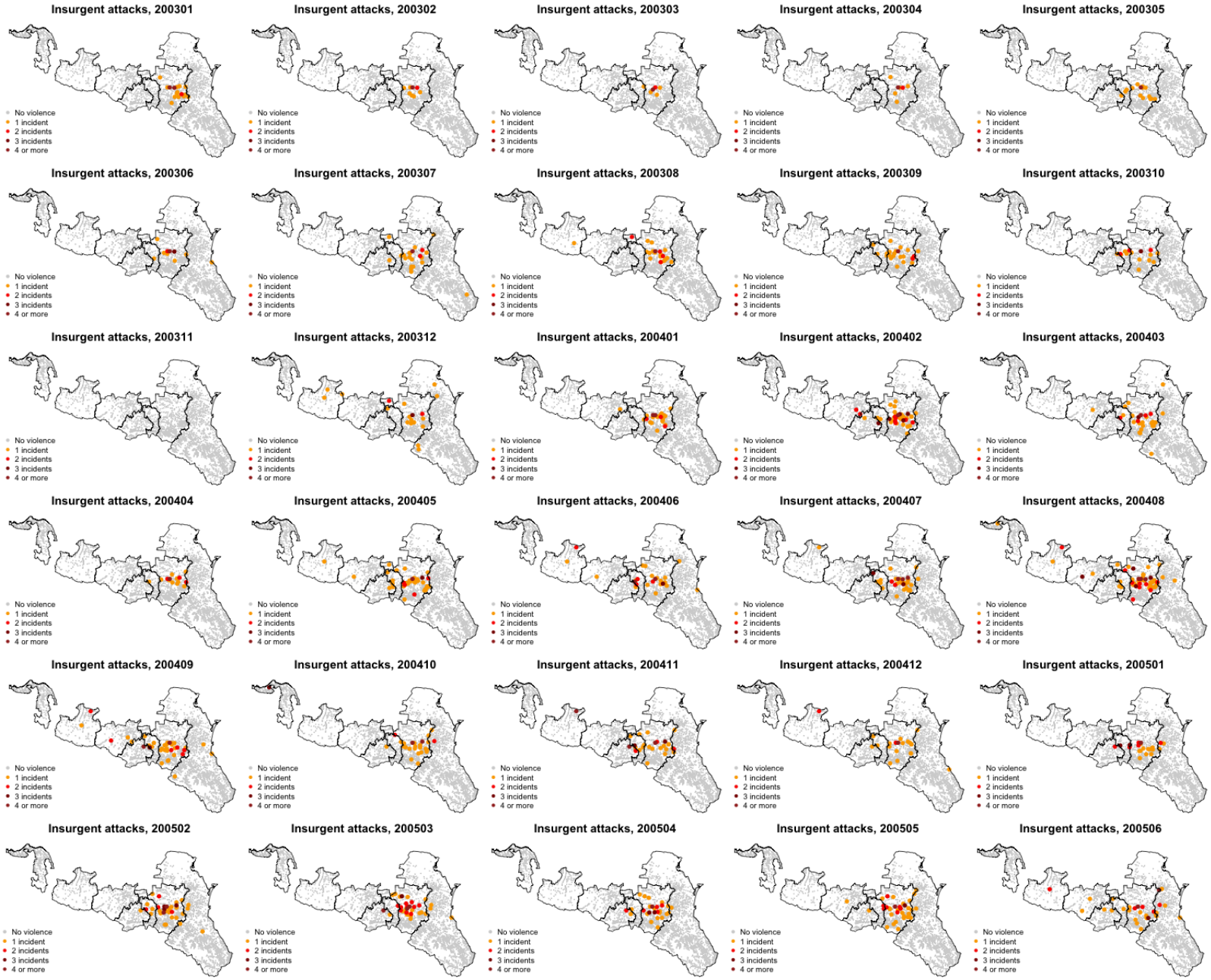


Figure 3: **Spatio-temporal distribution of insurgent attacks**, July 2005 - Dec 2007

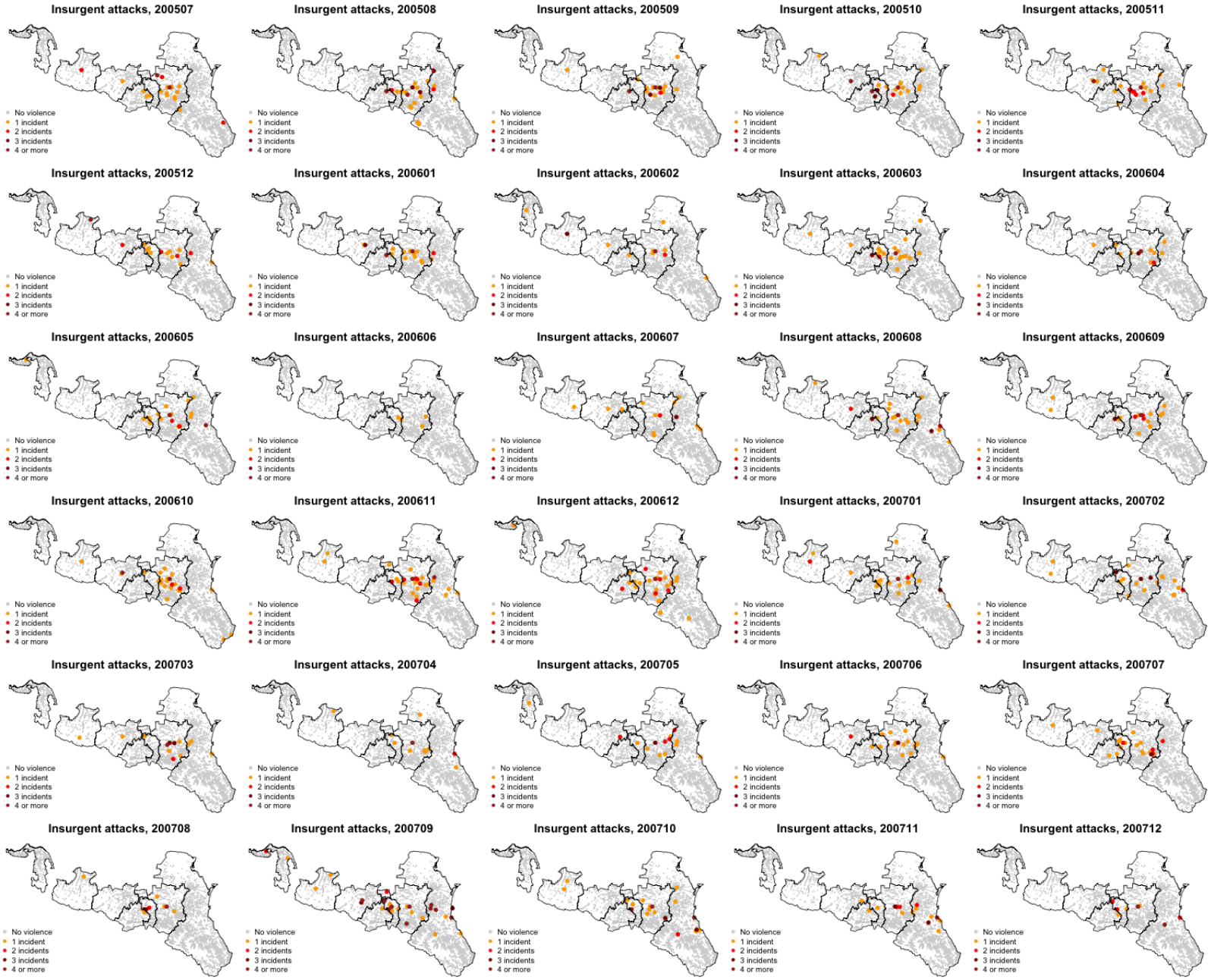


Figure 4: **Spatio-temporal distribution of insurgent attacks**, Jan 2008 - Dec 2008

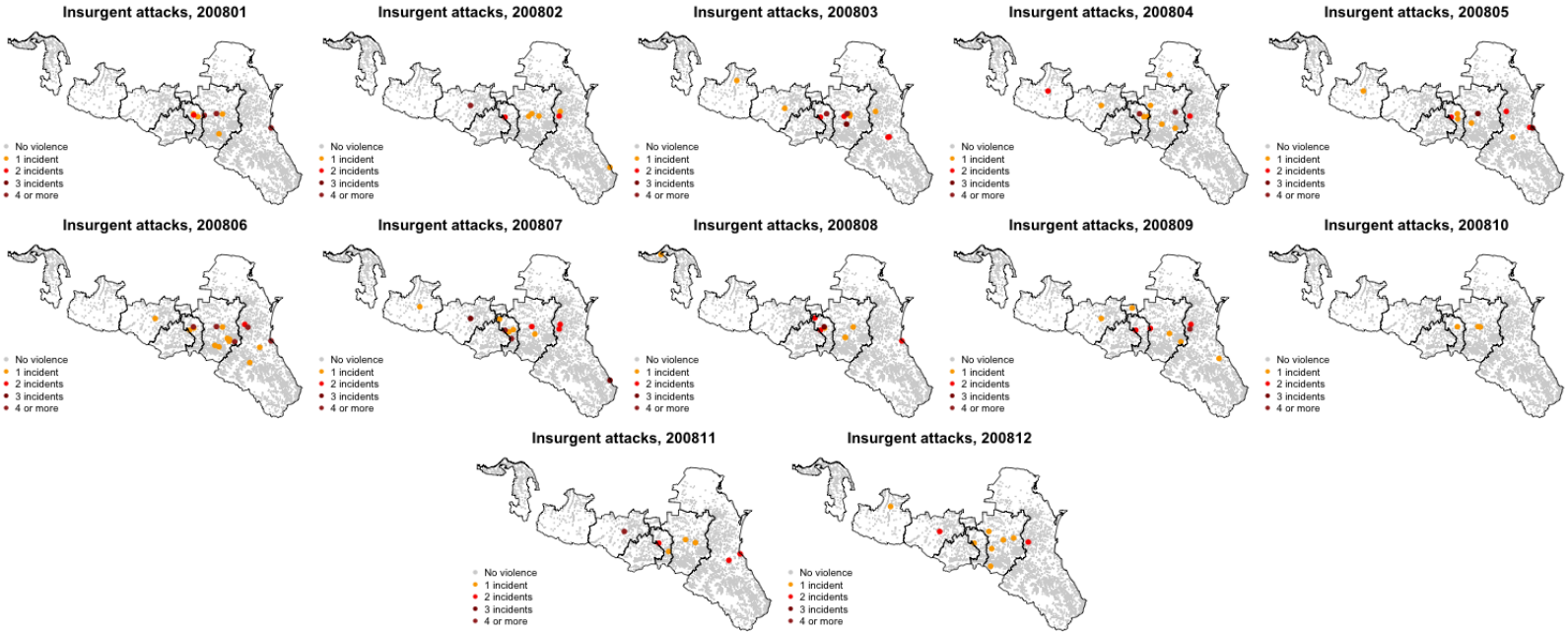
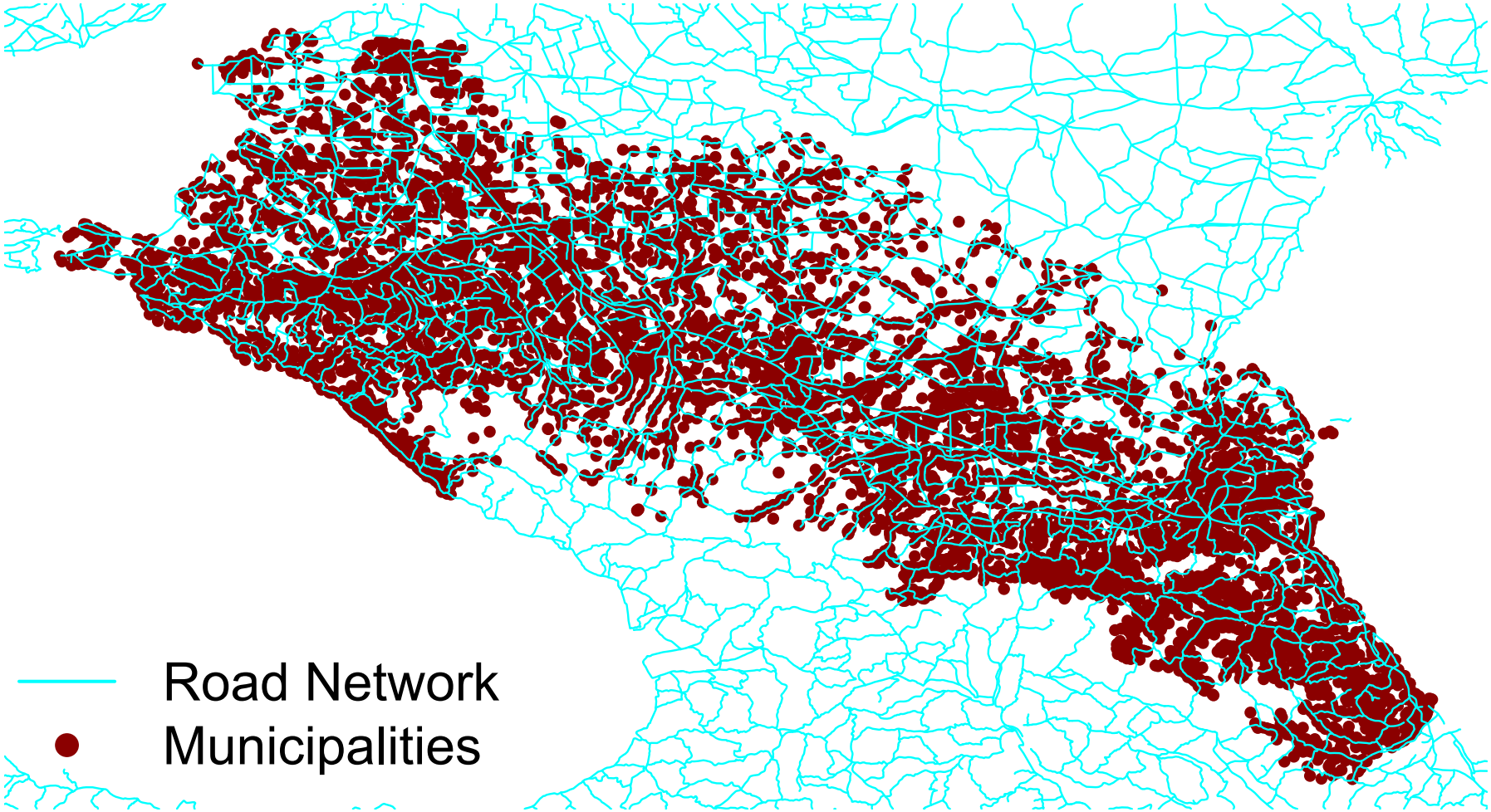


Figure 5: **Road Network**



3 Markov Transition Model with Spatial Spline

Following Amemiya (1985) and Jackman (2000), a logit link function was used to estimate the transition probabilities reported in the paper. The probability that a peaceful village i transitions to violence between times t and $t+1$ is expressed as

$$Pr_{i,t}(PV) = Pr(y_{i,t+1} = 1 | y_{i,t} = 0, \mathbf{x}_{i,t}) = \text{logit}^{-1}(\mathbf{x}_{i,t}\theta_0) \tag{1}$$

and the probability that a violent village remains violent is

$$Pr_{i,t}(VV) = Pr(y_{i,t+1} = 1 | y_{i,t} = 1, \mathbf{x}_{i,t}) = \text{logit}^{-1}(\mathbf{x}_{i,t}\theta_1) \tag{2}$$

where $y_{i,t} = 1$ indicates that location i is experiencing insurgent violence at time t, and $y_{i,t} = 0$ otherwise. θ_0 and θ_1 are sets of regression coefficients that capture the conditional effects of the covariates $\mathbf{x}$ under the two possible current states. These equations are reduced to

$$Pr_{i,t}(V) = Pr(y_{i,t+1} = 1 | \mathbf{x}_{i,t}) = \text{logit}^{-1}(\mathbf{x}_{i,t}\theta_0 + y_{i,t}\mathbf{x}_{i,t}\gamma) \tag{3}$$

where $\theta_1 = \theta_0 + \gamma$. Finally, the expression in (5) is used as the parametric portion of a GAM model

$$Pr_{i,t}(V) = \text{logit}^{-1}\big(\mathbf{x}_{i,t}\theta_0 + y_{i,t}\mathbf{x}_{i,t}\gamma + f(\text{Long}_i, \text{Lat}_i)\big) \tag{4}$$

where $f(\text{Long}_i, \text{Lat}_i)$ is a thin-plate regression spline of the geographic coordinates of village i.

GAMs assume that the mean of the dependent variable ($E[Y_{i,t}] = \mu_{i,t}$) depends on an additive predictor through a link function $g(\mu_{i,t})$, and that the linear predictor can include parametric model components and an unknown nonparametric smooth function $f()$:

$$E[Y_{i,t}] = \mu_{i,t} = g^{-1}(X^{*'}_{i,t}\beta + f(\text{Long}_i, \text{Lat}_i)\big) \tag{5}$$

where $X^*_{i,t}$ is the i,tth row of the model matrix for the strictly parametric model components, and $f(\text{Long}_i, \text{Lat}_i)$ is a thin-plate regression spline of the geographic coordinates of village i.

Thin-plate splines (Duchon, 1977; Wood, 2003) estimate f by minimizing

$$||\mathbf{y} - \mathbf{f}|| + \lambda J_{md}(f) \tag{6}$$

where $\mathbf{y}$ is a vector of y_i's, $\mathbf{f} = |f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)|'$, $\mathbf{x}$ is an $n \times d$ matrix of predictors (in this case, longitude and latitude), $||.||$ is the Euclidean norm, λ is a smoothing parameter, and J_{md} is a "wiggliness penalty" for f, defined as

$$J_{md} = \int \ldots \int_{\mathcal{R}_d} \sum_{\nu_1 + \cdots + \nu_d = m} \frac{m!}{\nu_1! \ldots \nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}} \right)^2 dx_1 \ldots dx_d \tag{7}$$

where m is the order of differentiation, satisfying $2m > d$. In the two predictor case, the wiggliness penalty becomes

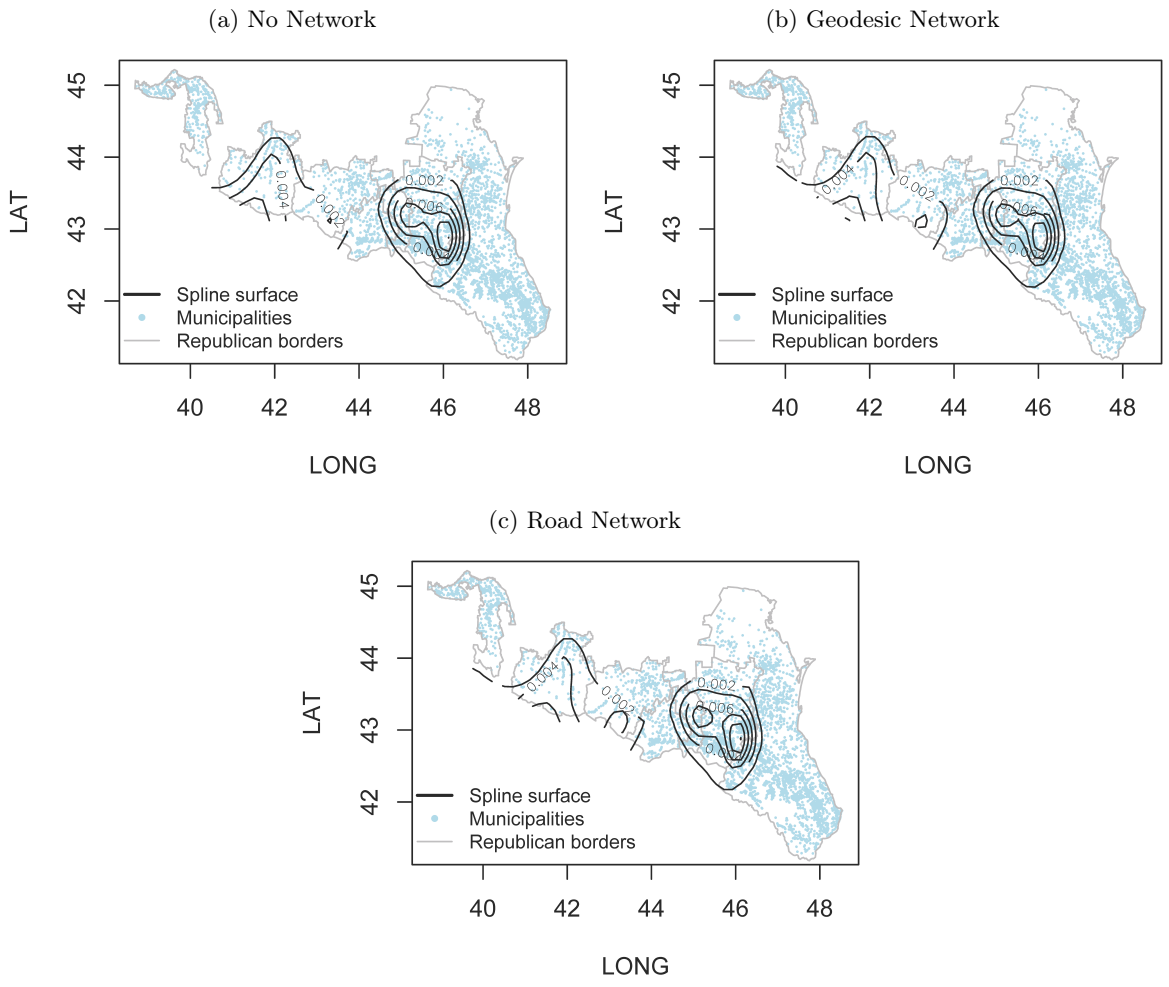
$$J_{22} = \int \int \Big(\frac{\partial^2 f}{\partial \text{Long}^2} \Big)^2 + 2 \Big(\frac{\partial^2 f}{\partial \text{Long}^2 \partial \text{Lat}^2} \Big)^2 + \Big(\frac{\partial^2 f}{\partial \text{Lat}^2} \Big)^2 d\text{Long} d\text{Lat} \tag{8}$$

When $\lambda = 0$, the expression in (2) can be treated as a pure regression spline. When $\lambda \neq 0$, the expression becomes a penalized regression spline. λ also governs the model degrees of freedom, and can be selected with criteria like generalized cross-validation or the Akaike information criterion (AIC).

The advantage of thin-plate regression splines is that they avoid the knot placement problems of conventional regression spline modeling, thus reducing the subjectivity of the model fitting process. They also nest smooths of lower rank within smooths of higher rank. GAM models can be estimated in `R` using the `mgcv` package developed by Simon Wood. See Wood (2006) for a detailed discussion of this class of models.

The surface estimated for Models 1-3 is shown in Figure 6 below. Areas with higher baseline risk of violence are identified around Chechnya and Ingushetia. A similar region is identified further west, to the south of Karachaevo-Cherkessiya and Kabardino-Balkaria. Because much of this last area lies across the border in Abkhazia and Georgia – and thus outside of our study region – this last set of extrapolated predictions does not influence simulations or other results.

Figure 6: **Spline surfaces for Models 1-3**



4 Republic-Level Fixed Effects and Residual Diagnostics

The paper reports the results of several sensitivity analyses that address potential substantive and methodological concerns. One these is a re-fitting of the three models with regional (republic-level) fixed effects. As reported in the paper, the inclusion of regional fixed effects changes neither the substantive results of the models, nor their relative levels of fit and accuracy. A plot of the residuals by republic (Figure 7) further suggests that the distribution of residuals does not vary significantly from zero within any of the seven regions. The horizontal grey lines are individual-level residuals for each village-month observation. The black squares represent 95% confidence intervals of each distribution and the thick black horizontal lines (barely visible) are the medians. The same statistics are reported in tabular form in Table 3. In each case, the 95% confidence interval covers the origin with precision of up to 4 decimal places. While the inclusion of fixed effects shifts the residuals' distribution toward zero in several cases (Ingushetia, North Ossetia, Dagestan, Kabardino-Balkaria), it lends little or no improvement to the others. Another view of the residuals of Model 3, aggregated by village, is shown on Figure 8. A Global Moran's I test of spatial autocorrelation indicates that the observed level of clustering in the residuals is not statistically distinguishable from zero ($I = 0.006, E[I] = -0.003, SD[I] = 0.014, z = 0.658, p\text{-value} = 0.255$).

Figure 7: **Residual plots.** Road network model before (a) and after (b) inclusion of republican fixed effects. Grey horizontal bars represent model residuals grouped by republic.

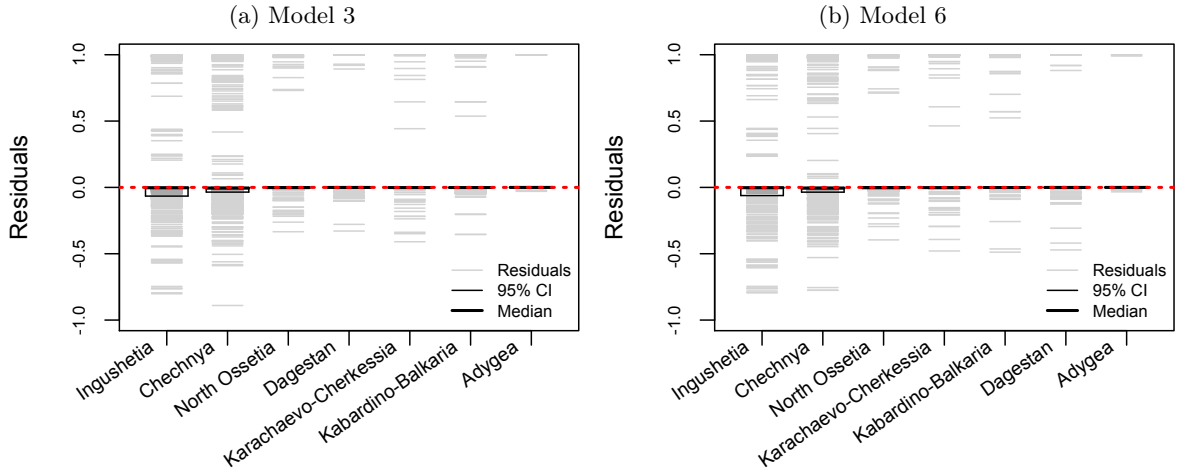
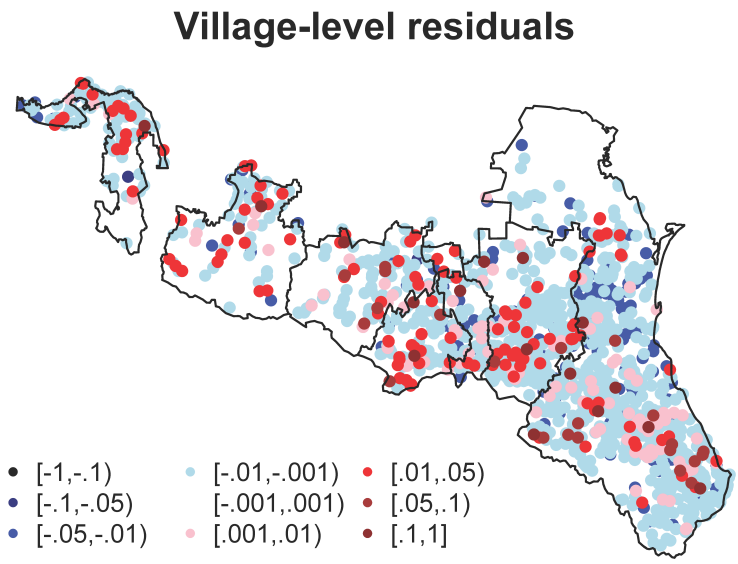


Table 3: **Additional residual diagnostics**.

Republic	Model 3			Model 6		
	Median	95% CI		Median	95% CI	
		Lower	Upper		Lower	Upper
Ingushetia	-0.003	-0.066	-1e-4	-0.002	-0.062	-8e-5
Chechnya	-0.006	-0.035	-2e-4	-0.006	-0.036	-2e-4
North Ossetia	-0.001	-0.007	-2e-5	-0.001	-0.008	-8e-6
Dagestan	-2e-4	-0.005	-8e-6	-2e-4	-0.004	-9e-6
Karachaevo-Cherkessia	-0.001	-0.005	-1e-4	-0.002	-0.005	-7e-6
Kabardino-Balkaria	-0.001	-0.003	-1e-4	-0.001	-0.004	-1e-5
Adygea	-4e-5	-0.002	-4e-7	-3e-5	-0.002	-3e-6

Figure 8: **Residual map**.



References

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Cherkassky, B., A. Goldberg, and T. Radzik (1996, 31 May). Shortest paths algorithms: Theory and experimental evaluation. *Mathematical Programming 73*(2), 129–174.

Davenport, C. and P. Ball (2002). Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995. *Journal of Conflict Resolution 46*, 427–450.

Davenport, C. and A. Stam (2006). Rashomon goes to Rwanda: Alternative Accounts of Political Violence and Their Implications for Policy and Analysis.

Duchon, J. (1977). *Construction Theory of Functions of Several Variables*, Chapter Splines minimizing rotation-invariant semi-norms in Solobev spaces. Berlin: Springer.

Feinerer, I. (2008, October). An introduction to text mining in R. *R News 8*(2).

Feinerer, I., K. Hornik, and D. Meyer (2008, March). Text mining infrastructure in R. *Journal of Statistical Software 25*(5).

GeoNames (2009). NGA GEOnet Names Server. U.S. National Geospatial Intelligence Agency.

Gerner, D. J., P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr (2002). The Creation of CAMEO (Conflict And Mediation Event Observations): An Event Data Framework For A Post Cold War World. Presented at the 2002 Annual Meeting of the American Political Science Association, 29 August – 1 September.

Goskomstat (2009). *Trud i zanyatost' v Rossii [Labor and employment in Russia]*. Moscow: Russian Federal Service of State Statistics.

Grimmer, J. and G. King (2009). Quantitative Discovery from Qualitative Information: A General-Purpose Document Clustering Methodology.

Han, J. and M. Kamber (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hearn, P., T. Hare, P. Schruben, D. Sherrill, C. LaMar, and P. Tsushima (2005). Global GIS: North Eurasia. U.S. Geological Survey & American Geological Institute.

Jackman, S. (2000). In and Out of War and Peace: Transitional Models of International Conflict.

Janko, E. (2009, September). Severo-kavkazskij voennyj okrug [North Caucasus Military District]. warfare.ru.

Kim, W., A. Aronson, and W. Wilbur (2001). Automatic MeSH term assignment and quality assessment. *Proceedings of the AMIA Symposium* (319-23).

King, G. and W. Lowe (2003). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization 57*, 617–642.

Kramer, M. (2004). The Perils of Counterinsurgency: Russia's War in Chechnya. *International Security 29*(3).

Kramer, M. (2005). Guerrilla Warfare, Counterinsurgency and Terrorism in the North Caucasus: The Military Dimension of the Russian-Chechen Conflict. *Europe-Asia Studies 57*(2).

Kuchins, A., M. Malarkey, and S. Markedonov (2011). *The North Caucasus: Russia's Volatile Frontier*. Washington, D.C.: Center for Strategic and International Studies.

LaFree, G. and L. Dugan (2007). Introducing the Global Terrorism Database. *Terrorism and Political Violence 19*, 181–204.

Lyall, J. (2009). Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya. *Journal of Conflict Resolution 53*(2).

Lyall, J. (2010). Are Coethnics More Effective Counterinsurgents? Evidence from the Second Chechen War. *American Political Science Review 104*(1).

Malashenko, A. and D. Trenin (2002). *Vremya Yuga: Rossiya v Chechnye, Chechnya v Rossii [Time of the South: Russia in Chechnya, Chechnya in Russia]*. Moscow: Gendalf.

Memorial (2009). Hronika nasilija [Chronicle of Violence]. Memorial Group, Moscow.

Mikhaylov, Slava, M. L. and K. Benoit (2008). Coder Reliability and Misclassification in Comparative Manifesto Project Codings. Paper presented at the Midwest Political Science Association, Chicago.

O'Loughlin, J. and F. Witmer (2011). The Localized Geographies of Violence in the North Caucasus of Russia, 1999-2007. *Annals, Association of American Geographers 101*(1).

Raleigh, C. and H. Hegre (2009). Population size, concentration, and civil war. A geographically disaggregated analysis. *Political Geography 28*.

Reeves, A. M., S. M. Shellman, and B. M. Stewart (2006). Media Generated Data: The Effects of Source Bias on Event Data Analysis. In *Presented at the International Studies Association annual convention*.

Sagramoso, D. (2007). Violence and conflict in the Russian North Caucasus. *International Affairs 83*(4), 681–705.

Schrodt, P. A. (2001). Automated Coding of International Event Data Using Sparse Parsing Techniques. In *Presented at the International Studies Association*, Chicago.

Schrodt, P. A. and D. J. Gerner (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science 38*, 825–854.

Shellman, S. M. (2008). Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors over Time and Space. *Political Analysis 16*(4).

Souleimanov, E. (2007). *Endless war: the Russian-Chechen conflict in perspective*. Frankfurt am Main: Peter Lang.

Steenberghen, T., K. Aerts, and I. Thomas (2009). Spatial clustering of events on a network. *Journal of Transport Geography*.

Thomas, A. C. and J. K. Blitzstein (2009, 24 October). The Thresholding Problem: Uncertainties Due to Dichotomization of Valued Ties.

Vendina, O. I., V. S. Belozerov, and A. Gustafson (2007). The Wars in Chechnya and Their Effects on Neighboring Regions. *Eurasian Geography and Economics 48*(2).

Weber, R. P. (1990). *Basic Content Analysis* (2nd ed.). Newbury Park, CA: Sage Publications.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B 65*(1), 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall.

Zhan, F. and C. Noon (1998, Feb). Shortest path algorithms: An evaluation using real road networks. *Transportation Science 32*(1), 65–73.