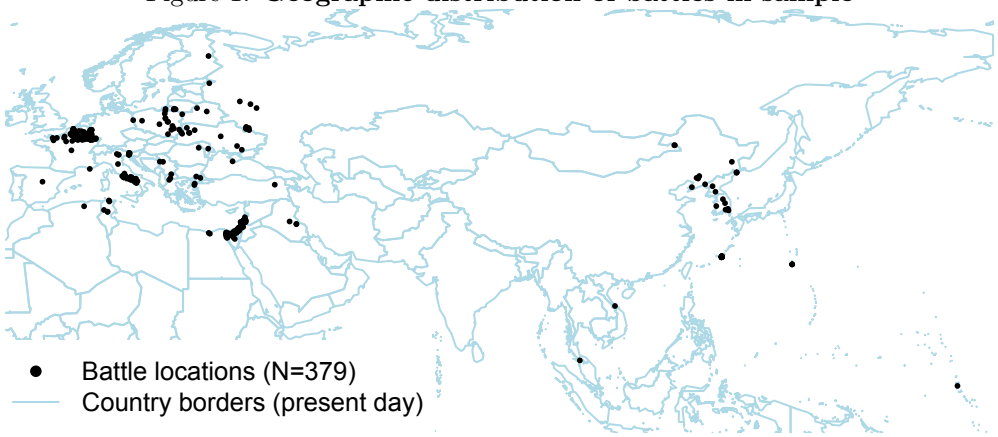
Battle Location Dataset: Codebook*

Yuri M. Zhukov
Department of Government, Harvard University†

March 21, 2010

This is a geocoded version of the U.S. Army's CBD90 (HERO) battle-level dataset (Dupuy, 1984). The sample includes 382 modern battles, original data on which were collected in 1984 by Trevor Dupuy and the Historical Evaluation and Research Organization (HERO) for the U.S. Army's Concepts Analysis Agency. An modified version of the dataset was compiled by the Institute for Defense Analyses in the late 1990's (Hamner, 1997; Biddle, 2004).[1] The locations of areas of operations were ascertained with reference to military atlases, campaign maps, GoogleEarth and secondary historical sources, particularly Dupuy (1984), Dupuy and Dupuy (1986) and Keegan (2003). The spatial distribution of the battles is shown in Figure 1.[2]

Figure 1: **Geographic distribution of battles in sample**



*I am grateful to Jeff Blossom and Sumeeta Srinivasan for their generous assistance and methodological advice in the development of this project. I am also grateful to Stephen Biddle and Stephen Long for sharing their IDA-modified version of CBD90 data.

†Contact: zhukov@fas.harvard.edu

[1]The original dataset included 660 battles in the period 1600-1973, but also carried numerous disadvantages, like double-counts, inconsistent military units of analysis, and selection issues. While little can be done about the latter two concerns, double counts were removed in an updated version of CBD90, compiled in the late 1990s by the Institute for Defense Analyses. The IDA-corrected subset includes only 20th century battles. This is the sample used here. See Hamner (1997) for correction specifics.

[2]A word of caution is in order here. Since selection of battles in CBD90 was neither random nor universal, there is an over-representation of U.S., Israeli and German units. Care should be taken by analysts in drawing inferences based on the full sample included here, although subsets of the data can potentially be quite useful in case studies of particular fronts and theaters of operation.

Case ID variables

iseq : Unique identification number assigned to each battle. Source: Dupuy (1984).

war : Name of war in the course of which the campaign and battle took place. Source: Dupuy (1984).

campn : Name of campaign during which the battle took place. Source: Dupuy (1984).

engagem : Name of battle. Source: Dupuy (1984).

atkr : Name/echelon of attacking unit. Source: Dupuy (1984).

atk_cmdr : Name/rank of commanding officer of attacking unit. Source: Dupuy (1984).

defr : Name/echelon of defending unit. Source: Dupuy (1984).

def_cmdr : Name/rank of commanding officer on defending unit. Source: Dupuy (1984).

date : Start date (continuous) of battle. Source: Dupuy (1984).

Geographic and terrain variables

long, lat : Geographic coordinates of battle. Due to its multi-continental coverage, the data's geographic coordinate system was defined as WGS84 and – to preserve distance – its projected coordinate system was defined as World Equidistant Cylindrical. Geographic coordinates were recorded as Decimal Degrees, with four decimal numbers – accurate to approximately 10 meters at the equator.[3]

dist_att : Euclidean distance (in kilometers) from the battle location to the attacker's capital city.

dist_def : Euclidean distance (in kilometers) from the battle location to the defender's capital city.

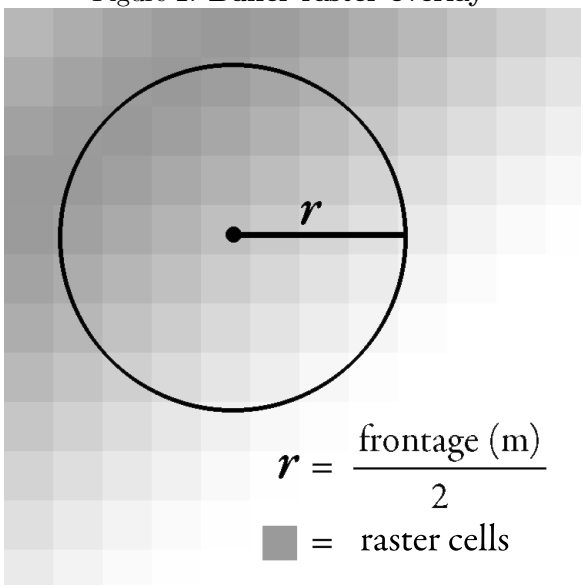
elev_mean_h : Mean elevation of battlefield (high resolution raster). Measured as $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$, where z_i is the elevation within a single raster cell in the battlefield (meters), $\bar{z}$ is the mean elevation (meters) in the battlefield and N is the total number of raster cells within the battlefield. The area of battle is defined as a circular region, the diameter of which is the battle's frontage, in meters (Figure 2). Estimates are based on a 30-second raster grid, where the cell length is approximately 1 kilometer at the equator. Source of elevation data: USGS.

elev_min_h : Lowest elevation point in battle area (high resolution raster). Source of elevation data: USGS.

elev_max_h : Highest elevation point in battle area (high resolution raster). Source of elevation data: USGS.

[3]The main difficulty in geocoding historical battles was occasional mismatch between place names – historical and present-day, source-language and transliterated. Three observations were dropped from the dataset due to insufficient information on their locations.

Figure 2: **Buffer-raster overlay**



elev_var_h : Variance of elevation in battle area (high resolution raster). This is a measure of the ruggedness of terrain, defined as $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2$, where z_i is the elevation within a single raster cell in the battlefield (meters), $\bar{z}$ is the mean elevation (meters) in the battlefield and N is the total number of raster cells within the battlefield. Higher values indicate rugged terrain. Lower values indicate flat terrain. Source of elevation data: USGS.

elev_sd_h : Standard deviation of elevation in battle area (high resolution raster). This is another measure of the ruggedness of terrain, defined as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2}$. Higher values indicate rugged terrain. Lower values indicate flat terrain. Source of elevation data: USGS.

elev_mean_l : Mean elevation of battlefield (low resolution raster). Measured as $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$, where z_i is the elevation within a single raster cell in the battlefield (meters), $\bar{z}$ is the mean elevation (meters) in the battlefield and N is the total number of raster cells within the battlefield. The area of battle is defined as a circular region, the diameter of which is the battle's frontage, in meters (Figure 2). Estimates are based on a 5-minute raster grid, where the cell length is approximately 9.3 kilometer at the equator. Source of elevation data: EEA.

elev_min_l : Lowest elevation point in battle area (low resolution raster). Source of elevation data: EEA.

elev_max_l : Highest elevation point in battle area (low resolution raster). Source of elevation data: EEA.

elev_var_l : Variance of elevation in battle area (low resolution raster). This is a measure of the ruggedness of terrain, defined as $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2$, where z_i is the elevation within a single raster cell in the battlefield (meters), $\bar{z}$ is the mean elevation (meters) in the battlefield and N is the total number of raster cells within the battlefield. Higher values indicate rugged terrain. Lower values indicate flat terrain. Source of elevation data: EEA.

elev_sd_l : Standard deviation of elevation in battle area (low resolution raster). This is another measure of the ruggedness of terrain, defined as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2}$. Higher values indicate rugged terrain. Lower values indicate flat terrain. Source of elevation data: EEA.

a_frontage : Attacker's width of battle front, in kilometers. Source: Dupuy (1984).

d_frontage : Defender's width of battle front, in kilometers. Source: Dupuy (1984).

soft_ar : Soft ground. $\frac{SfAr}{BtAr}$, where $SfAr$ is the portion of the battlefield are (square meters) covered by swamps, sand dunes and inundated areas and $BtAr$ is total area of the battlefield, measured as $\pi(\frac{Fr}{2})^2$, where Fr is the width of the battlefront (meters). Source: DCW.

tree_ar : Tree cover. $\frac{TrAr}{BtAr}$, where $TrAr$ is the portion of the battlefield area (square meters) covered by forest and $BtAr$ is total area of the battlefield. Source: GVM.

river : River crossing. Dummy variable, coded 1 if battle involved a river crossing, and 0 otherwise. Source: Dupuy (1984).

Weather variables

wet : Wet weather. Dummy variable, coded 0 if the battle was fought in dry weather conditions, and 1 in wet (rain, sleet, snow) weather conditions. Source: Dupuy (1984); Army (1986).

overcast : Cloud cover. Dummy variable, coded 0 if the battle was fought in clear sky conditions, and 1 if battle war fought under cloudy skies. Source: Dupuy (1984); Army (1986).

temp : Temperature. Ordinal variable, coded 0 if the battle was fought in cold weather conditions, 1 in temperate weather conditions, and 2 in hot weather conditions. Source: Dupuy (1984); Army (1986).

Force strength and density variables

unit_ar, unit_cp, unit_dv, unit_rg : Unit size. Dummy variables, coded 1 if battle was fought at the level of (a) *Army* and higher, (b) *Corps*, (c) *Division* or (d) *Regiment* and below; 0 otherwise. Source: Dupuy (1984); Army (1986).

amilpert : Total attacker personnel. Source: Dupuy (1984); Army (1986).

dmilpert : Total defender personnel. Source: Dupuy (1984); Army (1986).

ffr : Force-on-force ratio. Att/Def, where Att is total attacker personnel and Def is total defender personnel. Source: Dupuy (1984); Army (1986).

afsr: Attacker's force-to-space ratio. Att/Fr, where Att is total attacker personnel and Fr is the width of the battlefront (in kilometers). Source: Dupuy (1984); Army (1986).

dfsr: Defender's force-to-space ratio. Def/Fr, where Def is total defender personnel and Fr is the width of the battlefront (in kilometers). Source: Dupuy (1984); Army (1986).

Strategy variables

frontat : Direct strategy. Indicator variable, coded as 1 if the attacker executed a frontal attack, and 0 otherwise. Source: (Dupuy, 1984; Army, 1986).

envelop : Indirect strategy. Indicator variable, coded as 1 if the attacker executed an envelopment, double envelopment or flanking maneuver, and 0 otherwise. Source: Dupuy (1984); Army (1986).

defoff : Defender counterattack. Indicator variable, coded as 1 if the defender executed a counterattack, and 0 otherwise. Source: Dupuy (1984); Army (1986).

surprise : Level of surprise. Ordinal variable, coded 3 if "complete" surprise was achieved by the attacker, 2 if "substantial" surprise was achieved by the attacker, 1 if "minor" surprise was achieved by the attacker, 0 if "no surprise" was achieved by either side, -1 if "minor" surprise was achieved by the defender, -2 if "substantial" surprise was achieved by the defender, and -3 if "complete" surprise was achieved by the defender. Source: Dupuy (1984); Army (1986).

fort : Fortified defenses. Indicator variable, coded 1 if the defender employed planned entrenchments, field fortifications, or other obstacles, and 0 otherwise. Source: Dupuy (1984); Army (1986).

Technology variables

a_tks_to : Total armored vehicles (attacker). Source: Dupuy (1984); Army (1986).

a_tks_lt : Attacker's total number of motorized light armor vehicles. Source: Dupuy (1984); Army (1986).

a_tks_mb : Attacker's total number of main battle tanks. Source: Dupuy (1984); Army (1986).

a_sortie : Attacker's number of air sorties flown. Source: Dupuy (1984); Army (1986).

a_arty : Attacker's total number of artillery pieces. Source: Dupuy (1984); Army (1986).

d_tks_to : Total armored vehicles (defender). Source: Dupuy (1984); Army (1986).

d_tks_lt : Defender's total number of motorized light armor vehicles. Source: Dupuy (1984); Army (1986).

d_tks_mb : Defender's total number of main battle tanks. Source: Dupuy (1984); Army (1986).

d_sortie : Defender's number of air sorties flown. Source: Dupuy (1984); Army (1986).

d_arty : Defender's total number of artillery pieces. Source: Dupuy (1984); Army (1986).

Social and political variables

a_democ : Attacker's level of democracy. County's Polity IV DEMOC score, scaled 0-10, where higher values indicate higher levels of institutionalized democracy. Source: Marshall and Jaggers (2007).

d_democ : Defender's level of democracy. County's Polity IV DEMOC score, scaled 0-10, where higher values indicate higher levels of institutionalized democracy. Source: Marshall and Jaggers (2007).

a_prisec : Attacker's educational attainment. Level of primary and secondary education per capita. Source: Banks (1976).

d_prisec : Defender's educational attainment. Level of primary and secondary education per capita. Source: Banks (1976).

a_coup5 : Attacker's civil-military comity. Dummy variable, coded 1 if country experienced a military coup in the five years preceding the battle, and 0 otherwise. Source: Banks (1976).

d_coup5 : Defender's civil-military comity. Dummy variable, coded 1 if country experienced a military coup in the five years preceding the battle, and 0 otherwise. Source: Banks (1976).

Battle outcome variables

duration : Duration of battle, in days. Source: Dupuy (1984); Army (1986).

acas_per : Attacker's total unrecoverable battle casualties (personnel). Source: Dupuy (1984); Army (1986).

dcas_per : Attacker's total unrecoverable battle casualties (personnel). Source: Dupuy (1984); Army (1986).

ler : Loss-exchange ratio. $\frac{AttC}{DefC}$, where $AttC$ is the attacker's total unrecoverable battle casualties (personnel) and $DefC$ is the defender's total unrecoverable battle casualties (personnel). Higher values indicate defender advantage. Lower values indicate attacker advantage. Source: Dupuy (1984); Army (1986).

ln_ler : Loss-exchange ratio (natural log). Source: Dupuy (1984); Army (1986).

a_advance : Attacker's total penetration distance (meters). Higher values indicate attacker advantage. Lower values indicate defender advantage. Source: Dupuy (1984); Army (1986).

rtadv : Attacker's rate of advance. $\frac{Adv}{Dur}$, where Adv is the attacker's total penetration distance in kilometers and Dur is the duration of the battle in days. Higher values indicate attacker advantage. Lower values indicate defender advantage. Source: Dupuy (1984); Army (1986).

References

U.S. Army. *Combat History Analysis Study Effort (CHASE): Progress Report for the Period August 1984 - June 1985.* U.S. Army Concepts Analysis Agency, Bethesda, MD, 1986.

Arthur S. Banks. *Cross-National Time Series, 1815-1973.* Inter-university Consortium for Political and Social Research (ICPSR), Ann Arbor, MI, 1976.

Stephen D. Biddle. *Military Power: Explaining Victory and Defeat in Modern Battle.* Princeton University Press, Princeton, NJ, 2004.

R. Ernest Dupuy and Trevor N. Dupuy. *The Encyclopedia of Military History: From 3500 B.C. to the Present*. Harper and Row, New York, 1986.

Trevor Dupuy. *Analysis of Factors that have Influenced Outcomes of Battles and Wars*. Historical Evaluation and Research Organization, Dunn Loring, VA, 1984.

Christopher Hamner. Historical review of U.S. Army CBD90 dataset. Memorandum for the Record (unpublished material), November 1997.

GVM. *Global Landcover Classification for the Year 2000 (GLC2000)*. Joint Research Centre of the European Commission, Space Applications Institute, Global Vegetation Monitoring Unit, 2003.

John Keegan, editor. *Atlas of the Second World War*. Harper Collins, Slovenia, 2003.

Monty G. Marshall and Keith Jaggers. *Polity IV Project: Dataset Users' Manual*. Center for Global Policy, School of Public Policy, George Mason University and Center for Systemic Peace, Arlington, VA, October 2007.

EEA. *5-Minute Gridded Global Relief Data Collection (ETOPO5)*. National Geophysical Data Center, World Data Center for Geophysics and Marine Geology, Boulder, CO, 1988.

DCW. *Digital Chart of the World Landcover Polygons*. National Geospatial-Intelligence Agency, 1992.

USGS. *Global 30 Arc Second Elevation Data (GTOP030)*. U.S. Geological Survey, National Mapping Division, EROS Data Center, 1996.