Border Crossings of the World Codebook
1993-2020

Version 1.1[1]

Citation:
Michael R. Kenwick, Beth A. Simmons, and Richard McAlexander. 2024. "Infrastructure and Authority at the State's Edge: The Border Crossings of the World Dataset." *Journal of Peace Research* 61(3): 500-510.

Table of Contents

1 – Overview of The BCW Dataset 1
2 – The Condensed BCW Dataset Codebook 3
3 – The Full BCW Dataset Codebook 5
4 - The Coder-Level (Disaggregated) BCW Dataset Codebook 9
5 – Description of Interpolation and Aggregation Strategies 12

[1] This work was supported by the National Science Foundation Grant #1917573. Version 1.1 contains a series of relatively modest modifications and corrections to the initial release of the data. See the README file associated with this data release for additional information.

1 – Overview of The BCW Dataset

The Border Crossings of the World (BCW) dataset records information on international border crossings (1993-2020), defined as instances where a highway crosses an international border. This information includes features of the road itself (single-lane, multi-lane, etc.), as well as the presence or absence of filtering infrastructure used for search and inspection (e.g., official buildings, gates, inspections lanes). The data were collected based on visual inspections of publicly available satellite imagery, performed by a team of research assistants at Rutgers University and the University of Pennsylvania. See Kenwick, Simmons, and McAlexander (2024) for more information.

Files and Data Formats:

The BCW are available in three basic formats: The Condensed BCW Data, the Full BCW Data, and the Coder-Level BCW data, each described below. We expect most users will be primarily interested in the condensed version of the data. Note that the condensed and full versions of the data record information about border crossings based on the semi-crowdsourced classifications of individual research assistants from the coder-level data. These data sets also impute values for years when no satellite imagery data is available (See Section 5 for more information).

- **The Condensed BCW Data** (bcw_v1p1.csv) – This is the most condensed form of the data set for researchers interested in simply measuring the buildup of filtering infrastructure at international border crossings. The data is stored in a *directed border crossing dyad* format, such that every border crossing is recorded twice in each year, one for each country's side.[2] The data include composite measures of filtering infrastructure, after interpolating and aggregating coder-level data. *See Section 2 for more information.*
- **The Full BCW Data** (bcw_full_v1p1.csv) – This is a more comprehensive version of the BCW data set, which includes the full set of variables recorded for each border crossing year. This version is stored in an *undirected border crossing dyad* format, such that each row records information for both sides of a border crossing in a particular year. Composite variables filtering infrastructure are included along with more refined indicators for road size and type, and the various forms of filtering infrastructure (official buildings, gates, inspections lanes). Each variable is also made available using different interpolation strategies for imputing values when satellite imagery is unavailable. *See Section 3 for more information.*
- **Coder-Level BCW Data** (*bcw_coder_level_v1p1.csv*) – This is the most disaggregated version of the data, which contains the raw classifications provided by each research assistant (i.e., "coders" or "annotators"), such that each row is one assistant's classifications for a particular border crossing year.[3] This version of the data will be useful to researchers

[2] For example, a crossing between Canada and the United States would have two entries for each year, one recording information on the Canadian side and one for the US side.

[3]The coder-level data set also includes the years 1990-1992, which are retained for interpolation purposes, but removed from the primary data set since imagery during this time is often sparse and of low quality.

interested in generating their own aggregation or imputation procedures for combining the coder-specific information. *See Section 4 for more information.*

Auxiliary Files:

- **Interpolation and Aggregation Script** (bcw_transform.R) – This is a script which aggregates and interpolates the coder-level data (bcw_coder_level.csv) to generate the aggregated versions on the data.

2 – The Condensed BCW Dataset Codebook

This version of the data is in a *directed border crossing dyad* format such that each row pertains to one country's side of a border crossing in a particular year. The filtering infrastructure variables record the extent to which that country has built up gates, official buildings, and split-lane inspections areas on its side of the crossing. Users interested in specific types of infrastructure or features of the roads themselves should consult with the full BCW data (Section 3).

Variable description (column names)

Border crossing id (borderid)

A unique ID indicator for each border crossing.

Country names (country1, country2)

The names of the two countries that share a particular border crossing. Because the data are in directed dyad format, country 1 is the actor whose infrastructure is being measured, and country 2 is the "target" of that infrastructure.

COW country codes (cow1, cow2)

The Correlates of War codes for each country.

Geo-spatial coordinates (lat, long)

The geographic coordinates of a border crossing recorded in decimal degrees.

Year (year)

The calendar year for each observation.

Filtering infrastructure index (structure_tot)

An additive index of the filtering infrastructure on Country 1's side of a border crossing, including the presence of gates, the existence of a split-lane inspections area, and the number of official buildings (consult with the full version of the BCW data for how each is measured). Indicators for each of these three structures are rescaled to have a range between 0 and 1, producing an index with a range from 0 to 3. **A value of 0 indicates that there is a consensus among all coders that no filtering infrastructure is present, and a value of 3 indicates the opposite, with all types of infrastructure present at their maximal values (i.e., covered gates, split-lane areas, and several official buildings).**

This indicator is forward interpolated to impute values even for years when satellite imagery of a border crossing is absent. It is missing prior to the first year in which a border crossing became observable with satellite imagery (e.g., if a crossing is first observed in 2006, then the index is missing for all prior years).

Note: this variable was computed using the gate_locf, split_locf, and numbld_locf, variables discussed in Section 2. It is identical to structure_tot_locf in the full version of the data set.

Filtering infrastructure index, backward interpolated (structure_tot_back)

This is a backward interpolated version of the filtering infrastructure index above, with values imputed for years even prior to the first year of satellite imagery.

Note: this variable was computed using the gate_locf_back, split_locf_back, and numbld_locf_back, variables discussed in Section 2. It is identical to structure_tot_locf_back in the full version of the data set.

3 – The Full BCW Dataset Codebook

This is the complete version of the BCW data set, stored in an *undirected border crossing dyad* format, with one row per border crossing year, and variables recording infrastructure on each country's side (i.e., country1 and country2, and variables suffixed _1 and _2). The data are semi-crowdsourced, such that values assigned to infrastructure variables are weighted averages of the those provided by individual coders. Values are also interpolated for years in which satellite imagery is missing for a particular border crossing. Information on each process is briefly described below, with greater discussion in Section 5. Infrastructural variables are made available in multiple formats depending on which type of interpolation is used.

Weighting scheme: Most variables in this data set have values which are a weighted average of each research assistant's (coder's) classification. Weights are assigned based on whether a coder was a highly trained research assistant (RA), or a participant in a multi-day crowdsourcing ("codeathon") event. When all coders are in the same group, we take a simple average. When a border crossing is coded by both types of workers, we assign a weighted average, giving scores from the highly trained RAs twice the weight of the crowdsourced workers.[4]

Interpolation strategies: Satellite imagery data of border crossings is sparse, meaning that it is not available for every year for each border crossing, and so we make use of multiple interpolation strategies to impute values for several variables below. The first is ***last observation carried forward*** **interpolation** ***(variables suffixed _locf)***, which means that each variable retains is value during years in which satellite imagery is missing, until the next year in which satellite imagery is available. The second is ***linear imputation (variables suffixed _lin),*** where missing values are linearly imputed between years where satellite imagery is available. For each, variables are assigned missing values prior to the first year of satellite imagery. Variables that are suffixed ***_back,*** however, also perform **backward interpolation**, where structures that were present or absent in the first year of imagery were assumed to have remained so in prior years. These variables correlate very highly, although there may be some research contexts in which one strategy would be preferred over another.

Variable description (column names)

Border crossing id (borderid)

A unique ID indicator for each border crossing.

Country names (country1, country2)

The names of the two countries that share a particular border crossing. The "_1" and "_2" suffix map onto the variables for country 1 and country 2, respectively.

COW country codes (cow1, cow2)

The COW codes for each country.

[4] Note that we do not incorporate crowdsourced classifications for the split-lane indicator.

Geo-spatial coordinates (lat, long)

The geographic coordinates of a border crossing recorded in decimal degrees.

Year (year)

The calendar year for each observation.

Proportion of coders identifying imagery (imagery_prop_1, imagery_prop_2)

Records the proportion of coders that identified imagery as being present in a particular year.

Road status (road_status_1_locf, road_status_2_locf, road_status_1_locf_back, road_status_2_locf_back, road_status_1_lin, road_status_2_lin, road_status_1_lin_back, road_status_2_lin_back)

Records the average, interpolated values of the road status indicator, with values ranging from a minimum of 0 (no road is present) to a maximum of 3 (all coders indicate that a major road is present).

- 0 - **No Road (No Crossing):** no proximate road-border intersection. Note that this value is seldom observed in the final data set after removing false positive observations.
- 1 - **Unpaved road:** a road intersecting the border, but it appears to be unpaved
- 2 - **Paved road:** a clearly paved road intersecting the border, typically at least one lane each direction
- 3 - **Major road:** a clearly paved, major road intersecting the border

Split lane inspection area (split_1_locf, split_2_locf, split_1_locf_back, split_2_locf_back, split_1_lin, split_2_lin, split_1_lin_back, split_2_lin_back)

Records the average, interpolated values of the indicator for split-lane inspection facilities, with values ranging from 0 (all coders indicate that a split-lane inspection facility is absent) to 1 (all coders indicate that a split-lane inspection facility is present). Split-lanes are lane configurations that appear to accommodate search, interdiction, inspection, or pullover areas.

- 0 - **No split-lane inspection area.** The road does not widen or split in any way at the border, even if it has two lanes each direction.
- 1 – **Split-lane Inspection area.** Lanes split or proliferate for inspections at or very near the border. Auxiliary lanes may loop to the side and then rejoin the road, or there may be a turnout area for traffic at or near official-looking buildings, kiosks, gates, etc.

Gate or barricade on the road? (gate_1_locf, gate_2_locf, gate_1_locf_back, gate_2_locf_back, gate_1_lin, gate_2_lin, gate_1_lin_back, gate_2_lin_back)

Records the average, interpolated values of the gate indicator, with values ranging from 0 (all coders indicate that a gate is absent on a given side of a border crossing) to 2 (all coders indicate that a covered gate is present on a given side of a border crossing). Gates indicate that traffic on a road can be at least nominally stopped by a barrier placed in the road. Covered gates are meant to denote sheltered structures that straddle the road, where personnel not only could stop traffic, but under which it is plausible to check documents, question travelers, etc.

- 0 - **No gate**
- 1 - **Uncovered gate**
- 2 - **Covered gate**

Official building(s) (ofbld_1_locf, ofbld_2_locf, ofbld_1_locf_back, ofbld_2_locf_back, ofbld_1_lin, ofbld_2_lin, ofbld_1_lin_back, ofbld_2_lin_back)

Records the average, interpolated values of the dichotomous indicator for whether official buildings were present, with values ranging from 0 (all coders agree official buildings were absent) to 1 (all coders agree official buildings were present).

- 0 - **No official buildings**
- 1 - **One or more official buildings**

Note: Unlike privately owned buildings, official buildings tend to be

- at or near the border (proximity; nearer the border than residential or commercial structures).
- symmetrical on each side of the road.
- located on road loops that swing out from and then rejoin the main road.
- near to inspection areas; near to gates/barriers.
- one of a kind or one of a cluster of a kind around an inspection center/vehicle holding or parking area.
- linked/near to gates or barriers.
- marked with some form of state authority (e.g., flags).
- tagged as such in online mapping services.

Number of buildings (numbld_1_locf, numbld_2_locf, numbld_1_locf_back, numbld_2_locf_back, numbld_1_lin, numbld_2_lin, numbld_1_lin_back, numbld_2_lin_back)

Records the average, interpolated values of the ordinal indicator for how many official buildings were present, with values ranging from 0 (all coders agree official buildings were absent) to 4 (all coders agree six or more official buildings were present).

- 0 - **No official buildings**
- 1 - **One building** on a country's side of the border appears to qualify as official
- 2 - **Two to three** buildings on a country's side of the border appear to qualify as official
- 3 - **Four to five** buildings on a country's side of the border appear to qualify as official
- 4 - **Six or more** buildings on a country's side of the border appear to qualify as official

Multiple lanes (multilane_1_locf, multilane_2_locf, multilane_1_locf_back, multilane_2_locf_back, multilane_1_lin, multilane_2_lin, multilane_1_lin_back, multilane_2_lin_back)

Records the average, interpolated values of the multilane indicators, with values ranging from 0 (all coders indicate that a road with a single lane is present) to 3 (all coders indicate that four or more lanes are present).

- 0 - the road on a given state's side of the border has **one lane, total**
- 1 - the road on a given state's side of the border has **two lanes, total**
- 2 - the road on a given state's side of the border has **three or four lanes, total**

- 3 - the road on a given state's side of the border has **more than four lanes**

Filtering Infrastructure index (structure_tot_1_lin, structure_tot_2_lin, structure_tot_1_lin_back, structure_tot_2_lin_back, structure_tot_1_locf, structure_tot_2_locf, structure_tot_1_locf_back, structure_tot_2_locf_back)

An additive index of the filtering infrastructure on each side of a border crossing, including the presence of gates, the existence of a split-lane inspections area, and the number of official buildings. Indicators for each of these three structures are rescaled to have a range between 0 and 1, producing an index with a range from 0 to 3. **A value of 0 indicates that there is a consensus among all coders that no filtering infrastructure is present, and a value of 3 indicates the opposite, with all types of infrastructure present at their maximal values (i.e., covered gates, split-lane areas, and several official buildings).** Separate versions of the index are computed using the various interpolation strategies of the continuative indicators.

Data source (data_source)

Records the original data source that was used to identify a particular border crossing.

- **agi** – The border crossing was originally identified using roads data from the American Geosciences Institute
- **osm** – The border crossing was identified after expanding the data using Open Street Map

Number of coders (num_coders)

Records the number of coders that provided classifications for a particular border crossing.[5]

Observed Border Crossing Construction (construction_observed)

Records whether there was evidence that the border crossing was constructed during the period of observation. For this to be the case, the crossing must have been visibly absent in satellite imagery in earlier years, but present in later years.

- **0** – No evidence of border crossing construction
- **1** – Visible evidence of border crossing construction

Note: Years prior to road construction are omitted from the data because they do not meet the inclusion criteria of the BCW.[6] Note also that this variable only captures instances where road construction was observable and necessarily omits cases where road construction may have occurred prior to the first year in which satellite imagery was available.

[5] Note that this number omits coders who may have (incorrectly) identified a false positive at the border crossing location.

[6] For example, if satellite imagery is available in 2005 and 2007 and a crossing is absent in the former but present in the latter, then only years from 2007 onward are included in the final data set.

4 - The Coder-Level (Disaggregated) BCW Dataset Codebook

Each row is a single coder's classification of a particular border crossing year, prior to interpolation.

Variable description (column names)

Border crossing id (borderid)

A unique ID indicator for each border crossing.

Coder id (coder)

An anonymized ID for each human coder.

Country names (country1, country2)

The names of the two countries that share a particular border crossing. The "_1" and "_2" suffix map onto the variables for country 1 and country 2, respectively.

COW country codes (cow1, cow2)

The COW codes for each country.

Geo-spatial coordinates (lat, long)

The geographic coordinates of a border crossing recorded in decimal degrees.

Year (year)

The calendar year for each observation.

Imagery quality (imagery_earth_1, imagery_earth_2)

Records whether images are available and if so, their quality.

- 3 - **Good**: imagery is available and of high quality
- 2 - **Fair**: imagery is available and of intermediate quality
- 1 - **Poor**: the area is not blacked out, but the image is extremely blurry
- 0 - **Unavailable**: the area is blacked out

Road status (road_status_1, road_status_2)

Records the existence and features of the highway that constitutes the border crossing.

- 0 - **No road (no crossing):** no proximate road-border intersection. Note that this value is seldom observed in the final data set after removing false positive observations.
- 1 - **Unpaved road:** a road intersecting the border, but it appears to be unpaved
- 2 - **Paved road:** a clearly paved road intersecting the border, typically at least one lane each direction
- 3 - **Major road:** a clearly paved, major road intersecting the border

Gate or barricade on the road? (gate_1, gate_2)

Records whether it is likely that traffic on a road can be at least nominally stopped by a barrier placed in the road. Covered gates are meant to denote sheltered structures that straddle the road, where personnel not only could stop traffic, but under which it is plausible to check documents, question travelers, etc.

- 0 - **No gate**
- 1 - **Uncovered gate**
- 2 - **Covered gate**

Multiple lanes (multilane_1, multilane_2)

Records the number of lanes on a road, on each side of the border.

- 0 - the road on a given state's side of the border has **one lane, total**
- 1 - the road on a given state's side of the border has **two lanes, total**
- 2 - the road on a given state's side of the border has **three or four lanes, total**
- 3 - the road on a given state's side of the border has **more than four lanes**

Split lane inspection area (split_1, split_2)

Records lane configurations that appear to accommodate search, interdiction, inspection, or pullover areas. This is not the number of lanes per se; it is the configuration and purpose of those lanes.

- 0 - **No split-lane inspection area.** The road does not widen or split in any way at the border, even if it has two lanes each direction
- 1 – **Split-lane Inspection area.** Lanes split or proliferate for inspections at or very near the border. Auxiliary lanes may loop to the side and then rejoin the road, or there may be a turnout area for traffic at or near official-looking buildings, kiosks, gates, etc.

Official building(s) (ofbld_1, ofbld_2)

Records whether there are official looking buildings at or near the border.

- 0 - **No official buildings**
- 1 - **One or more official buildings**

Note: Unlike privately owned buildings, official buildings tend to be

- at or near the border (proximity; nearer the border than residential or commercial structures).
- symmetrical on each side of the road.
- located on road loops that swing out from and then rejoin the main road.
- near to inspection areas; near to gates/barriers.
- one of a kind or one of a cluster of a kind around an inspection center/vehicle holding or parking area.
- linked/near to gates or barriers.
- marked with some form of state authority (e.g., flags).
- tagged as such in online mapping services.

Number of buildings (numbld_1, numbld_2)

Records how many official-looking buildings are present.

- 0 - **No official buildings**
- 1 - **One building** on a country's side of the border appears to qualify as official
- 2 - **Two to three** buildings on a country's side of the border appear to qualify as official
- 3 - **Four to five** buildings on a country's side of the border appear to qualify as official
- 4 - **Six or more** buildings on a country's side of the border appear to qualify as official

Data source (data_source)

Records the original data source that was used to identify a particular border crossing.

- **agi** – The border crossing was originally identified using roads data from the American Geosciences Institute
- **osm** – The border crossing was identified after expanding the data using Open Street Map

Coder type (coder_type)

Records whether the coder providing a classification was a highly-trained research assistant or a participant in the multi-day "codeathon" crowdsourcing effort. Both types of coders completed significant training programs and were closely monitored, but research assistants developed greater expertise applying the coding rules.

- **RA** - The coder was a research assistant employed during the data collection effort
- **codeathon** - The coder was a participant in the crowdsourcing "codeathon" effort

Observed Border Crossing Construction (construction_observed)

Records whether there was evidence that the border crossing was constructed during the period of observation. For this to be the case, the crossing must have been visibly absent in satellite imagery in earlier years, but present in later years.

- **0** – No evidence of border crossing construction
- **1** – Visible evidence of border crossing construction

Note: This variable was coded once per border crossing after inspecting coder-level data (i.e., its values were not assigned by individual coders). Years prior to road construction are omitted from the data because they do not meet the inclusion criteria of the BCW.[7] Note also that this variable only captures instances where road construction was observable and necessarily omits cases where road construction may have occurred prior to the first year in which satellite imagery was available.

[7] For example, if satellite imagery is available in 2005 and 2007 and a crossing is absent in the former but present in the latter, then only years from 2007 onward are included in the final data set.

5 – Description of Interpolation and Aggregation Strategies

This section outlines interpolation and aggregation procedures in greater detail. By interpolation, we refer to our strategy for filling in missing values within the BCW that occur in years lacking satellite imagery. By aggregation, we refer to our strategy for moving from the individual classifications provided by each coder to the final, aggregated version of the data set with one row for every border crossing year. For all variables, interpolation is performed prior to aggregation.[8]

Our two interpolation strategies are visualized in Figure 1. The first strategy is stepwise or "last observation carried forward" interpolation ("_locf" in variable designations). Here, a structure that was present (absent) is assumed to remain present (absent) until the next year for which satellite imagery becomes available. The second strategy is linear interpolation ("_lin" in variable designations), where scores for each variable are imputed linearly between observation periods. There can be benefits and drawbacks to each method. The values produced by the "last observation carried forward" strategy may lag behind change occurring in the real world (e.g., a gate is constructed in 2010, but it is not visible until 2011). Linear interpolation avoids this problem but may create the illusion of incremental changes even when structures are built rather quickly (i.e., it typically does not take several years to build a gate). Which strategy is preferred will differ depending on the research context, but the scores produced using each correlate very highly and so the choice between them is often trivial.

We also make the data available with backward-interpolated variables, represented by the dashed blue line in Figure 1. Here the values of variables observed in the first year of satellite imagery are carried backward into previous years when no imagery is yet available. Given that structures are "sticky" we think interpolated left-censored values is plausible in most cases and conservative with respect to infrastructural increases. For example, if a crossing was first observed in 2005 and an official building was present (absent) we might assume it was also present (absent) in 2004.[9] Still, this assumption becomes increasingly heroic as one imputes further backward in time.

After interpolation, the scores provided by each coder are aggregated by taking a weighted average, with weights assigned based on whether an individual coder was a highly trained research assistant

[8] In principle, one could analyze the BCW without any form of imputation or interpolation, but doing so could be problematic since satellite imagery is non-randomly missing for relatively large subset of observations. Further, we perform interpolation prior to aggregation because failing to do so might would produce greater error in the data. By way of example, consider a border crossing coded by two individuals: Coder 1 and Coder 2. Both coders identify the same building near a crossing, but Coder 1 labels it "official" (ofbld=1) while Coder 2 thinks it is private and therefore does not code the structure (ofbld=0). Additionally, Coder 1 only identifies imagery in 2005, while Coder 2 finds imagery in 2005 and 2006. If we aggregated (averaged) the data prior to interpolation, this would create the appearance that an official building had been demolished (or decommissioned), since the average score of the official building indicator would be 0.5 in 2005, but 0 in 2006. Conversely, if interpolation occurs first, then the score would be 0.5 in each year, more accurately reflecting the disagreement between the coders.

[9] Backward interpolation also somewhat reduces within-case variability caused by coders who begin coding crossings in different years and disagree about the presence of a given structure. For example, if one coder first finds imagery in 2006 and identifies a split-lane, while another first finds imagery in 2007 but does not, then the value on the split lane indicator will reduce when only using forward-only interpolation, but will remain constant when backward interpolation is also employed.

(RA) or a participant in the crowdsourcing ("codeathon") exercise. Highly trained research assistants each took part in a multi-week training exercise and typically worked on the project for several months or years, giving them greater expertise. Participants in the crowdsourcing exercise were also provided with a multi-week training regimen and subject to quality checks, but worked for a shorter duration and were expected to have less expertise in the coding rules. We therefore assign twice the weight to the RA classification than we did to the crowdsourced classifications.[10]

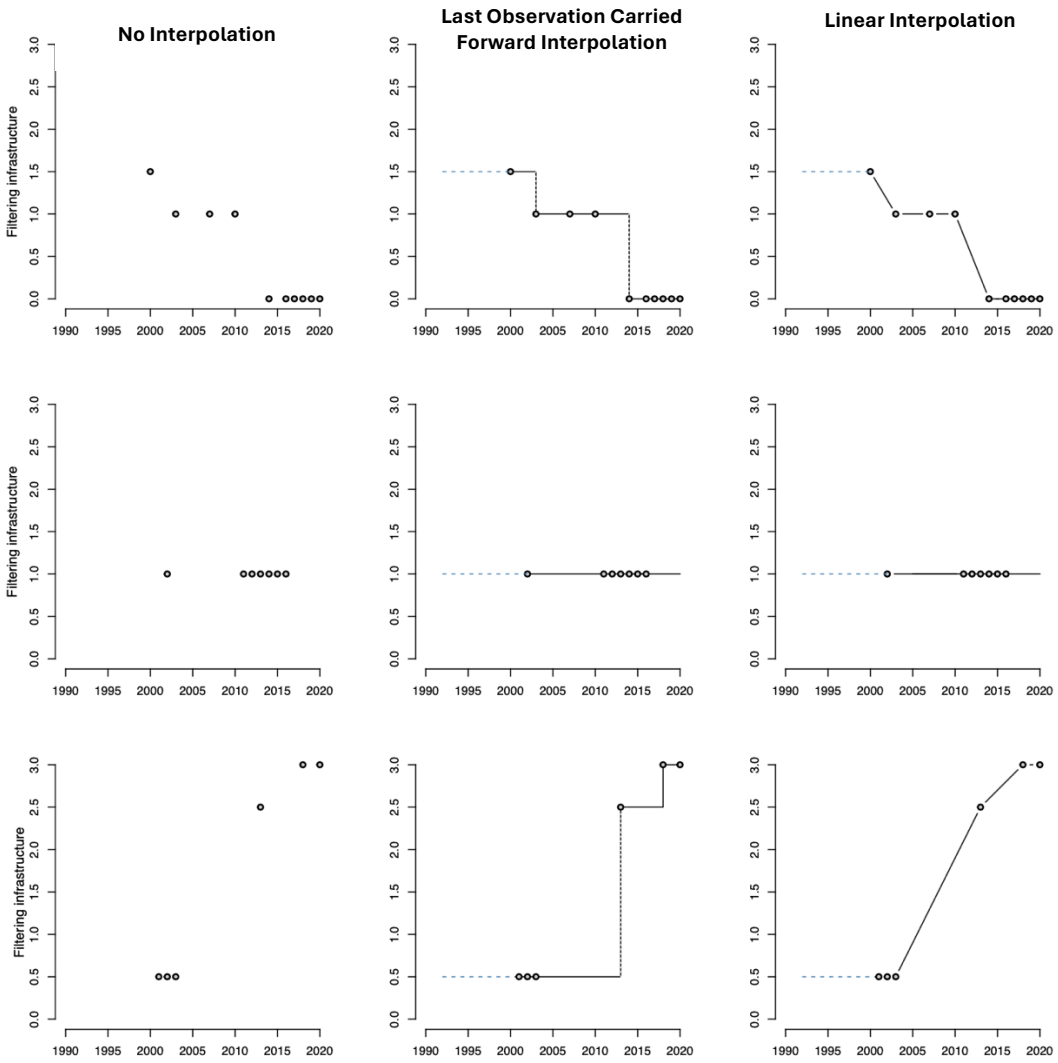

Figure 1: Data interpolation strategies

Note: Plot displays an additive index of filtering structures (gates, split-lanes, official buildings) for three sample border crossings using different strategies for interpolation. Left-censored periods are marked with a blue, dashed line.

[10] As discussed in the article of record, we omit information on the split-lane indicator from the crowdsourcing event, since this variable was recorded with a higher degree of error than others.