

Technical Note

Longitudinal Study of Refugees from Ukraine

REG2204

Ukraine, September 2026

1. Executive Summary

This technical note accompanies the public release of an anonymised longitudinal panel dataset and documents the methodology used to prepare the dataset for publication. It briefly describes the transformations applied to the source data to reduce re-identification risks and explains how these transformations changed the structure and level of detail of the published dataset compared with the original data.

Tracking individuals over multiple interview rounds provides valuable insights into long-term displacement and socio-economic trends. However, the longitudinal structure of the data also creates specific data protection risks. Removing direct identifiers, such as names or phone numbers, is not sufficient to make the dataset anonymous, as combinations of other respondent characteristics may still allow records to be linked to information available from external sources.

Since respondents' characteristics may change between rounds, each respondent has a sequence of values across time — a trajectory — which can identify them more narrowly than an individual characteristic or a combination of characteristics observed in a single round. If a trajectory is rare, an outside observer who already knows a few facts about a respondent may be able to locate their record and infer other information contained in the dataset. This risk increases as information accumulates across survey rounds and respondents' trajectories become more distinctive.

To mitigate this risk while preserving the dataset's research value, the data were processed through an anonymisation pipeline. A record passes through the following stages:

1. Reduction in the Number of Longitudinal Rounds: As the number of interview rounds increases, the number of possible combinations of respondent characteristics across time grows, making individual longitudinal trajectories potentially more distinctive and increasing re-identification risk. To reduce this risk, the pipeline limits the number of interview rounds included in the published dataset: the baseline round is retained in full, and for each subsequent round group, one round per respondent is selected.

2. Structural Categorisation and Aggregation: To reduce the granularity of the source data, the pipeline groups certain variables into broader categories.

3. FCLA Anonymisation ¹: At this stage, the Fast Clustering Longitudinal Anonymisation (FCLA) method is applied to identify and reduce disclosure risks in the longitudinal data. The method uses five variables that may indirectly identify individuals (referred to as quasi-identifiers; see Chapter 4): host country group, region of origin, current region (for respondents residing in Ukraine only), displacement period and gender. These variables are used to construct each respondent's trajectory across windows of three consecutive survey rounds. Each window is evaluated against two types of risk: whether too few respondents share the same combination of quasi-identifier values, potentially allowing a record to be identified, and whether particular values of sensitive variables, such as income⁶ documentation status or other (see Chapter 4), are over-represented within such a

¹ The core principles of the Fast Clustering Longitudinal Anonymisation (FCLA) algorithm are described [here](#)

group relative to their share in the dataset as a whole. Where a window is flagged, FCLA reduces the identified risk by further generalising quasi-identifiers, i.e. replacing more detailed values with broader categories where needed. Wherever possible, generalisation is applied only to the quasi-identifier values contributing to the flag, rather than to the entire column.

The process is iterative: at each iteration, the engine ranks the available generalisation steps according to how many flagged windows they resolve relative to how much detail they remove and applies the highest-ranked step.

4. Fallback Suppression: Where generalisation cannot resolve a flagged window, the pipeline deletes the affected rows.

2. Original Longitudinal Dataset Structure

The dataset follows the same group of respondents, each identified by a unique LONGIT_ID, across consecutive data-collection waves, or "rounds" (R). The panel covers rounds R1 to R34, conducted between March 2022 and October 2025, and follows 13,730 respondents: refugees interviewed outside Ukraine and returnees interviewed inside the country. Each row is one completed interview; rounds in which a respondent did not take part are not represented.

3. Preprocessing the Dataset Before Anonymisation

The sections below outline the additional steps taken with the original dataset prior to implementing the anonymisation pipeline. This includes the unification of the structure of the survey rounds, the reduction in the number of rounds, as well as generalisation of certain response options.

Unification of the survey rounds

In the initial stages of the survey, refugees (those outside Ukraine) and returnees (those who had returned to Ukraine) were interviewed within shared, synchronised survey rounds. Starting from Round 17, data collection was conducted separately for the refugee and returnee subsamples, resulting in two parallel sets of survey rounds with different timing.

Before reducing the number of longitudinal rounds, these separate refugee and returnee rounds were grouped back into common survey periods. This was necessary to create a consistent set of longitudinal periods across both subsamples, so that the subsequent round reduction could be applied uniformly.

Reduction in number of rounds

Over four years of data collection, the study included 34 survey rounds, with respondents interviewed as frequently as once a month during some periods. Retaining all available rounds would therefore result in long and increasingly detailed longitudinal trajectories, with a large number of possible combinations of respondent characteristics across time. Reducing the number of rounds limits the distinctiveness of these trajectories and the associated re-identification risk, while retaining observations from across the study period.

The framework groups individual interview waves into ten chronological windows, referred to as *round groups* (see Table 1). The ten round groups vary in size and duration, reflecting differences in how frequently respondents' circumstances changed over the study period. More rounds were retained for the earlier periods, when respondents' circumstances changed more frequently, while later rounds were combined into broader groups as changes between consecutive rounds

became less frequent. As a result, the number of rounds and the time period covered by each group vary across the study period, as shown in the table below.

Table 1. Selected round grouping type

Round Group	Rounds (R) included	Start date of the Round Group	End date of the Round Group	Months, total
1	R1	27.03.2022	30.06.2022	3.2
2	R3–R4	07.07.2022	03.09.2022	1.9
3	R5–R6	14.09.2022	06.11.2022	1.8
4	R7–R8	16.11.2022	04.01.2023	1.6
5	R9–R12	13.01.2023	01.05.2023	3.6
6	R13–R16	09.05.2023	10.09.2023	4.1
7	R17–R20	21.09.2023	04.01.2024	3.5
8	R21–R26	17.01.2024	24.06.2024	5.3
9	R27–R31	09.07.2024	20.12.2024	5.5
10	R32–R34	15.05.2025	22.10.2025	5.3

The reduction of rounds' frequency was applied by randomly selecting one round per participant within each subsequent round group. As a result, each respondent contributes at most one record per group. Across the ten groups, this leaves each respondent with at most ten records instead of up to thirty-four, shortening the sequences processed by the privacy engine. The configuration retains 60,400 of the original 120,467, a reduction of 50%. The impact of round reduction on the distribution of key demographic and socio-economic variables is outlined in Chapter 5.

Pre-emptive Granularity Reduction

Some variables outside the quasi-identifier set may still contain highly specific or rare response categories that could contribute to re-identification. To limit residual disclosure risk, selected variables were therefore generalised before entering the

privacy engine. This included merging highly specific response categories into broader groups, particularly where a category applied to very few respondents and retaining it separately provided limited analytical value. This pre-emptive reduction preserves the analytical utility of the variables while avoiding unnecessary detail that could facilitate linkage with external information. For example, households with a high number of members or children were grouped into broader categories. Similar recoding was applied to selected variables on documentation status, accommodation, rent, occupation, needs and other characteristics where individual response categories were rare.

4. Applying the FCLA algorithm

Core Logic of the (K, β)-L Privacy Framework

The framework divides the variables into two functional categories according to their disclosure risk: quasi-identifiers and sensitive attributes.

Quasi-Identifiers (QIs). Quasi-identifiers are variables that do not directly identify an individual but may enable re-identification when combined with other information. For example, a combination of *age*, *location*, *occupation*, and *household composition* may correspond to a specific individual when matched against information available from external sources. In longitudinal datasets, re-identification risk may increase because quasi-identifiers are recorded repeatedly across rounds. While individual values may be common, their combination and changes over time can create a distinctive sequence shared by only a small number of respondents.

The core tracking model selects five critical QIs to form this multi-dimensional sequence matrix:

- *country_group*: The respondent's host country, initially grouped into seven individual countries and two broader categories (Other Europe and Other world), and subsequently generalised to five individual countries and the same two broader categories where required.
- *region_group*: The oblast where the respondent lived before displacement, initially mapped to one of ten geographic groups and subsequently generalised to broader groupings where required.
- *region_current_group*: The oblast where the respondent lives at the time of the interview, mapped to the same macro-regions and subject to the same generalisation hierarchy. Recorded only for respondents inside Ukraine.
- *date_displ_q*: The period in which displacement began, grouped by four-month blocks for 2022 (Jan-Apr 2022; May-Aug 2022), and a single block covering the remainder (Sep 2022 and later).
- *gender_group*: Gender of the respondent.

Sensitive Attributes (SAs). Unlike quasi-identifiers, which may enable respondents to be distinguished or linked to external information, sensitive attributes are variables whose values are themselves considered potentially harmful if inferred or disclosed. Rather than being generalised, these variables are therefore protected through constraints on their distribution within groups defined by identical quasi-identifier sequences. The engine does not modify sensitive attribute values themselves. Instead, for each such group, it compares the relative frequency of each sensitive value with its frequency in the dataset as a whole and requires the within-group share not to exceed the overall share by more than a specified multiple. This constraint limits the extent to which group membership can increase an observer's ability to infer a respondent's sensitive attributes.

Four variables are treated as sensitive attributes:

- *income_bucket*: household income, recoded in income bands;
- *doco_status*: legal status in the host country;
- *debts*: household debt availability;
- *ranq1_income_source*: primary source of household income.

The pipeline applies a specialised (K, β) -L privacy framework implemented through the Fast Clustering Longitudinal Anonymisation (FCLA) engine. Rather than evaluating individual records independently, FCLA operates over a moving temporal window of length L , combining a respondent's values across consecutive survey rounds into a single longitudinal trajectory for privacy assessment. This allows disclosure risks arising from combinations of information over time to be identified and addressed, including risks that would not be detected when each survey round is assessed separately.

The framework uses three parameters to define the required level of protection while preserving analytical utility:

- **The K-Anonymity Metric ($K = 10$):** Each three-round sequence of quasi-identifier values must occur at least ten times across the panel. Sequences occurring fewer than ten times are treated as K-anonymity violations and flagged by the engine.
- **The β -Privacy Constraint:** While K-anonymity ensures that a sufficient number of respondents share each quasi-identifier sequence, it does not prevent a sensitive attribute from being highly concentrated within that group. A group may therefore satisfy $K = 10$ while still creating a disclosure risk if its members disproportionately share the same sensitive value, such as a particular income band or legal status. β limits this concentration by constraining the extent to which a sensitive value may be over-represented within a group of identical quasi-identifier sequences relative to its share in the dataset as a whole. β is not fixed in advance: the engine searches upward from 1.2 and reports the lowest

value at which no group breaches the limit. For this release, that value is 6.0. In practice, most group–value combinations remain well below this threshold: across the four sensitive attributes, the median within-group prevalence of each value is close to its dataset-wide prevalence, while fewer than five per cent of group–value combinations exceed five times the corresponding dataset-wide rate. Ratios approaching the $\beta = 6.0$ bound are associated with values that are rare in the overall dataset, for which even a single occurrence within a small group can produce a relatively large ratio.

- **The Moving Window Length ($L = 3$):** This defines the continuous chronological span monitored by the algorithm. An L value of 3 means the engine evaluates every three consecutive round groups per respondent as one sequence.

Rationalisation of Parameters Choice

Multiple configurations of K and L were tested to assess the trade-off between privacy protection and data utility. The final configuration of $K = 10$ and $L = 3$ was selected based on the resulting levels of suppression, statistical distortion, and protection against longitudinal disclosure risks.

1. Selection of K :

- **$K = 5$:** The lower threshold resulted in limited suppression and higher record retention but provided smaller anonymity groups, increasing the potential for re-identification where quasi-identifier sequences correspond to relatively small or localised respondent groups.
- **$K = 15$:** Increasing the threshold required substantially larger anonymity groups and resulted in considerably higher suppression. The associated loss of observations also increased distributional distortion, measured using Jensen–

Shannon divergence (JSD; See Chapter 5 for details), across the main country cohorts.

- **K = 10:** The selected value provided an intermediate level of protection while maintaining suppression and distributional distortion within the defined data-quality thresholds. It was therefore retained as the baseline K value for the production configuration.

2. Selection of L:

- **L = 2:** A two-round window generated fewer unique longitudinal sequences and consequently required relatively little suppression. However, it only protects against disclosure risks observable across pairs of consecutive rounds and does not account for patterns that emerge over longer sequences.
- **L = 4:** Extending the window to four rounds substantially increased the number of unique quasi-identifier sequences. As fewer respondents shared the same four-round trajectories, considerably more suppression was required to satisfy the K threshold, reducing both record retention and longitudinal continuity.
- **L = 3:** The three-round window provided a balance between longitudinal protection and data retention, capturing disclosure risks across multiple consecutive rounds without the substantially higher suppression observed under L = 4. It was therefore selected for the production configuration.

Generalisation and Suppression

When the engine identifies a privacy violation, it applies a two-stage procedure in which local generalisation is attempted first, and suppression is used only when generalisation cannot resolve the violation.

Phase 1: Generalisation

For flagged sequences, the FCLA engine uses predefined generalisation hierarchies to replace specific quasi-identifier values with broader categories where required to satisfy the privacy constraints. Generalisation is applied locally at the record level rather than across an entire variable, allowing unaffected records to retain their original level of detail.

For example, *region_group* and *region_current_group* initially classify oblasts into ten geographic groups. Kharkivska, Kyiv, Odeska, Dnipropetrovska and Donetska are retained as separate categories, while the remaining oblasts are grouped into Other South, Other East, Centre, West, and Other North. At subsequent levels of the generalisation hierarchy, selected categories are progressively combined, reducing the number of distinct regional groups from ten to six. For example, Centre and West are combined at one level, while at a later level Odeska is combined with Other South. This allows the engine to move from a more detailed to a broader geographic classification where necessary to satisfy the privacy constraints, before resorting to suppression.

The same principle applies to *country_group*, while *gender_group* and *date_displ_q* remain fixed and are not generalised.

Phase 2: Suppression

Where all available levels of generalisation have been exhausted, and a sequence still does not satisfy the privacy constraints, the affected records are suppressed. Suppression therefore acts as the final step of the procedure for violations that cannot be resolved while retaining the record.

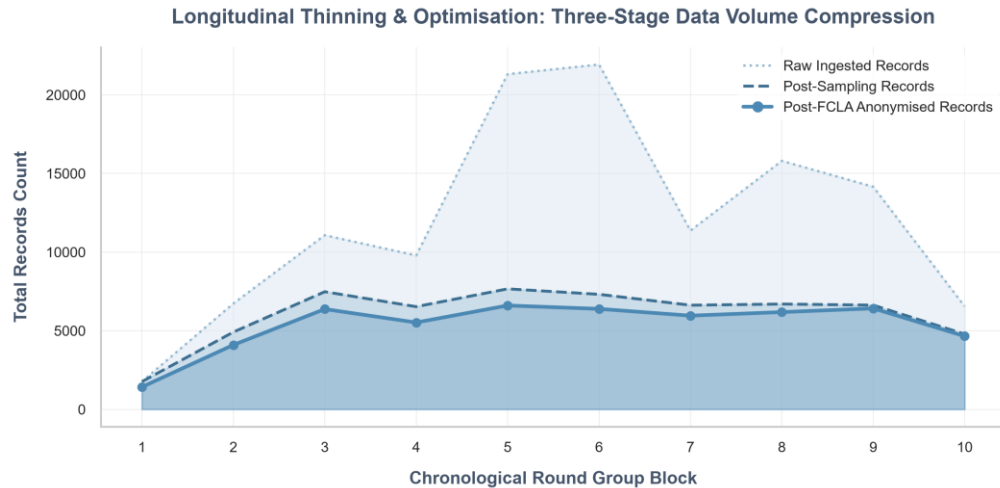
Local generalisation resolved 761 of the 2,556 flagged sequences over seven iterations, with all changes applied at the record level rather than across entire variables. The remaining violations required suppression, resulting in the removal of 6,755 records, equivalent to 11.2% of the records entering the engine. Although generalisation is prioritised to preserve observations and minimise information loss, suppression accounted for the larger share of adjustments required to satisfy the privacy constraints in the production run.

5. Assessment of Data Loss and Distributional Impact

The preparation of the released dataset involves two stages that may affect its size and distributions: round sampling, which reduces the longitudinal history to ten round groups by retaining one observation per respondent within each group, and the subsequent FCLA anonymisation, which applies generalisation and suppression to satisfy the privacy constraints. The impact of these transformations was assessed through record retention, the number of unresolved privacy violations, and changes in variable distributions before and after processing.

Of the 60,400 records entering the privacy engine, 53,645 were retained in the released dataset, corresponding to a suppression rate of 11.2%. The number of sequences below the K threshold decreased from 2,556 before processing to zero in the final dataset, with no remaining K or β violations. Figure 1 shows the reduction in records across the preceding round-sampling step and subsequent FCLA processing for each round group.

Figure 1. Records retained at successive stages of the pipeline, by round group. The figure compares the number of records initially available, retained after round sampling, and retained after FCLA anonymisation.

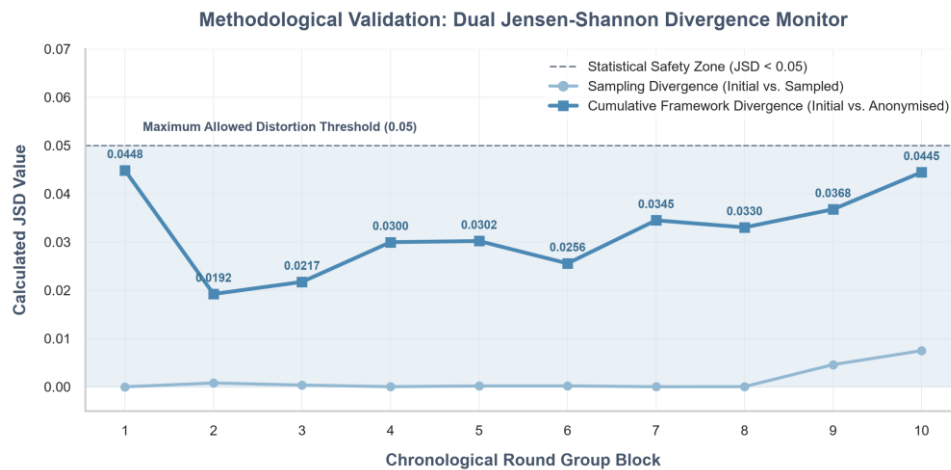


Distributional change was assessed using the Jensen–Shannon Divergence (JSD)², with 0.05 adopted as the threshold for material distortion for this release. The assessment distinguishes between the effect of round sampling, which reduces repeated observations within round groups, and the combined effect of round sampling and subsequent suppression by the privacy engine.

Round sampling had a limited effect on country distributions within round groups, with JSD values ranging from 0 to 0.0075. After suppression by the privacy engine was also taken into account, divergence increased but remained below the adopted threshold in all round groups, ranging from 0.019 to 0.045 (Figure 2).

² JSD is an information-theoretic measure of the difference between two probability distributions (Lin, 1991), where a value of 0 indicates identical distributions and larger values indicate greater divergence.

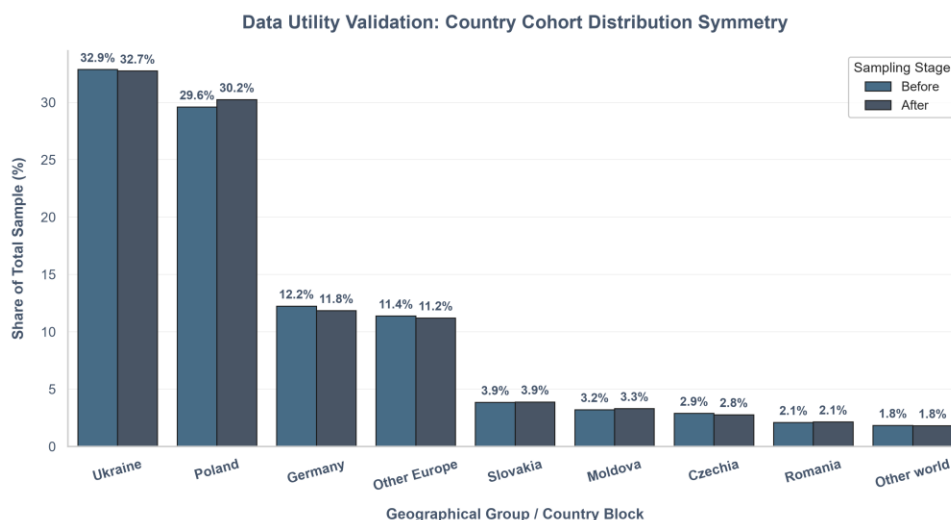
Figure 2. Jensen–Shannon Divergence in country distributions by round group.



Legend: The light blue line shows the change associated with round sampling, while the dark blue line shows the cumulative change after both round sampling and FCLA anonymisation. The dashed line indicates the 0.05 threshold adopted for this release.

The limited effect of round sampling on country composition is also visible in the aggregate country distribution, with only small differences in country shares before and after sampling (Figure 3).

Figure 3. Country distribution before and after round sampling. Bars show each country or country group as a share of the sample before and after the round-sampling procedure.



As an additional check, JSD was calculated across the variables processed by the pipeline that were not subject to generalisation. Across 92 such variables, the median divergence was 0.0001 and the maximum was 0.015, with no variable exceeding the 0.05 threshold. Quasi-identifiers were assessed separately because changes in their distributions may result directly from the generalisation procedure and therefore represent an intended transformation rather than unintended distortion.

Taken together, these checks show that the final dataset satisfies the specified privacy constraints while the observed distributional changes remain within the thresholds adopted for this release.