Kurdish Studies
December 2021
Volume: 09, No: 02, pp 419-447
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.67145/ks.v9i2.4142

Certified bipartite alignment and cross-database transfer in electronic health record integration

Luis Eduardo Muñoz Guerrero[1]

[1]Universidad Tecnológica de Pereira, Pereira, Colombia. Email: lemunozg@utp.edu.co. ORCID: https://orcid.org/0000-0002-9414-6187

Abstract

Background: three governorate profiling exercises documented displacement in the Kurdistan Region of Iraq, and their household microdata have never been reanalysed. Methods: an instrumented pipeline audits provenance, aligns the two admissible instruments by maximum-weight bipartite matching with a linear-relaxation optimality certificate, and tests cross-governorate transfer of stated non-return. Results: identical labels recover 6 of 84 correspondences; the certified alignment recovers 60 with no false pair and all 84 at a precision of 0.966. Non-return is predictable within a governorate and no better than chance across them. Conclusions: comparability must be computed, not assumed.

Keywords: bipartite matching; optimality certificate; instrument alignment; forced displacement; Kurdistan Region of Iraq

1. Introduction

By the middle of 2016 the Kurdistan Region of Iraq was hosting the largest share of Iraq's displaced population. According to the profiling reports themselves, the arrival of Iraqis fleeing the advance of the Islamic State and of Syrians fleeing the war next door raised the urban population of some governorates by as much as thirty per cent in two years, and most of the unfinished buildings, invisible to the camp-based information systems that the response had been built around.

In response, the governorate authorities of Erbil, Duhok and Sulaymaniyah, working with the Kurdistan Region Statistics Office and with UNHCR, UN-Habitat, OCHA, UNFPA and IOM, carried out three collaborative urban profiling exercises between December 2015 and June 2016. Each surveyed roughly twelve hundred households drawn in equal parts from displaced Iraqis, Syrian refugees and the host community, and each produced a descriptive report. The household microdata were released on the Humanitarian Data Exchange under a Creative Commons Attribution licence and have been publicly downloadable since February 2017.

They have also been almost entirely unused. Searching the indexed literature for the exact names of the three exercises, in Semantic Scholar and OpenAIRE, returns the profiling reports themselves and nothing else, and a search of public code repositories for the same names returns none that loads the files. Research on displacement in the Region has drawn on purpose-built surveys, on clinical and demographic records and on qualitative interviewing, but not on these files. The one report examined here in full contains frequencies, cross-tabulations and maps, and no model, index or interval in its eighty-nine pages. The Region's most detailed picture of out-of-camp displacement has therefore never been asked a question that the reports did not already answer.

One such question is whether the three governorates can be spoken about together. Policy documents, appeals and scholarship routinely generalise from one governorate to the Region, and the profiling data are in principle the right evidence to test that practice.

In practice the test cannot begin until the instruments are made comparable, and they are not comparable as published.

The Erbil questionnaire carries 101 household-level items and the Sulaymaniyah questionnaire 105, and only 6 of them share an identical label. Everything else that is in fact the same question is worded differently, so the comparison rests on a correspondence that somebody has to establish.

Establishing it by hand is neither reproducible nor neutral. Establishing it by computation is a problem with a name: given a similarity between every pair of items, choose the set of pairings of maximum total similarity in which no item is used twice.

That is maximum-weight bipartite matching, it is solvable exactly in polynomial time, and, because the assignment polytope is integral, the linear relaxation supplies a certificate that the solution found is the best one and not merely a good one. The distinction matters here more than it usually does, because an alignment is an assumption that everything downstream inherits.

The question of comparability is not a technicality peculiar to these three files. Profiling exercises are designed collaboratively with the authorities of the place they cover, which is what makes them useful and also what makes each one slightly different from the last.

Questions are added because a governorate asks for them, dropped because the enumeration window is short, and reworded because a translation reads badly. The result is a family of instruments that are recognisably siblings and nowhere identical, and a body of evidence whose internal comparability is nobody's explicit responsibility. The same pattern appears wherever

humanitarian assessment is decentralised, so the machinery reported here is not specific to the Kurdistan Region even though the finding is.

Making heterogeneous instruments comparable is a recognised problem outside humanitarian work. Database research calls it schema matching and has studied it for decades, combining evidence from element names, data types, value distributions and structure, and the record-linkage literature has developed a parallel apparatus for deciding when two records describe the same entity.

What that work supplies here is not a new technique but a discipline: the matching is a first-class artefact, it is evaluated against a labelled reference rather than asserted, and the evidence it uses is stated so that a reader can disagree with it precisely.

There is also a reason to care about the particular question studied on the aligned data. Whether displaced households intend to return is the hinge of every durable-solutions argument: it decides whether a response is framed as temporary relief or as integration, and in the Kurdistan Region in 2016 it decided whether municipal services were planned for a population that would leave or one that would stay.

Intentions are imperfect predictors of behaviour, but they are what a household survey can collect, and they were collected here from more than a thousand displaced households in two governorates at a moment when the answer was genuinely open.

This paper reports four things. First, an audit of the three published files, run before any analysis, which established that one of them cannot be used as labelled. Second, a certified optimal alignment of the two admissible instruments, evaluated against a manual adjudication of every pair it proposes.

Third, a test of whether a household's stated intention not to return to its place of origin can be predicted from its circumstances, and whether that prediction survives being moved from one governorate to the other.

Fourth, a pipeline that produces all of it from the published files with one command, with measured cost and with automated gates that fail the build rather than warn.

2. Materials and methods

Three files were retrieved from the Humanitarian Data Exchange, 8,493,482 bytes in all (8.1 MB), each recorded with its retrieval address, licence, publication date, coverage window and SHA-256 digest before it was read. Table 1 is that record. A download whose digest does not match the recorded value fails the build rather than warning, because a file that is not the file the study was run on is a different file.

Table 1. The three files as retrieved, with the leading hexadecimal digits of the SHA-256 digest recorded before each was read. All three were published on 2017-02-09 under a CC BY 4.0 licence. The household counts are those announced on the dataset pages, not those found in the files; Table 2 reports what the files contain.

File as published	Erbil	Duhok	Sulaymaniyah
Bytes	2,622,675	2,976,637	2,894,170
SHA-256, first eight digits	526a8a9d	3f6567b4	0a5b7f81
Coverage from	2015-12-01	2016-06-01	2016-06-01
Coverage to	2016-01-31	2016-06-30	2016-06-30
Households announced	1,163	1,205	1,201

Each file holds three sheets: a description, a data dictionary and the records. The records are not households. They are household members, between five and six rows per household, with the household-level answers repeated down the block, so the first step is an aggregation keyed on the household identifier.

In the Sulaymaniyah file the dictionary's name column is positional, the data sheet's header row is the integers one to 105 plus the individual-level items, and the household identifier is an undocumented column near the end; the schema has to be reconstructed by position before anything can be read from it.

Provenance was then audited by assigning every row to a governorate on the evidence of its own district field rather than on the name of the file that contained it. The audit is reported in Table 2. Two gates follow it.

The first compares the resulting distribution against the distribution recorded when the study was run and fails on any change. The second checks, for each site, the fraction of households whose declared size equals the number of member records actually present, and admits a site only above 99 per cent.

Table 2. What each published file actually contains, by the district recorded on each row rather than by the name of the file. The size check is the percentage of households whose declared size equals the number of member records present; admission requires 99 per cent. The file published as Duhok carries 5,974 rows of Erbil districts, under the same household identifiers as the Erbil file.

File as published	Rows	Rows in own districts	Households	Size check	Admitted
Erbil	5,974	5,974	1,163	100.0%	yes
Duhok	7,062	1,088	213	51.2%	no

Sulaymaniyah	6,258	6,258	1,201	100.0%	yes

Notation is collected in Table 3.

Table 3. Symbols used in the formulation of the alignment problem and in the evaluation of transfer. The metadata term of the similarity is written with a Greek mu, and the mixture of two response distributions with a bar, so that the capital M is left to mean the matching and nothing else.

Symbol	Meaning
A, B	household-level item sets of the two instruments
l_i	normalised label of item i
T_i, G_i	token set and character trigram set of that label
R_i, V_i	observed response set of item i, and the values it is built from
θ_i, k_i	topic and question type recorded for item i
ν_i	fraction of the values of item i that parse as numbers
Σ	stop list removed from the token sets
s_{ij}	similarity between item i in A and item j in B
μ_{ij}	the metadata term of that similarity
x_{ij}	one if items i and j are paired, zero otherwise
M, M^{+}	the matching selected, and the pairs of it adjudicated correct
u, v	dual potentials of the assignment relaxation
τ	similarity threshold above which a pair is accepted
y	one if the household would not consider return
π	prevalence of the positive class
$\hat{s}$	score the model gives a household
β, ω	model coefficients and class weights
F, S_X	terms common to both encoders, and those site X retains
$P, Q, \overline{P}$	two response distributions and their mixture
JS	Jensen-Shannon divergence between two distributions
AP	average precision, the area under the precision-recall curve

The two instruments are recognisably siblings and nowhere identical, and Figure 1 says how.
They ask about the same topics and allocate a different number of questions to each, and they record their answers in different formats. Neither difference is a defect; both are what makes an alignment necessary.

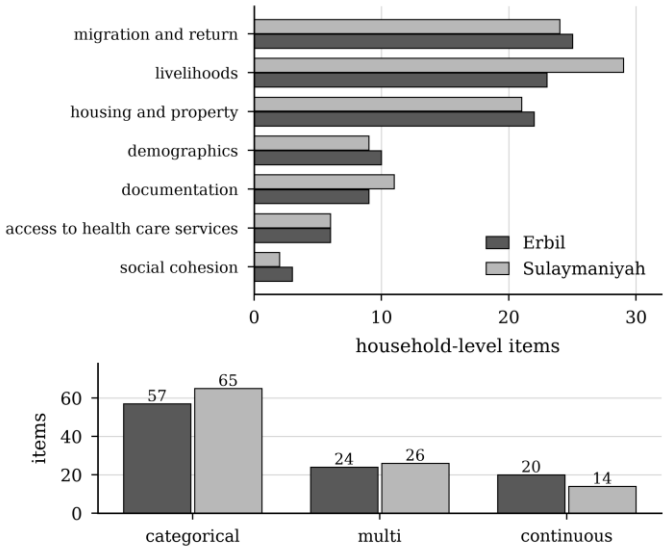

Figure 1. What each questionnaire contains. Upper panel: household-level items by the topic the data dictionary files them under. Lower panel: how the answers are recorded. Sulaymaniyah devotes more questions to livelihoods and Erbil more to migration history, and Erbil records more of its answers as free numbers where Sulaymaniyah offers categories. Only 6 items of roughly a hundred carry an identical label.

Let A and B be the household-level item sets of the two admissible instruments, with $|A| = 101$ and $|B| = 105$. For every ordered pair the similarity combines four pieces of evidence, three drawn from the documentation and one from the records themselves, as in Equation 1.

$$s_{ij} = w_1 J(T_i, T_j) + w_2 D(G_i, G_j) + w_3 \mu_{ij} + w_4 J(R_i, R_j) \qquad (1)$$

Table 4 sets out what each term compares and the weight it was given, fixed before the alignment was inspected. The fourth term is the only one that does not depend on how somebody happened to word the dictionary, and it is what separates two items whose labels look alike but whose answer sets do not.

Table 4. The four terms of the similarity of Equation 1 and the weights they were given before the alignment was inspected. Three read the documentation and would rank two items as similar because somebody worded their labels alike. The fourth reads the records and is the only one that cannot be fooled that way, which is why it carries as much weight as the token overlap.

Term	Compares	Evidence	Weight
$J(T_i, T_j$	content-word token sets of the two labels	dictionary	0.30
$D(G_i, G$	character trigram sets of the two labels	dictionary	0.25
μ_{ij}	topic and question type recorded for each item	dictionary	0.15
$J(R_i, R$	response values actually observed in the records	records	0.30

The two set coefficients are the usual ones, Equation 2 and Equation 3, with one convention worth stating. The Jaccard coefficient of two empty sets is taken to be zero rather than one. Under the usual convention two items with no observed responses at all would score the maximum on the fourth term, which is precisely the term the study leans on as the one that cannot be talked into agreement.

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \qquad J(\emptyset, \emptyset) = 0 \qquad 2)$$

$$D(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \qquad D(X,\emptyset) = D(\emptyset,Y) = 0 \tag{3}$$

The two label sets are Equation 4 and Equation 5. Labels are first normalised: accents stripped, bracketed text removed, lower-cased, and everything that is not a letter, a digit or a space discarded. The token set then drops a stop list and any token of two characters or fewer. The stop list contains the word household, which appears in most labels of both instruments and would otherwise make every pair of items look alike. The trigram set is taken over the label padded with two blanks at each end, written with a tilde, so that the first and last characters of a label carry weight instead of appearing in fewer trigrams than the ones in the middle.

$$T_i = \{t \in \mathrm{tok}(\ell_i) : t \notin \Sigma,\ |t| > 2\} \tag{4}$$

$$G_i = \left\{ \tilde{\ell}_i[k : k+2] \mid 1 \le k \le \left|\tilde{\ell}_i\right| - 2 \right\} \tag{5}$$

The third term, Equation 6, rewards agreement of the topic and of the question type recorded for each item in its dictionary, half a point each.

$$\mu_{ij} = \frac{1}{2}\mathbf{1}[\theta_i = \theta_j] + \frac{1}{2}\mathbf{1}[\kappa_i = \kappa_j] \tag{6}$$

The fourth term compares the sets of responses the two items actually received, and how that set is formed is a modelling choice rather than a fact about the data. Equation 7 gives it. Writing V for the non-empty values recorded for an item and for the fraction of them that parse as numbers, an item that is numeric in more than four fifths of its values collapses to a single marker, because comparing two sets of numeric literals measures nothing; otherwise the set is capped at the forty most frequent normalised responses, so that a long tail of typing errors cannot dominate the overlap.

$$R_i = \mathrm{top}_{40}(V_i) \ \text{ if } \nu_i \le 0.8, \qquad R_i = \{\mathrm{num}\} \ \text{ if } \nu_i > 0.8 \tag{7}$$

The alignment is then the maximum-weight bipartite matching whose objective is Equation 8, subject to the constraints of Equation 9, which say that no item of either instrument is used more than once.

$$\max_{x} \sum_{i \in A} \sum_{j \in B} s_{ij}\ x_{ij} \tag{8}$$

$$\sum_{j \in B} x_{ij} \le 1 \ \ \forall i, \qquad \sum_{i \in A} x_{ij} \le 1 \ \ \forall j, \qquad x_{ij} \in \{0,1\} \tag{9}$$

This is the classical assignment problem. It is solved here by the Jonker-Volgenant shortest augmenting path algorithm, a refinement of the Hungarian method, in time cubic in the number of items, which for instruments of this size is immaterial. What is not immaterial is that the solution can be certified. Relaxing the integrality constraint gives the linear program of Equation 10.

$$z_{\mathrm{LP}} = \max\{ s^{\top} x : Ax \le \mathbf{1},\ x \ge 0 \} \tag{10}$$

The constraint matrix of the assignment problem is totally unimodular and its polytope has integral vertices, so the relaxation and the integer problem share an optimum. A matching whose weight equals the value of the relaxation is therefore optimal.

The dual of that relaxation, Equation 11, assigns a price to every item of both instruments and asks for the cheapest set of prices under which no pair is worth more than the sum of its two.

$$z_{\mathrm{D}} = \min\{\mathbf{1}^{\top}u + \mathbf{1}^{\top}v : u_i + v_j \geq s_{ij} \;\; \forall (i,j),\; u \geq 0,\; v \geq 0\} \tag{11}$$

Its optimal prices are potentials u and v that satisfy the two conditions of Equation 12: dual feasibility everywhere and complementary slackness on the chosen pairs.

$$u_i + v_j \geq s_{ij} \;\; \forall (i,j), \qquad u_i + v_j = s_{ij} \;\; \forall (i,j) \in M \tag{12}$$

Strong duality closes the argument, as in Equation 13. The three quantities coincide, so exhibiting the prices exhibits a proof: any matching at all has weight at most the sum of the potentials, and this one attains it.

$$z_{\mathrm{LP}} = z_{\mathrm{D}} = \sum_{i \in A} u_i + \sum_{j \in B} v_j \tag{13}$$

The certificate is worth the few lines it costs because the whole study rests on the alignment, and a reader is entitled to know that no better one exists under the stated similarity rather than to be told that a good one was found. It is also worth noting where the tractability ends. With two instruments the problem is polynomial. With three it becomes axial three-dimensional assignment, which is NP-hard, and the exact certificate of Equation 12 is no longer available; approximation would be the only option.

The present study needs only the two-instrument case, and it needs it for a reason given in the results rather than by design. The three-instrument case deserves more than a sentence, since it was the original design and would be the design again if the Duhok file were repaired.

With three instruments the natural objective, Equation 14, sums the three pairwise similarities of each triple, subject to the constraints of Equation 15, which say that no item of any instrument is used more than once. This is axial three-dimensional assignment, for which no polynomial algorithm is known.

$$\max_{x} \sum_{i} \sum_{j} \sum_{k} \left(s_{ij}^{12} + s_{ik}^{13} + s_{jk}^{23} \right) x_{ijk} \tag{14}$$

$$\sum_{j,k} x_{ijk} \leq 1 \;\; \forall i, \qquad \sum_{i,k} x_{ijk} \leq 1 \;\; \forall j, \qquad \sum_{i,j} x_{ijk} \leq 1 \;\; \forall k \tag{15}$$

A useful bound survives the loss of the exact certificate. Let OPT be the value of the best three-way assignment and let z12, z13 and z23 be the optima of the three two-instrument problems, each of which is computable exactly.

Projecting an optimal three-way assignment onto a pair of instruments yields a feasible two-instrument assignment, so the weight it contributes is at most that pair's optimum; summing the three projections recovers OPT exactly, which gives Equation 16.

Any heuristic three-way solution can thus be certified against a bound obtained from three polynomial computations, which is weaker than the exact certificate available here but is not nothing.

$$\mathrm{OPT} \leq z_{12} + z_{13} + z_{23} \tag{16}$$

Algorithm 1 states the procedure, including the certificate check that makes the build fail if the matching cannot be shown optimal.

```
Algorithm 1  Certified instrument alignment

Input  item sets A, B; labels, topics, types;
       response sets R; weights w; cut t
Output matching M plus optimality certificate
 1  for i in A, for j in B
 2    s[i][j] <- similarity, Equation 1
 3  M <- max-weight matching on s
 4  z <- optimum of relaxation, Equation 4
 5  (u,v) <- dual prices of relaxation
 6  assert |sum of s over M - z| < eps
 7  assert u[i]+v[j] >= s[i][j] all i,j
 8  assert u[i]+v[j] == s[i][j] on M
 9  return {(i,j) in M : s[i][j] >= t}
Invariant  after line 3 no item is paired
twice; lines 6-8 hold iff M is optimal.
```

Two thresholds are used in what follows and both are chosen by a rule rather than by eye. Writing M for the matching and M-plus for the subset of it the adjudication accepted, the pairs kept at a threshold are Equation 17, and precision and recall against the adjudication are Equation 18.

$$M_\tau = \{(i,j) \in M : s_{ij} \geq \tau\} \tag{17}$$

$$\mathrm{pre}(\tau) = \frac{|M_\tau \cap M^+|}{|M_\tau|}, \qquad \mathrm{rec}(\tau) = \frac{|M_\tau \cap M^+|}{|M^+|} \tag{18}$$

The analysis uses the loosest threshold at which no accepted pair was adjudicated wrong, Equation 19, so that nothing downstream rests on a pairing the author judged wrong.

The other threshold reported is the one that maximises the harmonic mean of precision and recall, Equation 20; it recovers every genuine correspondence at the cost of a few false pairs, and it is quoted to show what the matching can do rather than to carry the analysis.

$$\tau^* = \min\{\tau : \mathrm{pre}(\tau) = 1\} \tag{19}$$

$$F_1(\tau) = \frac{2\,\mathrm{pre}(\tau)\,\mathrm{rec}(\tau)}{\mathrm{pre}(\tau) + \mathrm{rec}(\tau)}, \qquad \tau_F = \operatorname{argmax}_\tau F_1(\tau) \tag{20}$$

A matching is a perfect pairing on the smaller side, so it pairs every item whether or not a counterpart exists. Some pairs are therefore forced rather than genuine, and the similarity threshold is what separates the two.

Rather than choose the threshold by inspection, all 101 proposed pairs were adjudicated by hand by the author, reading the two labels of each pair against the questionnaire, before any threshold was fixed. A pair counts as correct when the two labels designate the same underlying question, even if the wording or the reference period differs.

The adjudication is published with the code, so the threshold can be chosen against measured precision and recall instead of taste. The adjudication protocol was fixed before it began. The pairs were read in descending order of similarity, which is the order the matching itself produces, and each was judged on the two labels and, where the labels were ambiguous, on the question text carried in the dictionaries.

A pair was marked correct when the two items ask about the same construct, so that an item on expenditure over the previous week and one on expenditure over the previous month count as a correspondence even though their reference periods differ, and marked incorrect when no correspondence exists in either direction. No threshold was contemplated until every pair had been judged, and the judgements were not revised afterwards.

One property of the formulation deserves a warning. A matching maximises total similarity, which means it can accept a mediocre pair in order to free a better one elsewhere, and it has no notion of an item that simply has no counterpart. Both

behaviours are correct for the problem as posed and wrong for the problem as intended, and the threshold is what reconciles them.
An alternative formulation would add a dummy item on each side with a fixed similarity, which is equivalent to thresholding and hides the choice inside the objective rather than exposing it; the threshold is kept explicit here precisely so that it can be swept.
The outcome studied on the aligned core is a displaced household's answer to whether it would consider returning to its place of origin, coded one for no and zero for yes. The question is put only to the displaced strata, so a household with no answer recorded is excluded, as are the small number who answered that they did not know. The item itself and its two follow-up questions about conditions for return are removed from the predictors, as are the sampling weight and every geographic identifier, since the last would let a model recover the site instead of the household.
Two model families are fitted: an L1-penalised logistic regression, which yields a sparse and inspectable set of coefficients, and histogram gradient boosting, which does not assume additivity. The whole preprocessing chain is fitted inside each training fold, so that nothing learned from a test household reaches the model that scores it, and across governorates the model is fitted on the whole of one site and evaluated on the whole of the other, so no household ever appears on both sides.
The sparse model solves Equation 21. The penalty on the sum of absolute coefficients is what drives most of them to exactly zero, and the comparison of which ones survive in each governorate is the mechanism this paper ends up reporting, so the objective is worth writing down rather than naming.
The class weights of Equation 22 compensate for the small positive class, and the labels are recoded to plus and minus one for the loss.

$$\hat{\beta}, \hat{b} = \operatorname{argmin}_{\beta, b} \sum_{i} \omega_{y_i} \log\!\left(1 + e^{-\tilde{y}_i (\beta^{\top} x_i + b)}\right) + \frac{1}{C} \lVert \beta \rVert_1 \tag{21}$$

$$\omega_c = \frac{n}{2 n_c}, \qquad \tilde{y}_i = 2 y_i - 1 \tag{22}$$

With 16.1 per cent of the sample in the positive class, accuracy would be uninformative and the area under the receiver operating characteristic curve would flatter every model.
The headline metric is therefore average precision. Precision and recall at a score threshold are Equation 23, and average precision is the area they sweep out, Equation 24, which a constant predictor attains only at the prevalence.

$$P(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FP}(t)}, \qquad R(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)} \tag{23}$$

$$\mathrm{AP} = \sum_{k} (R_k - R_{k-1})\, P_k \tag{24}$$

Two threshold metrics are reported alongside it, both at the score cut of Equation 25, which is the quantile that makes the model predict as many positives as the data contain.
They are the Matthews correlation coefficient, Equation 26, which is the correlation between prediction and truth and is zero for any uninformative rule; and balanced accuracy, Equation 27, the mean of the two class-wise rates, which is one half for the same rules.

$$t^{*} = Q_{\hat{s}}(1 - \hat{\pi}), \qquad \hat{\pi} = \frac{1}{n} \sum_{i} y_i \tag{25}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (26)$$

$$\mathrm{BA} = \frac{1}{2}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{1}{2}\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \quad (27)$$

Three deliberately unintelligent baselines are reported for each site: the majority rule, a uniform random score, and the single best item chosen with hindsight on the evaluation set itself, Equation 28. The last maximises over every numeric item and both signs, on the very data it is scored on, which is why it is a demanding floor rather than a fair competitor.

$$\mathrm{AP}_{\max} = \max_{j} \max_{\varepsilon=\pm 1} \mathrm{AP}\left(y, \varepsilon\, x_{\cdot j}\right) \quad (28)$$

Uncertainty comes from 2,000 bootstrap resamples of the evaluation set, reported as the percentile interval of Equation 29. Every comparison also carries an explicit null: 1,000 permutations of the labels against the fitted scores, which gives the distribution of the metric when the model has learned nothing, summarised by the p-value of Equation 30.
The one added to numerator and denominator is not decoration: it keeps the p-value from ever being zero, and it is why the smallest value this paper can report is 0.0010 rather than nothing at all. Every setting either model makes is collected in Table 5.

$$\mathrm{CI}_{95} = \left[\, q_{2.5}(\theta_1^*, \ldots, \theta_B^*),\ q_{97.5}(\theta_1^*, \ldots, \theta_B^*) \,\right] \quad (29)$$

$$p = \frac{1 + \left|\left\{b : \mathrm{AP}\left(y^{(b)}, \hat{s}\right) \geq \mathrm{AP}(y, \hat{s})\right\}\right|}{B+1} \quad (30)$$

Table 5. Every choice the models make that is not forced by the data. The library versions are recorded because a default that changes between minor releases moves a result without anyone editing a line.

Choice	Setting
Sparse model	L1-penalised logistic regression, C = 0.10, liblinear, balanced class weights
Non-additive model	histogram gradient boosting, 300 iterations, depth 4, learning rate 0.06, L2 1.0
Numeric items	median imputation, then standardisation
Categorical items	one-hot encoding, levels seen fewer than 10 times pooled
Within a governorate	stratified 5-fold cross-validation, whole chain fitted inside each fold
Across governorates	fitted on all of one site, evaluated on all of the other
Uncertainty	2,000 bootstrap resamples of the evaluation set
Null model	1,000 permutations of the labels against the fitted scores
Items barred as leakage	12 patterns: the outcome, its follow-ups, the sampling weight, every geographic identifier
Random seed	20191231, fixed throughout
Software	Python 3.12.10, NumPy 2.5.3, SciPy 1.18.1, scikit-learn 1.9.1

Failure to transfer has two very different explanations, and separating them is the main analytical difficulty. The determinants of non-return may genuinely differ between governorates, which is a claim about households; or the same item may simply be coded differently in the two questionnaires, which is a claim about paperwork. Three diagnostics separate them. The first asks how well the site itself can be predicted from the aligned predictors. The second measures, item by item, the Jensen-Shannon divergence between the two sites' response distributions. It is built from the Kullback-Leibler divergence and the mixture of the two distributions, Equation 31, and is Equation 32. The logarithm is taken in base two, which bounds the divergence between zero and one and makes the numbers reported later readable as a fraction of the maximum possible disagreement.

$$D_{\mathrm{KL}}(P\|Q) = \sum_{r} P\,(r) \log_2 \frac{P(r)}{Q(r)}, \qquad \bar{P} = \frac{1}{2}(P+Q) \tag{31}$$

$$\mathrm{JS}(P,Q) = \frac{1}{2} D_{\mathrm{KL}}(P\|\bar{P}) + \frac{1}{2} D_{\mathrm{KL}}(Q\|\bar{P}) \tag{32}$$

Divergence is not the only way an item can fail to carry across. A model trained on one site can meet, in the other, response categories it has never seen, and those carry no coefficient at all. Equation 33 measures that separately as the share of the evaluation site's responses that lie outside the support of the training site's. An item can score high on one of the two and zero on the other, so both are reported.

$$u_j = \sum_{r \notin \mathrm{supp}(P_j)} Q_j\,(r) \tag{33}$$

The third repeats the whole transfer experiment after discarding the items whose divergence exceeds a sequence of ever stricter bounds. If the failure is an artefact of coding, it should ease as the incompatible items are removed. Every arbitrary choice in the pipeline is subjected to the same treatment: the alignment threshold is swept across six values and the divergence bound across five. The random seed is fixed at 20191231 throughout, and two runs of the pipeline produce identical artefacts. Every result the paper reports is written to a JSON artefact under results/ and the document interpolates from it; no figure reads a number that a person typed.
Wall-clock time and peak allocated memory are recorded for each stage and reported with the results. The pipeline is shown in Figure 2.

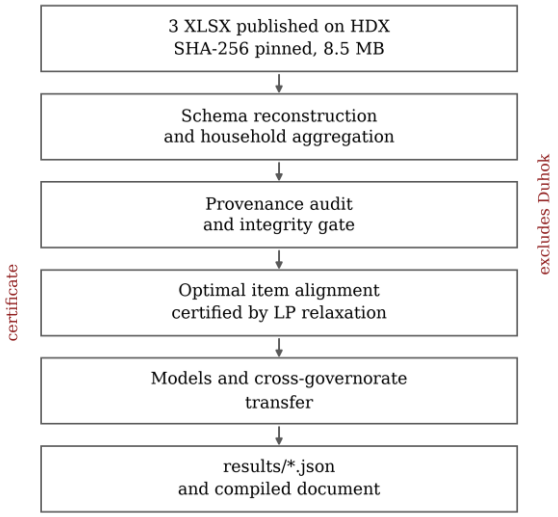

Figure 2. Stages of the pipeline, from the three published files to the compiled document. The integrity gate removes one of the three governorate files; the alignment stage emits an optimality certificate that the build checks before continuing.

3. Results

The provenance audit did not behave as expected. Of the 7,062 rows in the file published as the Duhok exercise, 5,974 (84.6 per cent) carry districts of Erbil governorate: Hawler, Deshti Hawler Bnslaho, Shaqlawa, Khabat, Koya and Soran, in exactly the counts found in the Erbil file, and under exactly the 1,163 household identifiers of the Erbil file. Only 1,088 rows lie in districts of Duhok. The dataset page announces a sample of 1,205 Duhok households; the file contains 213.
Figure 3 shows both the composition and the gate that judged it.

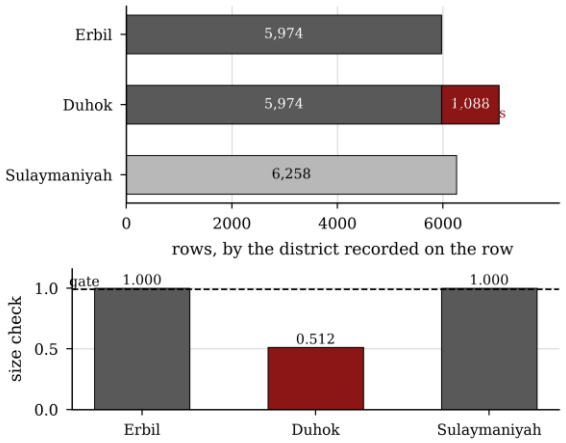

Figure 3. What the three published files contain and what the integrity gate makes of it. Upper panel: rows by the governorate recorded on the row itself, not by the name of the file. Lower panel: the fraction of households whose declared size equals the number of member records present, against the admission threshold. The file published as Duhok is 84.6 per cent Erbil rows, and the part that is genuinely its own fails the gate.

The integrity gate then settled what to do about it. In the Erbil and Sulaymaniyah files the declared household size equals the number of member records for every single household, 1,163 of 1,163 and 1,201 of 1,201. In the Duhok block it holds for 109 of 213, or 51.2 per cent, so the block does not group into coherent households at all and the gate excludes it.
The study therefore rests on two sites rather than three, and this is the reason the alignment problem stayed within the polynomial case described in Equation 8 rather than becoming the NP-hard three-way one.
Alignment quality is reported in Figure 4. Of the 101 pairs the matching proposes, 84 were adjudicated correct and 17 forced. The baseline that a reader might reasonably expect somebody to use, pairing items whose labels are identical, recovers 6 of the 84 genuine correspondences, a recall of 7.1 per cent. The certified matching recovers all 84 at a precision of 0.966 when the threshold is set at 0.35, and recovers 60 of them with no false pair at all from 0.60 upward. The analysis uses the latter, so that nothing downstream rests on a pairing the author judged wrong.

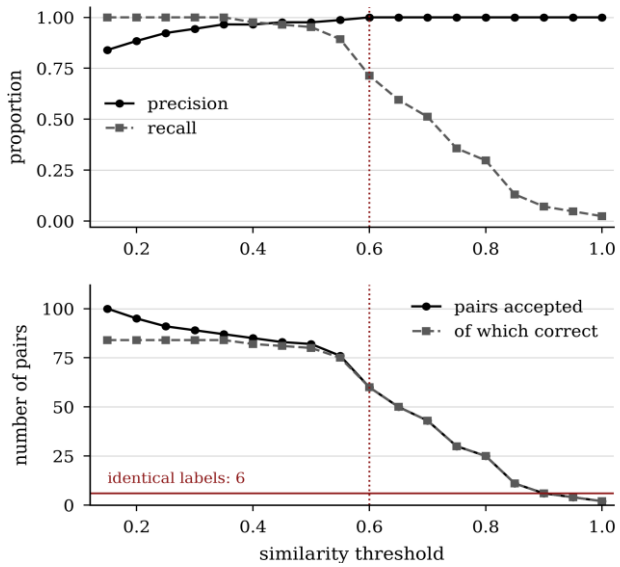

Figure 4. Alignment quality against the manual adjudication of all 101 proposed pairs, as the similarity threshold varies. Upper panel: precision and recall. Lower panel: pairs accepted and the number of them adjudicated correct, with the horizontal line marking the 6 pairs that identical labels alone would have found. The dotted vertical line is the threshold of 0.60 used in the rest of the paper.

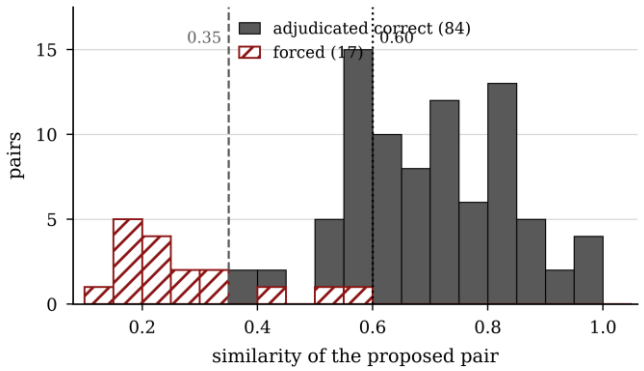

Figure 5. Why a threshold works at all. The similarities of the 84 pairs the adjudication accepted and of the 17 it rejected, as two distributions. They overlap only between the two thresholds the paper uses: below 0.35 nothing genuine is lost, and above 0.60 nothing forced is admitted.

Table 6. Alignment quality under four selection rules, against the manual adjudication of all 101 proposed pairs. The last row is the baseline a reader comparing the two questionnaires by eye would produce, and it recovers 7.1 per cent of the 84 genuine correspondences. It is not the same as the row above it: only 2 of those 6 pairs reach a similarity of 1.000, because the similarity also weighs the dictionary topic and the question type. The analysis uses the second row, the loosest threshold at which no accepted pair was adjudicated wrong.

Selection rule	Accepted	Correct	Precision	Recall
similarity at least 0.35	87	84	0.966	1.000
similarity at least 0.60	60	60	1.000	0.714
similarity exactly 1	2	2	1.000	0.024
identical labels only	6	6	1.000	0.071

Quality is not spread evenly over the questionnaire. Table 7 breaks the same adjudication down by the topic under which the data dictionary files each item.

The instruments agree best where the wording is administrative and least where it is descriptive: documentation recovers 0.889 of its genuine correspondences and housing 0.524, and it is housing that carries most of the study's substantive items.

Where the matching gives up is also visible, in livelihoods, which proposes 23 pairs of which only 13 are genuine, because the two instruments itemise spending and income differently and the matching pairs what is left.

Table 7. Where the alignment works, by the topic the data dictionary records for the Erbil item. Items is how many pairs the matching proposes in that topic, Correct how many of them the adjudication accepts, and Recovered how many of those survive the threshold of 0.60. Recall is the last divided by the one before it. Housing is the weak topic, where the two instruments word the same questions least alike; documentation is the strong one.

Questionnaire topic	Items	Correct	Recovered	Recall
Migration history plans and return	25	23	18	0.783
Livelihoods	23	13	11	0.846
Housing shelter and property	22	21	11	0.524
Demographics locations	10	8	6	0.750
Documentation	9	9	8	0.889
Access to health care services	6	5	4	0.800
Social cohesion	3	2	1	0.500
Education	1	1	0	0.000
Extrapolation factor	1	1	0	0.000
Population group	1	1	1	1.000

Table 8 gives the flavour of what the matching finds and where it fails. The easy correspondences are trivially easy, and identical labels would have found them. The interesting ones are the items whose labels share almost no words and whose correspondence is established mainly by the fact that their observed answer sets coincide, of which the outcome variable of this study is itself an example: Erbil asks whether a displaced household is considering return and Sulaymaniyah asks the

reason a household would consider returning, and only the shared answer set of yes, no and do not know reveals that the second is in fact the same yes-or-no question. The failures are the forced pairs at the bottom of the ranking, where an Erbil expenditure item is married to a Sulaymaniyah income item because nothing better is left.

Table 8. Pairs proposed by the certified matching, with their similarity and the adjudication. The first three would also be found by comparing labels. The fourth and fifth would not: the outcome variable of this study is recovered mainly because the two items share the same observed answer set. The last three are forced pairings of items that have no counterpart, and the threshold of 0.60 excludes them.

Erbil item	Sulaymaniyah item	s	Correct
Sex of household head	Sex of household head	1.000	yes
Households that came directly to current location	Household came directly to current location	0.974	yes
Displaced household left assets in place of origin	Household left assets in place of origin	0.915	yes
Displaced household considering return to place of origin	Reason household would consider returning to place of origin	0.552	yes
Education level of household head	Highest level of education completed by household head	0.506	yes
Approximate household expenditure on fuel in previous month	Amount household spent on repaying loans in past month	0.552	no
Household has children with friends from other population groups	Household contains smokers	0.318	no
Approximate household expenditure on clothing in previous month	Amount of household income from begging in past month	0.190	no

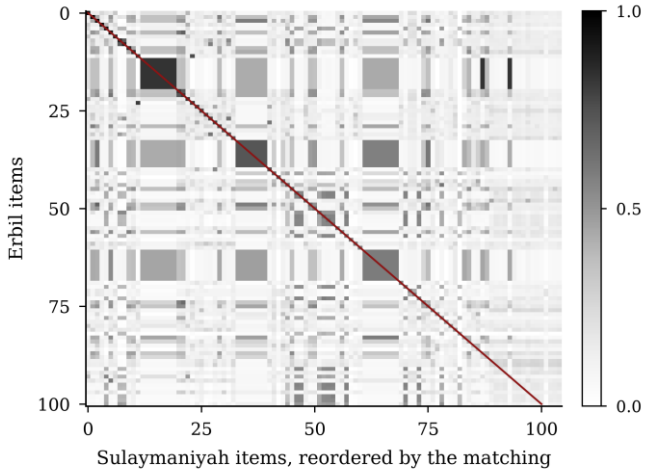

Figure 6. The problem the matching solves. Every cell is the similarity of one Erbil item with one Sulaymaniyah item, and the rows and columns are reordered so that the chosen pairs run down the diagonal. The dark blocks off the diagonal are groups of repeated items, such as the eight document questions, which resemble each other as much as they resemble their counterpart: they are exactly the pairs a rule that took the best match for each item in turn would get wrong.

The matching is optimal and not merely good, and Table 9 is the certificate. It reports the three checks of Equation 10 and Equation 12 with the tolerance each is held to, and the build refuses to continue if any of them fails.

Table 9. The optimality certificate, with the tolerance each check is held to. The matching weighs 63.493431, the relaxation of Equation 10 is worth 63.493431, and the dual prices of Equation 11 sum to 63.493431: the three coincide, which is the strong duality of Equation 13. All four measurements are at the level of floating-point noise, and the build stops if any of them exceeds its tolerance.

Check	Measured	Tolerance	Holds
Matching weight minus LP optimum	0.0e+00	1e-07	yes
Smallest dual slack over all pairs	-5.6e-17	1e-07	yes
Complementary slackness on M	1.1e-16	1e-07	yes
Dual objective minus LP optimum	3.6e-14	1e-07	yes

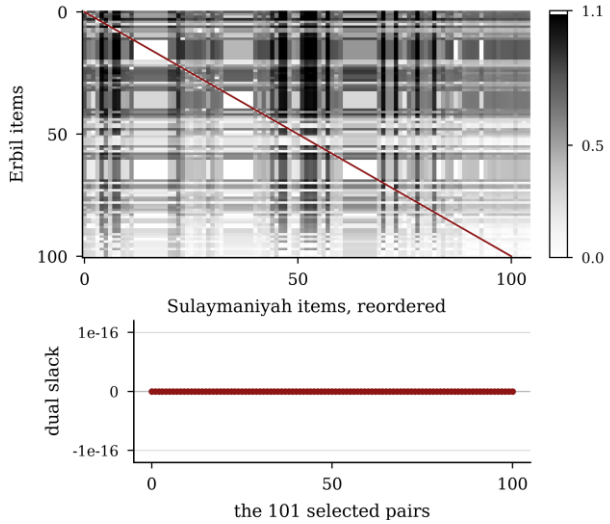

Figure 7. The certificate, drawn. Upper panel: the dual slack of every possible pair, which dual feasibility requires to be non-negative. Lower panel: the same quantity on the 101 pairs the matching chose, where complementary slackness requires it to be zero. It is zero to the last bit in every one of them, not merely within the tolerance the build enforces.

After alignment at the certified-precision threshold, 57 items are available as predictors, 12 of them numeric and 45 categorical. Table 10 traces the households from the admitted files to the analysis sample of 1,437, of which 232 (16.1 per cent) said they would not consider returning.

The two sites are close on this, at prevalences of 0.166 and 0.157. The table also shows how much of each file never reaches a model: 882 households of 2,364 have no answer recorded, most of them because they are not displaced and the question is not put to them.

Table 10. From the households in each admitted file to the households the models are fitted on. Rows two and three partition the file by stratum and rows four to six partition it by what the household answered, so each group sums to the first row. The question is asked only of the displaced strata, which is why most of the missing answers are host or local households. The table also exposes an inconsistency in the published data: 2 Sulaymaniyah households are recorded in the local stratum and still answer the question. They are kept, because an answer is recorded, and at this size they cannot move an estimate.

Households	Erbil	Sulaymaniyah	Both
In the admitted file	1,163	1,201	2,364
Host or local stratum	390	399	789
Displaced stratum	773	802	1,575
No answer recorded	437	445	882
Answered do not know	22	23	45
Analysis sample	704	733	1,437
Would consider return	587	618	1,205
Would not consider return	117	115	232

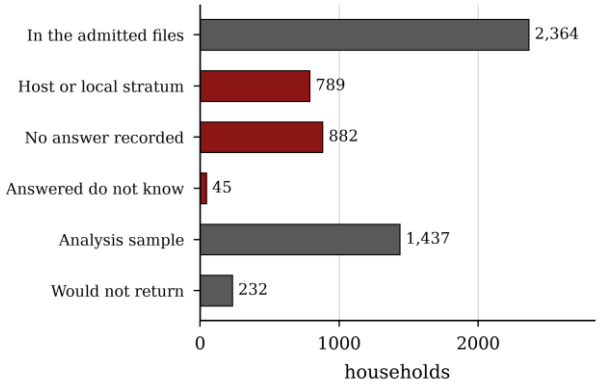

Figure 8. The same funnel drawn to scale. Of 2,364 households in the two admitted files, 1,437 reach a model: the bars in the accent colour are what is removed at each step. Most of the loss is the host and local strata, who are not asked whether they would return.

Within a governorate the answer is weakly but reliably predictable. The sparse logistic model reaches an average precision of 0.302 in Erbil, with a bootstrap interval of 0.242 to 0.385 against a prevalence of 0.166, and 0.229 (0.179 to 0.305) in Sulaymaniyah. Both lie far outside their permutation nulls, at p = 0.0010 and p = 0.0010, and both beat the best single item chosen with hindsight (0.219 and 0.192). Gradient boosting does no better than the linear model, which is what one expects at this sample size. These are modest effects and are reported as such: a screening instrument built on them would still be wrong most of the time.

Across governorates the same models collapse. Fitted on Erbil and evaluated on Sulaymaniyah, average precision falls to 0.140 (0.112 to 0.180), which is the prevalence; the permutation test cannot reject the null at p = 0.9670. In the other direction the figure is 0.162 at p = 0.7562. The receiver operating characteristic is more pointed still: 0.436 with an interval of 0.384 to 0.490, which excludes one half from below, as does the interval in the other direction, 0.381 to 0.495. For the sparse model the transported classifier is therefore not merely uninformative but slightly inverted. The claim is model-dependent: gradient boosting transported from Sulaymaniyah to Erbil reaches 0.501, an interval of 0.444 to 0.555 that contains one half.

What both models agree on is the absence of useful transfer. Figure 9 shows the four estimates against the prevalence.

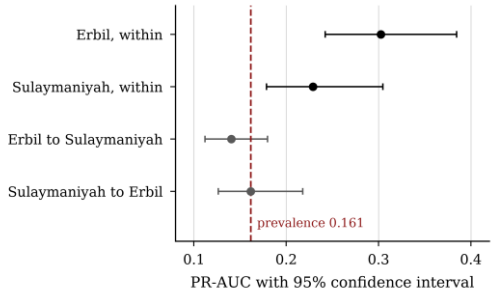

Figure 9. Average precision of the sparse logistic model within each governorate and transported between them, with bootstrap intervals from 2,000 resamples. The dashed line is the prevalence of non-return in the pooled sample, 0.161, which is what a model that has learned nothing attains.

Two of those four estimates are far outside what chance produces and two are inside it. Figure 10 shows the nulls themselves rather than summarising them as a p-value, Figure 11 shows the resampling behind the intervals, and Figure 12 shows what the models actually do to the households they score.

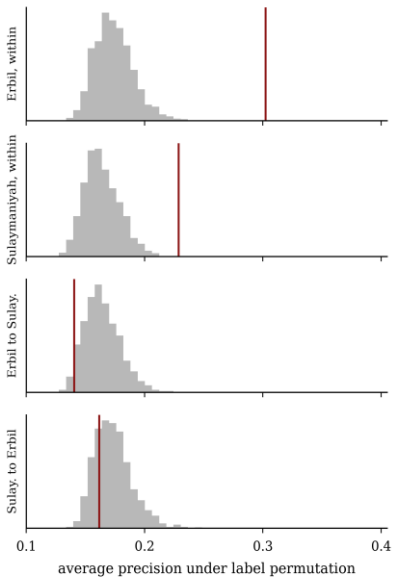

Figure 10. The null model, drawn. Each panel is the distribution of average precision over 1,000 permutations of the labels, with the observed value in the accent colour. Within a governorate the observed value is off the right of the null. Transported, it sits inside it, and in the Erbil-to-Sulaymaniyah panel it sits on the left tail, which is what a p-value of 0.9670 means.

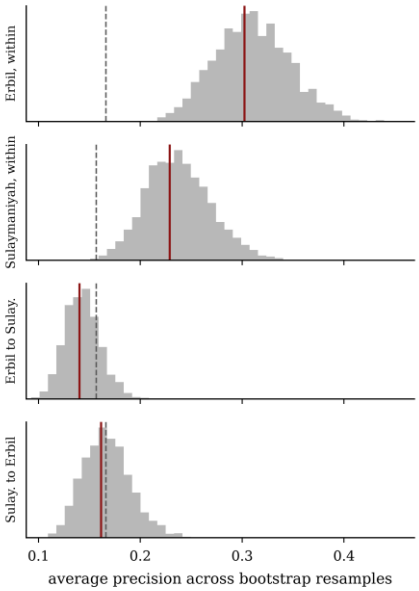

Figure 11. Where the intervals come from. Each panel is the distribution of average precision over 2,000 bootstrap resamples of the evaluation set; the accent line is the value on the sample itself and the dashed line is the prevalence. The two transported distributions straddle the prevalence; the two within-site ones do not reach it.

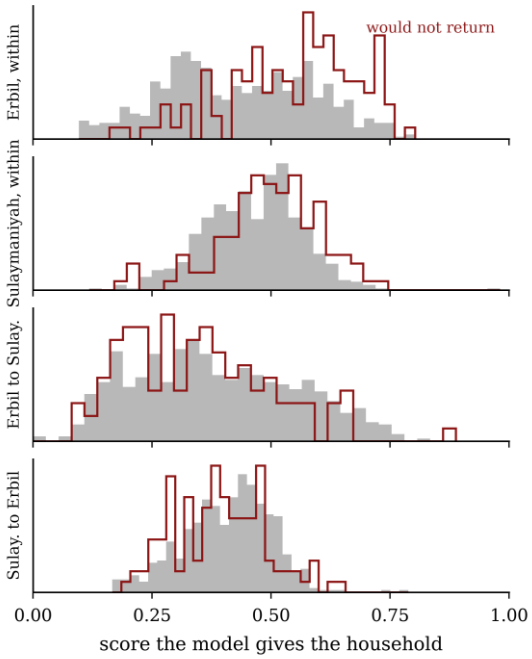

Figure 12. What the models do to the households. In each panel the filled shape is the score given to households that would consider return and the outlined one the score given to those that would not. Within a governorate the outlined distribution sits to the right of the filled one. Transported, the two lie on top of each other, which is the same fact the curves report in another currency.

Table 11 collects the four fits with their intervals and their nulls, against three deliberately unintelligent baselines.

Table 11. Prediction of stated non-return with the sparse logistic model, within each governorate and transported between them, against three deliberately unintelligent baselines. Both columns are ranking metrics and need no operating threshold. Intervals are the percentile bootstrap of Equation 29 over 2,000 resamples, and p is the permutation test of Equation 30 over 1,000 relabellings, which is why 0.0010 is the smallest value it can take. The prevalence of non-return is 0.161, which is what a model that has learned nothing attains; the two transported models do not beat it.

Fit and evaluation	PR-AUC	95% interval	ROC	p
Erbil, cross-validated	0.302	0.242-0.385	0.707	0.0010
Sulaymaniyah, cross-validated	0.229	0.179-0.305	0.604	0.0010
Erbil to Sulaymaniyah	0.140	0.112-0.180	0.436	0.9670
Sulaymaniyah to Erbil	0.162	0.126-0.218	0.438	0.7562
Erbil, majority rule	0.166			
Erbil, random score	0.173			
Erbil, best single item	0.219			
Sulaymaniyah, majority rule	0.157			
Sulaymaniyah, random score	0.169			
Sulaymaniyah, best single item	0.192			

Table 12 gives the two metrics that need an operating threshold, at the cut of Equation 25. They say the same thing in a different currency, and the transported rows say it more bluntly: a negative correlation between prediction and truth.

Table 12. The two metrics that do need an operating threshold, both measured at the cut of Equation 25, which makes the model call as many households positive as the data contain. They are the Matthews correlation coefficient of Equation 26 and the balanced accuracy of Equation 27. An uninformative rule scores zero and one half respectively. Within a governorate both sit above those floors and both are small; transported, the correlation turns negative, which is the same inversion the areas under the curve show.

Fit and evaluation	MCC	Balanced accuracy
Erbil, cross-validated	0.190	0.595
Sulaymaniyah, cross-validated	0.123	0.562
Erbil to Sulaymaniyah	-0.083	0.459
Sulaymaniyah to Erbil	-0.066	0.467

The headline metric is the area under a curve, and Figure 13 is the curve. Figure 14 is the receiver operating characteristic of the same four fits, where the claim about inversion is easiest to check: a curve that sits below the diagonal ranks households backwards.

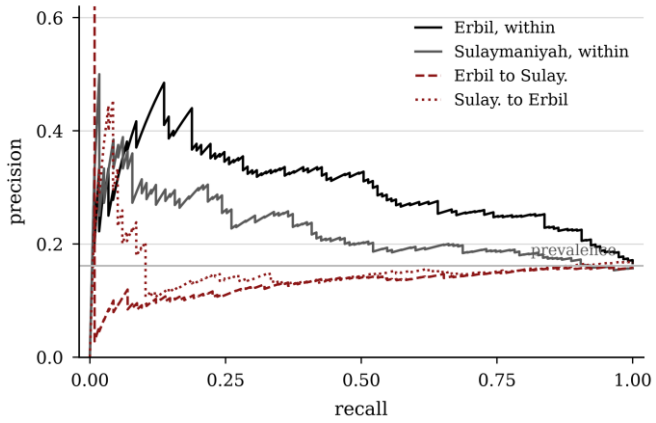

Figure 13. Precision against recall for the four fits. The two cross-validated curves rise above the prevalence at low recall, which is what makes them usable for ranking at all. The two transported curves sit on the prevalence across the whole range: picking the households the model scores highest is no better than picking at random.

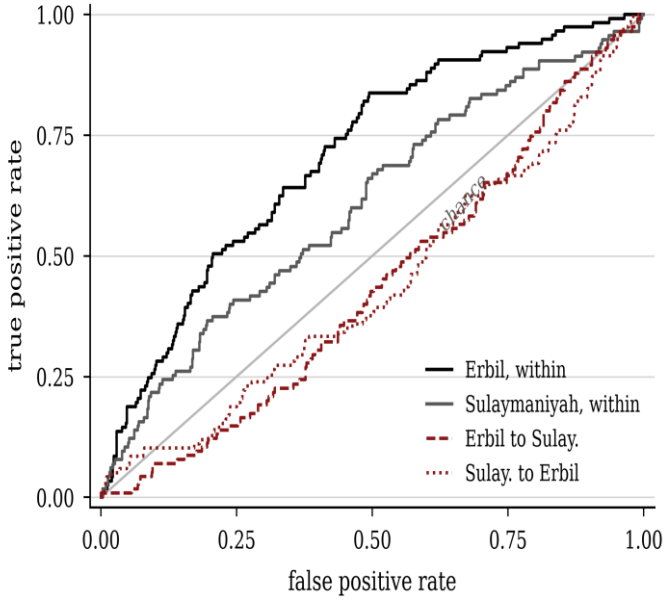

Figure 14. The same four fits as receiver operating characteristics. The two within-site curves are above the diagonal along their whole length. The two transported curves are below it along almost all of theirs, which is the inversion the intervals in Table 11 report as an area under one half.

Table 13 repeats the four fits with the non-additive model, so that the reader can see how much of the result belongs to the choice of model family.

Table 13. The same four fits with histogram gradient boosting, the companion to Table 12. It does no better than the linear model within a site and no better than the prevalence across sites, so the conclusion does not depend on

the model family. One detail does: transported from Sulaymaniyah to Erbil its ROC interval contains one half, where the sparse model's excludes it, so the inversion is a property of the sparse model and not of the data alone.

Fit and evaluation	PR-AUC	95% interval	ROC	ROC interval
Erbil, cross-validated	0.254	0.206-0.328	0.670	0.618-0.720
Sulaymaniyah, cross-validated	0.227	0.181-0.298	0.619	0.563-0.674
Erbil to Sulaymaniyah	0.136	0.110-0.181	0.415	0.360-0.473
Sulaymaniyah to Erbil	0.161	0.131-0.203	0.501	0.444-0.555

The obvious objection is that the transfer fails because the aligned items are not really comparable. Three diagnostics say otherwise.
Predicting the site rather than the outcome from the same 57 predictors reaches an area under the curve of 1.000, so the two samples are separable; but the median item-level Jensen-Shannon divergence of Equation 32 has a median of 0.041 over all 57 of them, and on average 7.1 per cent of the response mass in the evaluation site is unseen during training, measured over the 45 categorical items where that quantity is defined.
No single item is a site fingerprint; the separation is built from many small margins.

Table 14 lists the items that differ most, and separates two faults that are easy to confuse: an item can diverge because the two instruments offer different answer options, which is coding, or because the same options are chosen in different proportions, which is population. Most decisively, discarding every item whose divergence exceeds 0.05 leaves 32 predictors, holds the within-site performance at 0.278 and 0.252, and leaves the transfer exactly where it was, at 0.132 and 0.172, while the site is still recovered at 0.983. Figure 15 shows the sweep.

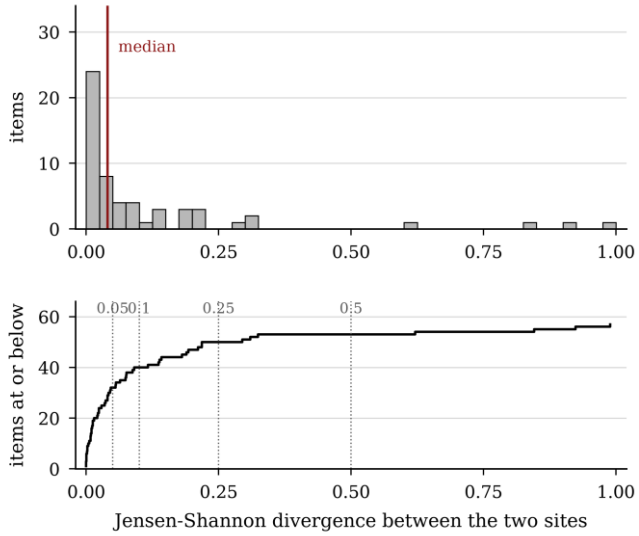

Figure 15. How the divergence is spread over the 57 predictors. Upper panel: the distribution, with the median in the accent colour. Lower panel: the same as a cumulative count, with the bounds of the filter sweep marked. 32 of 57 items sit below the strictest bound, so the separation between the two sites is not carried by a handful of badly coded questions.

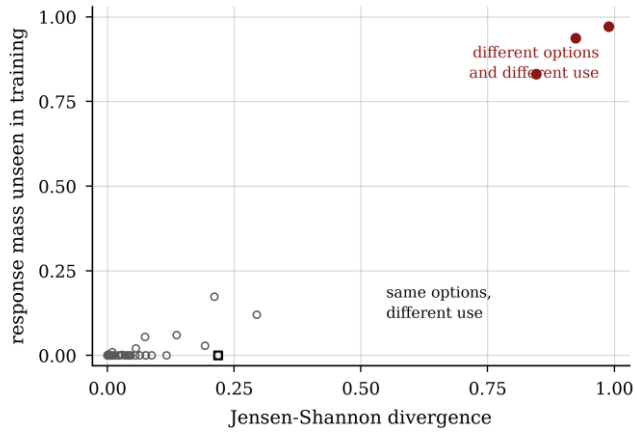

Figure 16. Two different faults, told apart. The horizontal axis is how much the two response distributions differ; the vertical axis is how much of the evaluation site's response mass the training site never saw. The three items at the top right differ because the two instruments offer different answer options. The item marked with a square differs with the same options in both, which is a statement about the households and not about the paperwork.

Table 14. The items whose response distributions differ most between the two governorates. JS is the divergence of Equation 32, Unseen is the share of Sulaymaniyah responses never seen in Erbil, and Levels is how many distinct responses each instrument records. The two columns separate two different faults.

Item	JS	Unseen	Levels
Type of dwelling	0.989	0.971	9 / 3
Access to health care	0.924	0.937	2 / 2
Monthly rent	0.846	0.831	11 / 9
Secondary heating fuel	0.295	0.120	6 / 6
Written tenure document	0.218	0.000	2 / 2
Spending on healthcare medicines treatment	0.211	0.173	10 / 7
Member absent from location (2)	0.193	0.029	12 / 9
Member absent from location (5)	0.137	0.060	2 / 3
Member absent from location (4)	0.117	0.000	5 / 3
Spending on house shelter repairs	0.087	0.000	3 / 3

The first three rows are coding: the Erbil instrument offers nine kinds of dwelling and the Sulaymaniyah one three, so almost no response carries over. Written tenure document is the other kind, the same two levels in both and a genuinely different distribution. Only 32 of 57 items survive the strictest filter, and the transfer does not improve.

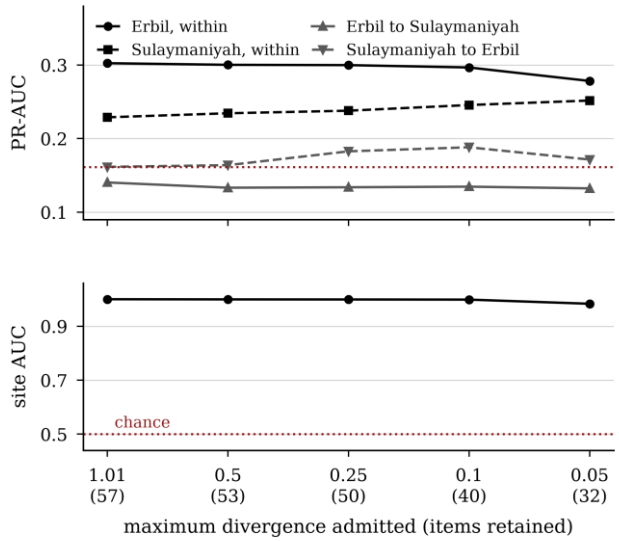

Figure 17. Robustness of the transfer failure to removing items whose response distributions differ between sites. Upper panel: average precision within each site and transported between them, against the prevalence. Lower panel: how well the site itself can still be predicted from the surviving items. Removing the incompatible items does not recover the transfer.

Table 15. The transfer experiment repeated after discarding every item whose divergence exceeds the bound in the first column, reported as average precision. This is Figure 17 in numbers. Dropping the incompatible items costs almost nothing within a site and buys nothing across sites: the two transported columns stay at the prevalence of 0.161 throughout, while the site itself remains recoverable at an area under the curve of 0.983 even from the 32 items that survive.

Max JS	Items	Erbil CV	Sulay. CV	E to S	S to E	Site AUC
1.01	57	0.302	0.229	0.140	0.162	1.000
0.50	53	0.300	0.234	0.133	0.164	0.999
0.25	50	0.300	0.238	0.134	0.183	0.999
0.10	40	0.297	0.246	0.135	0.188	0.999
0.05	32	0.278	0.252	0.132	0.172	0.983

What does change between sites is which circumstances matter. The two encoders agree on 131 terms, and the set each model actually retains is Equation 34. Fitted separately, the sparse model keeps 24 terms in Erbil and 30 in Sulaymaniyah, of which only 11 are kept in both, a Jaccard overlap of 0.256. Of those 11 shared terms, 4 point the same way and 7 point in opposite directions. The correlation of Equation 35, taken over all 131 common terms and not only over the 11 retained ones, is -0.072. Figure 18 summarises this and Table 16 names the shared terms one by one. A model carrying associations of the wrong sign into the other governorate is exactly what produces an area under the curve below one half.

$$S_X = \{f \in F : \beta_f^X \neq 0\}, \qquad J(S_A, S_B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \quad (34)$$

$$\rho = \mathrm{corr}\{(\beta_f^A, \beta_f^B) : f \in F\}, \qquad F = F_A \cap F_B \quad (35)$$

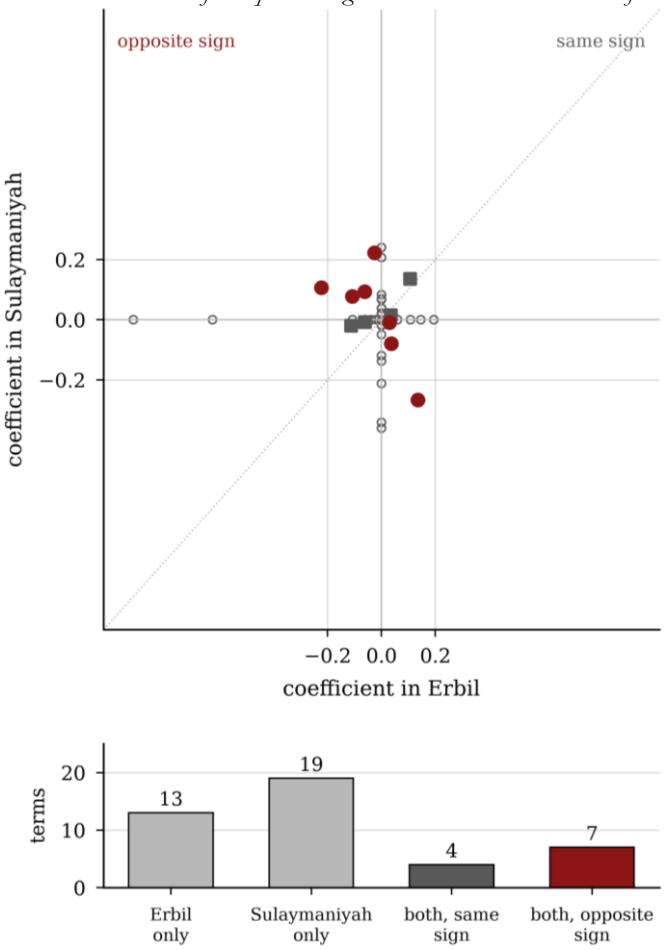

Figure 18. The coefficients of the two governorates against each other. Upper panel: the 131 terms common to both encoders, with the ones retained in both filled; the accent colour marks the pairs whose signs disagree, which fall in the two quadrants off the dotted line. Most terms sit on an axis, retained in one site and dropped in the other. Lower panel: the same four counts. Of the 11 retained in both, 7 carry opposite signs, which is why the transported model is inverted rather than merely uninformative.

Table 16. The 11 terms the sparse model retains in both governorates, with the coefficient it receives in each. This is the mechanism behind the failed transfer, item by item. Of them, 4 keep their sign. Those are the household-size terms and spending on utilities and on other needs. What flips are circumstances a reader would expect to be stable: holding documents, the year of arrival and the age of the head all carry the opposite association in the other governorate. A model that transports these terms does not merely fail to rank; it ranks backwards.

Term retained in both	Erbil	Sulaymaniyah	Sign
Documents held by head = No	0.134	-0.267	opposite
Year of arrival	-0.224	0.107	opposite
Age group of head = 41 - 50	-0.026	0.223	opposite
Documents held by head = Yes	-0.108	0.078	opposite
Spending on house shelter repairs	-0.063	0.093	opposite
Documents held by head (6) = No	0.036	-0.079	opposite
Evicted in past year = No	0.029	-0.009	opposite
Household size, grouped = 2 to 4 members	0.106	0.136	same
Household size	-0.112	-0.020	same
Spending on electricity utilities	-0.061	-0.008	same
Spending on other needs	0.035	0.015	same

A sceptical reader will ask whether the inversion is a property of the households or of the penalty that produced the coefficients.

Figure 19 answers without a model in the way.

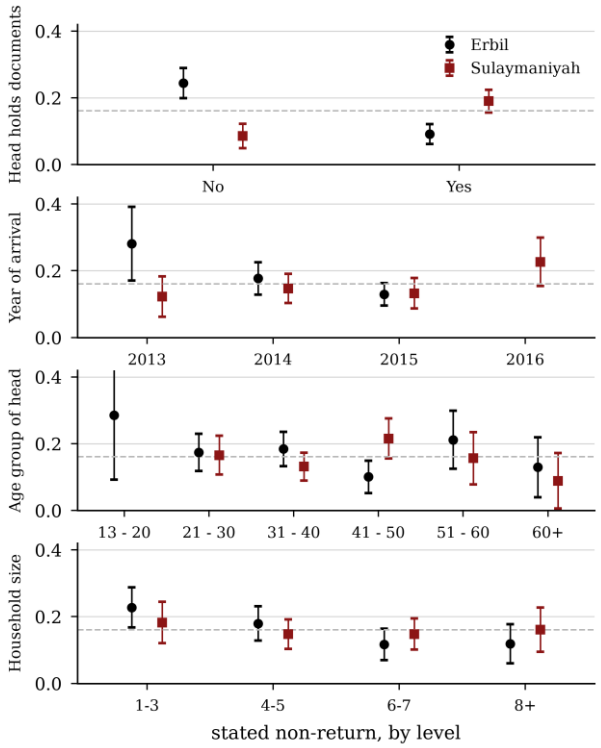

Figure 19. The inversion in the raw proportions. For four of the terms the model retains in both governorates, the share of households saying they would not return, by level, with ninety-five per cent intervals; the dashed line is the pooled prevalence. In the top panel the two lines cross: a head who holds documents is less likely to refuse return in Erbil and more likely in Sulaymaniyah, and neither interval covers the other site's estimate. No regularisation, no cross-validation and no model stand between these proportions and the data.

The alignment threshold is the one free parameter that could plausibly drive the result, since it decides how many items the models see. Table 17 sweeps it from the strictest setting, which admits 23 predictors, to the loosest, which admits 80. Across that range the within-site average precision in Erbil stays between 0.236 and 0.318, always well above the prevalence, while the Erbil-to-Sulaymaniyah transfer stays between 0.134 and 0.155 and the reverse direction between 0.158 and 0.193, never leaving the neighbourhood of the prevalence. Loosening the threshold admits more items, including pairs the adjudication judged wrong, and still does not make the transfer work. The conclusion does not depend on where the cut is placed.

Table 17. Sensitivity of every headline figure to the alignment threshold, reported as average precision. The first two numeric columns are cross-validated within a governorate and the last two are transported between them. Across the whole range the within-site estimates stay well above the prevalence of 0.161 and the transported estimates stay at it.

Threshold	Items	Erbil CV	Sulay. CV	E to S	S to E
0.35	80	0.318	0.246	0.155	0.166
0.45	76	0.318	0.246	0.155	0.165
0.55	70	0.311	0.226	0.150	0.158
0.60	57	0.302	0.229	0.140	0.162
0.70	40	0.291	0.242	0.134	0.183
0.80	23	0.236	0.246	0.136	0.193

Three choices remain that the reader has to take on trust unless they are measured: whether this design could have seen a transfer had one existed, whether the alignment depends on the four weights it was given, and whether the clean adjudication depends on the single reader who made it. Each is measured in turn. A null result is only informative if the design could have found the effect it failed to find.

The transfer experiment was therefore rerun on the real predictors with a shared signal planted in them, at eight strengths with 120 replications each, holding the prevalence and the two sample sizes at their observed values. The signal is planted with the same coefficients in both governorates, which is the most favourable case there is for transfer. Power is the quantity of Equation 36 and the minimum detectable effect is the smallest signal that reaches 80 per cent of it.

$$\Pi(\delta) = \Pr(p < \alpha \mid \delta), \qquad \delta^{*} = \min\{\delta : \Pi(\delta) \geq 0.8\} \quad (36)$$

With no signal planted the test fires 0.033 of the time, which is the level it is set to and confirms the experiment is calibrated. Power reaches 80 per cent against a signal strong enough to lift the transported average precision to 0.236, or 1.46 times the prevalence. The two transported estimates actually observed are 0.140 and 0.162, both at the prevalence and far below that bound. The absence reported here is therefore an absence of any transfer this design would have been able to see, not an absence of data. Table 18 and Figure 20 give the whole curve.

Table 18. What this design could have seen. A shared signal of the given strength is planted in the real predictors, with the same coefficients in both governorates, and the whole transfer experiment is rerun 120 times at each strength. Power is the share of those runs in which the permutation test of Equation 30 rejects at 0.05. The first row is the control: with no signal at all the test fires 0.033 of the time, which is the level it is set to. Reading across, the design reaches 80 per cent power against any signal strong enough to lift the transported average precision to 0.236.

Signal strength	Transported PR-AUC	Power
0.00	0.169	0.033
0.15	0.170	0.042
0.30	0.178	0.167
0.45	0.194	0.383
0.60	0.207	0.617
0.80	0.240	0.825
1.00	0.268	0.900
1.30	0.321	0.950

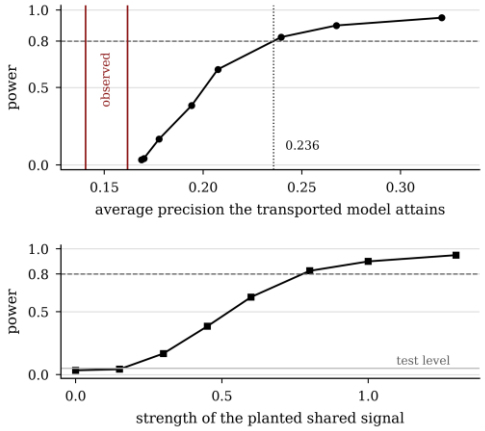

Figure 20. What a transfer would have had to be worth for this design to detect it. Upper panel: power against the average precision the transported model attains, with the 80 per cent line and the two observed transfers in the accent colour. Lower panel: the same against the strength of the planted signal, with the level of the test marked. At zero signal the curve sits on that level, which is the control.

The four weights of Equation 1 are the one arbitrary choice the paper had not swept. Table 19 and Figure 21 report 800 alternative weight vectors, half drawn tightly around the chosen ones and half drawn uniformly over the simplex, which is the difference between asking what happens if the weights move a little and asking what happens if somebody else picks them. Moving them a little changes nothing: precision stays at or above 0.95 in 99.8 per cent of those draws and the matching keeps 99.4 of its 101 pairs.

Picking them arbitrarily does cost something, and the paper should say so: precision at the fixed cut holds in only 55.8 per cent of draws.

But the loss is in the cut, not in the matching. Applying the selection rule of Equation 19 rather than the constant it produced, an arbitrary reweighting still recovers 55.4 genuine correspondences against 73 for the chosen weights, and 6 for identical labels.

Table 19. The alignment under 800 other weight vectors, half drawn tightly around the chosen ones and half drawn uniformly over the whole simplex. Two readings are given because they answer different questions. At the

fixed cut of 0.60 the comparison is unfair to the alternatives: changing the weights changes the scale of the similarity, so the same number no longer means the same thing. Applying the paper's own rule instead, the loosest cut at which no accepted pair is wrong, the chosen weights recover 73 pairs and an arbitrary reweighting still recovers 55.4 on average, against 6 for identical labels.

	Around the chosen weights	Anywhere on the simplex
Precision at 0.95 or above	99.8%	55.8%
Mean precision at the fixed cut	0.994	0.938
Pairs recovered by the rule, mean	71.6	55.4
Pairs recovered, lower fifth of draws	63	20
Pairs identical to the chosen matching	99.4	91.3

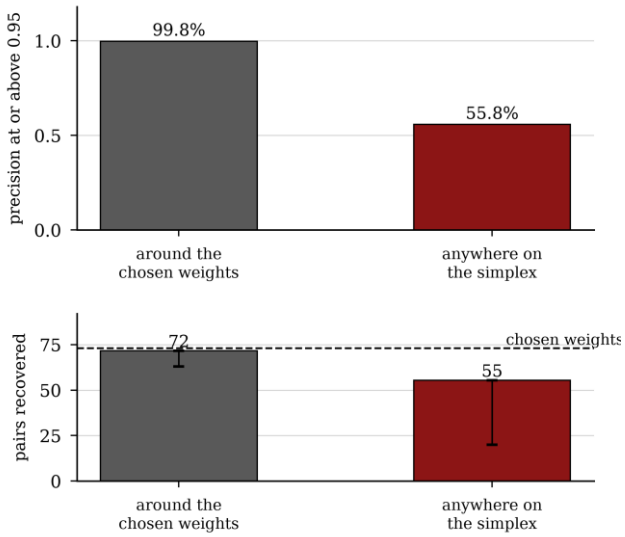

Figure 21. The alignment under other weights. Upper panel: how often precision at the fixed cut of 0.60 stays at or above 0.95. Lower panel: how many genuine correspondences survive when the paper's own selection rule is applied instead of the constant, with the lower fifth of draws marked and the chosen weights as the dashed line. The conclusion the study rests on is the second panel: the matching is robust, the particular threshold is not.

The adjudication was made by one reader and the study leans on it, so Table 20 measures how far it can be wrong. The judgements are overturned adversarially, worst first, rather than at random.

A precision of 1.000 falls to 0.983 on the first overturned judgement, as any perfect score does; it takes 4 to bring it below 0.95, and those 4 pairs all sit within 0.015 of the threshold, which is exactly the band where the Limitations predicted a second reader would disagree.

Overturning the same number at random leaves precision at 0.950. The claim that survives unconditionally is the weaker one: the accepted set is clean unless the reader erred specifically on the pairs closest to the cut.

Table 20. How much the clean result depends on the reader. The adjudication is attacked deliberately rather than perturbed at random: at each step the accepted pair with the lowest similarity is declared wrong, which is both where a second reader would most plausibly disagree and where the damage is greatest. A precision of 1.000 is broken by a single error, as any perfect score is. It takes 4 errors, chosen adversarially, to bring it below 0.95, and all 4 lie within 0.015 of the threshold. Drawing the same number of errors at random instead leaves precision at 0.950 on average.

Judgements overturned	Precision	Similarity of the pair
1	0.983	0.601
2	0.967	0.608
3	0.950	0.612
4	0.933	0.615
5	0.917	0.619

The whole pipeline runs in 113.4 seconds across 7 instrumented stages on the hardware declared in Table 21, with a peak measured allocation well under a tenth of available memory. The certified alignment, the part of the work with a formal guarantee attached, is also the cheapest.

Table 21. Wall-clock time and peak allocated memory of each instrumented stage, measured on an AMD Ryzen 7 9800X3D with 16 GB of DDR5-5200 memory under Windows, with 2,000 bootstrap resamples and 1,000 permutations per comparison. The certified alignment is the cheapest stage in the pipeline.

Stage	Seconds	Peak MiB
Load and align	0.16	47.4
Build design matrix	0.21	2.9
Models and transfer	85.07	11.1
Covariate-shift diagnostics	2.37	6.6
Divergence-filtered transfer	18.50	10.9
Per-site coefficients	0.14	3.1
Threshold sensitivity	6.98	6.5
Total	113.4	

4. Discussion

Three findings follow from this, in ascending order of consequence for how the Region's displacement is described.

The first is about the evidence base itself. A file that has sat on a public portal since 2017 under the name of one governorate turns out to consist, in 84.6 per cent of its rows, of another governorate's sample, and its own remainder does not group into households. Nobody noticed, and the reason nobody noticed is that nobody looked: the reports were written from the survey firm's own tables, and the published microdata appear never to have had a second reader. This is an argument for the cheap, unglamorous machinery the pipeline puts around the data, not for distrust of the exercises, which were carefully done. A gate that assigns rows to places on the evidence of the rows rather than the filename costs a dozen lines and would have caught it on the day of publication.

The second is about comparability. A pair of questionnaires designed half a year apart by overlapping teams for the same purpose in adjacent governorates share 6 identical item labels out of roughly a hundred. Anyone comparing the two governorates is therefore performing an alignment, whether or not they say so, and the alignment they perform is invisible, unrecorded and unevaluated. Computing it instead costs milliseconds, comes with a certificate that no better alignment exists under the stated similarity, and can be checked against an adjudication that anyone may disagree with in public.

The general lesson is not that bipartite matching is clever, which it is not; it is that the step exists and ought to be an artefact rather than an assumption.

The third is the substantive one. Among displaced households in the Kurdistan Region in 2016, the decision not to contemplate return was related to circumstances in ways that a model can pick up inside a governorate and cannot carry across one. With the sparse model the transported classifier is not merely uninformative but slightly inverted, because 7 of the 11 associations it carries have the opposite sign in the other site. Erbil and Sulaymaniyah were not two samples of one displacement situation, and the profiling data record the difference clearly enough that the site can be recovered from a household's circumstances almost perfectly. Why the two differ is not something these data can settle. Origins of displacement, labour markets, distance to the areas under Islamic State control and governorate politics all differed between the two, and any of them could be responsible; none is measured here, and the candidates are offered as hypotheses for work that can test them.

A fourth observation follows from the three taken together. The exclusion of the Duhok file is not only the loss of a third site; it is also what kept the alignment problem in its polynomial regime. Had all three files been usable, the exact certificate of Equation 12 would not have been available, and the study would have had to rely on the weaker bound from the three pairwise optima. That a data-quality defect decided a question of computational complexity is an accident, but it is the kind of accident that recurs: the shape of the evidence determines which guarantees are affordable, and a pipeline that does not audit the evidence first cannot know which regime it is in.

The practical implication runs against a common habit. Findings from a profiling exercise in one governorate of the Region should not be extended to another without a test, and the test is cheap: fit where the data are, predict where they are not, and compare against the prevalence and a permutation null. Where the transfer fails, as it does here, an instrument tuned in one governorate will misrank households in the next, and a targeting rule built on it will send assistance to the wrong doors in a way that looks principled.

It is worth being precise about what the site classifier does and does not show. That the governorate can be recovered from a household's circumstances at an area under the curve of 1.000 is not in itself surprising: two purposive samples drawn six months apart in different cities would differ on many margins even if the underlying process were identical. What matters is that removing the items responsible for the largest distributional differences leaves both the separation and the transfer failure intact. A shift that could be cured by dropping a handful of badly coded items would be a paperwork problem; one that survives dropping almost half of them is a statement about the populations.

The engineering that produced these findings is deliberately unremarkable in its parts. A digest recorded before a file is read, a gate that assigns rows to places on their own evidence, a matching solved exactly with its certificate checked, a null model

attached to every comparison, and a document whose every number is interpolated from a machine-written artefact are all standard practice somewhere. What is unusual is applying them together to a humanitarian dataset, where the incentives run towards producing a report quickly and the cost of a silent error is borne by people who will never read it. The provenance defect reported here had been public since the files were released, and cost twelve lines of code to find.

None of this says that the household's situation is irrelevant to the decision to return. It says that the mapping from situation to decision is local. Qualitative accounts of displacement have long described return decisions as situated in a particular place and moment rather than as a general function of need; the contribution here is a number and an interval, in a setting where the data to produce them had been published and left unexamined.

5. Limitations

The adjudication of the 101 proposed pairs was done by a single reader, the author, and a second reader would disagree on some of them. The disagreements would fall almost entirely among the middling similarities, where a pair joins two items that ask about the same construct over different reference periods; the pairs that carry the analysis, above 0.60, are not of that kind. The adjudication is published so that the disagreement can be specific.

The similarity of Equation 1 and its four weights are a modelling choice. They were fixed before the alignment was inspected, and the threshold sweep shows the conclusions do not turn on where the cut is placed, but a different similarity would propose a different matching, and the certificate guarantees optimality only with respect to the similarity given to it. A certificate of optimality is not a certificate of correctness.

The outcome is a stated intention recorded at one moment, not an observed return. Intentions are known to be unstable and to respond to information about conditions at origin that the survey does not capture, and a household that says it will not return may be reporting an assessment of the security situation rather than a plan. The finding is therefore about the predictability of the statement, which is what a profiling exercise can actually collect.

The two sites were surveyed at different moments, Erbil at the turn of 2015 into 2016 and Sulaymaniyah in June 2016, so the difference between them mixes place with time. Six months is short against the displacement itself but not against the military situation, and with two sites there is no way to separate the two sources of variation.

The Duhok exercise would have helped, and could not be used.

The positive class is small in absolute terms: 117 households in Erbil and 115 in Sulaymaniyah. Bootstrap intervals on average precision at that size are wide, and the within-site estimates should be read as evidence that a relationship exists rather than as a usable calibration. The cross-site conclusion is the more secure of the two, because it rests on an absence that is confirmed by the permutation test, by the divergence-filtered replication and by the coefficient comparison, three pieces of evidence that would have to fail together for the conclusion to be wrong.

The sampling weights published with the exercises were not used in the models. They are extrapolation factors designed for population estimates of the kind the profiling reports produce, and applying them inside a cross-validated classifier changes what is being estimated without making the transfer question any clearer. Every figure reported here is therefore a statement about the sampled households, not a population estimate for the governorates, and should not be read as one.

Finally, the transfer test uses only household-level items. The individual-level blocks on employment, education and health were left aside because summarising them to the household introduces choices that would need their own sensitivity analysis. It is possible, though the coefficient comparison makes it unlikely, that a well-chosen individual-level summary transfers where the household-level items do not.

6. Conclusion

The Kurdistan Region's own picture of out-of-camp displacement is three files on a public portal. One of them cannot be used as labelled, and a dozen lines of provenance checking establish that before any analysis begins. The remaining pair can be compared only through an alignment that must be computed: identical labels recover 6 of 84 genuine correspondences, whereas a maximum-weight bipartite matching recovers 60 of them without a single false pair, and all 84 at a precision of 0.966, with a certificate that no better alignment exists under the similarity used.

On the aligned core, a displaced household's stated refusal to contemplate return is weakly predictable inside a governorate and no better than chance across governorates, and the failure survives removing every item whose response distribution differs between sites. The reason is visible in the coefficients: of the 11 associations retained in both governorates, only 4 point the same way. Generalising a profiling finding from one governorate of the Region to another is an empirical claim, it is testable with the data already published, and in this instance it does not hold.

Declarations

Data availability. The three source files are published by the Joint IDP Profiling Service on the Humanitarian Data Exchange under a Creative Commons Attribution licence.

In total three files are used, 8,493,482 bytes in all, each recorded with its retrieval address, licence and SHA-256 digest before it was read.

Code availability. The whole pipeline accompanies this manuscript as a supplementary archive: the analysis code, the automated tests, the quality gates, the manual adjudication of the alignment and the builder that produced this document, together with the intermediate JSON artefacts that every number printed here is interpolated from.

The archive carries no survey data. The three source files are downloaded from the addresses recorded above by the first stage of the pipeline and checked against their digests before anything reads them, so the run can be repeated from the published sources rather than from a copy of them. On acceptance the archive will be deposited in a public repository under a permanent identifier.

Funding. This research received no external funding.

Conflicts of interest.

The author declares no conflict of interest.

Ethics. The study is a secondary analysis of de-identified, publicly released household microdata and involved no contact with human participants.

Reproducibility. The pipeline runs end to end from the published files with a single command.

Every number in this paper is interpolated from a JSON artefact produced by that run, and the gates of Table 22 run on the assembled document before it is accepted. The random seed is 20191231 and two runs produce identical artefacts.

Table 22. The gates the build runs on the assembled document. Each one fails the build rather than printing a warning, because a warning in a log nobody reads is the same as no check at all. The last one is the reason no number in this paper was typed by a person.

Gate	What it refuses to build
Dates	a document with any date later than the cut-off in its visible text, equations included
Order of appearance	a document that cites a figure or table that does not exist, or numbers them out of order
Abstract	an abstract longer than the journal allows
Abbreviations	an acronym used before the words it stands for
Number format	an ordinal in digits, a thousand without its separator, or a digit and a spelled number counting things in one sentence
Equations	a document whose nth equation is not the one its nth key names
Numbers	a document containing a figure that no artefact under results/ produced

References

1. Abdulkhalk Abass, L. "Assessment of Knowledge and Practices of Internally Displaced Pregnant Women." Kurdistan Journal of Applied Research 3, no. 1 (2018): 55–60. https://doi.org/10.24017/science.2018.1.10.
2. Aziz, I. A., C. V. Hutchinson, and J. Maltby. "Quality of Life of Syrian Refugees Living in Camps in the Kurdistan Region of Iraq." PeerJ 2 (2014): e670. https://doi.org/10.7717/peerj.670.
3. Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. "A Theory of Learning from Different Domains." Machine Learning 79, no. 1–2 (2010): 151–175. https://doi.org/10.1007/s10994-009-5152-4.
4. Birkhoff, G. "Tres observaciones sobre el álgebra lineal." Universidad Nacional de Tucumán Revista Serie A 5 (1946): 147–151.
5. Buck, K. "Displacement and Dispossession: Redefining Forced Displacement and Identifying Protection Gaps." Journal of International Humanitarian Action 2, no. 1 (2017): 3. https://doi.org/10.1186/s41018-016-0016-6.
6. Cetorelli, V., I. Sasson, N. Shabila, and G. Burnham. "Mortality and Kidnapping Estimates for the Yazidi Population in the Area of Mount Sinjar, Iraq, in August 2014." PLOS Medicine 14, no. 5 (2017): e1002297. https://doi.org/10.1371/journal.pmed.1002297.
7. Christen, P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, 2012.
8. Crama, Y., and F. C. R. Spieksma. "Approximation Algorithms for Three-Dimensional Assignment Problems with Triangle Inequalities." European Journal of Operational Research 60, no. 3 (1992): 273–279. https://doi.org/10.1016/0377-2217(92)90078-N.
9. Davis, J., and M. Goadrich. "The Relationship between Precision-Recall and ROC Curves." In Proceedings of the 23rd International Conference on Machine Learning, 233–240. 2006. https://doi.org/10.1145/1143844.1143874.
10. Garey, M. R., and D. S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman, 1979.
11. Joint IDP Profiling Service. Displacement as Challenge and Opportunity. Urban Profile: Refugees, Internally Displaced Persons and Host Community, Erbil Governorate, Kurdistan Region of Iraq. Joint IDP Profiling Service, 2016.
12. Joint IDP Profiling Service. Displacement as Challenge and Opportunity. Urban Profile, Duhok Governorate, Kurdistan Region of Iraq. Joint IDP Profiling Service, 2016.
13. Joint IDP Profiling Service. Displacement as Challenge and Opportunity. Urban Profile, Sulaymaniyah Governorate and Garmian Administration, Kurdistan Region of Iraq. Joint IDP Profiling Service, 2016.
14. Jonker, R., and A. Volgenant. "A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems." Computing 38, no. 4 (1987): 325–340. https://doi.org/10.1007/BF02278710.

15. Karp, R. M. "Reducibility among Combinatorial Problems." In Complexity of Computer Computations, edited by R. E. Miller and J. W. Thatcher, 85–103. Plenum Press, 1972. https://doi.org/10.1007/978-1-4684-2001-2_9.
16. Kevers, R., and P. Rober. "The Role of Collective Identifications in Family Processes of Post-Trauma Recovery." Kurdish Studies 5, no. 2 (2017): 155–176. https://doi.org/10.33182/ks.v5i2.440.
17. Kuhn, H. W. "The Hungarian Method for the Assignment Problem." Naval Research Logistics Quarterly 2, no. 1–2 (1955): 83–97. https://doi.org/10.1002/nav.3800020109.
18. Lin, J. "Divergence Measures Based on the Shannon Entropy." IEEE Transactions on Information Theory 37, no. 1 (1991): 145–151. https://doi.org/10.1109/18.61115.
19. Matthews, B. W. "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme." Biochimica et Biophysica Acta 405, no. 2 (1975): 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.
20. Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. "A Unifying View on Dataset Shift in Classification." Pattern Recognition 45, no. 1 (2012): 521–530. https://doi.org/10.1016/j.patcog.2011.06.019.
21. Munkres, J. "Algorithms for the Assignment and Transportation Problems." Journal of the Society for Industrial and Applied Mathematics 5, no. 1 (1957): 32–38. https://doi.org/10.1137/0105003.
22. Peng, R. D. "Reproducible Research in Computational Science." Science 334, no. 6060 (2011): 1226–1227. https://doi.org/10.1126/science.1213847.
23. Rahm, E., and P. A. Bernstein. "A Survey of Approaches to Automatic Schema Matching." The VLDB Journal 10, no. 4 (2001): 334–350. https://doi.org/10.1007/s007780100057.
24. Saito, T., and M. Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." PLOS ONE 10, no. 3 (2015): e0118432. https://doi.org/10.1371/journal.pone.0118432.
25. Schrijver, A. Combinatorial Optimization: Polyhedra and Efficiency. Springer, 2003.
26. Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society Series B 58, no. 1 (1996): 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
27. Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.