Using Longitudinal Data for Plausible Counterfactual Explanations

Alexander Asemota[1],*, Giles Hooker[2]

[1]University of California Berkeley, Berkeley, California, United States
[2]University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Counterfactual explanations are a common approach to providing recourse to data subjects. However, current methodology can produce counterfactuals that cannot be achieved by the subject, making the use of counterfactuals for recourse difficult to justify in practice. Though there is agreement that plausibility is an important quality when using counterfactuals for algorithmic recourse, ground truth plausibility continues to be difficult to quantify. In this paper, we propose using longitudinal data to assess and improve plausibility in counterfactuals. In particular, we develop a metric that compares longitudinal differences to counterfactual differences, allowing us to evaluate how similar a counterfactual is to prior observed changes. Furthermore, we use this metric to generate plausible counterfactuals. Finally, we discuss some of the inherent difficulties of using counterfactuals for recourse.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

Over the past two decades, machine learning and artificial intelligence have become intertwined with broad swaths of society, from education to criminal justice to consumer finance. Throughout this transition away from human decision-makers and towards algorithmic decision-makers, researchers, practitioners, and advocates have emphasized the need for explainability and transparency. Approaches to explainability have varied widely, from creating novel ’glass-box’ model architectures to developing post-hoc local explainability techniques [1] [2]. Of particular interest in the past five years are *counterfactual explanations*, which explain an individual prediction by finding an, in some sense, small change to achieve a desired prediction[3]. In contrast to most explainability techniques, counterfactual explanations seek to explain algorithmic decisions to data subjects.

Although there has been substantial work in the domain of machine learning explainability, significant gaps exist regarding the utility of explanations to data subjects. Unlike concepts such as accuracy or sparsity, subject utility has neither a simple nor agreed upon mathematical definition. Consequently, counterfactual explanation methods optimize subject utility using disparate approaches [4] [5]. Terms such as plausibility, validity, and actionability are used to describe different aspects of the utility of counterfactuals. Plausibility, the main focus of this

HI-AI@KDD, Human-Interpretable AI Workshop at the KDD 2024, 26th of August 2024, Barcelona, Spain

*Corresponding author.

alexander.asemota@berkeley.edu (A. Asemota); ghooker@upenn.edu (G. Hooker)

paper, requires that a counterfactual is a possible state of being [6]. Nonetheless, generating plausible counterfactuals is not a simple task. Substantial effort has been devoted to developing methods for plausible counterfactuals, but there are no agreed upon approaches or even metrics for plausibility. Additional effort has gone into using counterfactuals to provide recourse to data subjects [7]. Recourse is a stricter goal, requiring that a counterfactual be useful to a data subject in pursuing a desired decision. Therefore, plausibility is necessary for counterfactuals to be used for recourse.

This paper proposes a novel approach to evaluating plausibility using longitudinal data. We begin by briefly reviewing approaches to improving plausibility in counterfactuals, discussing in particular persistent pitfalls. We then introduce a longitudinal distance metric for counterfactual explanations. In introducing our metric, we bring forth the benefits of using longitudinal data as a proxy for plausibility and mention some limitations. Next, we perform experiments with our metric to evaluate the use of longitudinal data for plausibility. We also explore some of the consequences of requiring plausibility. Finally, we discuss the implications of our results in the broader context of providing recourse through counterfactual explanations.

2. Offering Recourse Through Counterfactuals

A common motivation for counterfactual explanations is to provide data subjects with a path to recourse. Counterfactuals are unique in their ability to not only explain algorithmic decisions to a lay audience, but also explain how someone could receive a desired decision. That is, a counterfactual informs a subject not only *why* they received a decision, but *what* to change and *how much* to change. Therefore, counterfactuals have the potential to greatly increase transparency and accountability in algorithmic decision-making.

However, persistent gaps exist between the ideal scenario and counterfactuals in practice. Centrally, current methodology fails to consistently produce plausible or achievable explanations. Here, we use the terms plausible and achievable to refer to *objective* and *subjective* perspectives of the difficulty of pursuing a given counterfactual. A counterfactual is *plausible* if it respects constraints on reality, for example, not changing ethnicity or decreasing age. On the other hand, a counterfactual is *achievable* if the relevant subject can achieve it. It is generally plausible for someone to increase their level of education, but it may not be achievable for a given individual. These definitions themselves elucidate the difficulty of offering recourse through counterfactual explanations; how do we know if a data subject can act on a particular recommendation?

Existing counterfactual explanation methods use proxies for plausibility and achievability in an attempt to avoid implausible recommendations. Two proxies are most common: relying on user constraints and leveraging structure in data[5] [8] [4] [9] [10] [7]. Users (i.e. the person using counterfactuals to explain an algorithm) often have domain expertise on how data subjects can change. However, relying solely on users risks introducing social bias to explanations. Data offer some opportunity to craft constraints objectively, but existing methods to enforce plausibility using data are insufficient. Current data-based plausibility constraints either assume individuals are interchangeable or require causal modeling. The former approach does not reflect the complexities of recommending changes to people, and the latter requires significant (and often unavailable) knowledge on the part of the user.

Ultimately, proxies are needed to produce plausible or achievable counterfactuals at scale. We may not know what an individual can or cannot achieve, or we may have incomplete information of relationships between the features in our model. However, existing methodologies often use proxies that insufficiently penalize implausible explanations.

3. Longitudinal Data as a Proxy for Plausibility

As discussed in 2, counterfactual explanations often are motivated by the goal of offering recourse to data subjects, though there are persistent issues that prevent most methods from providing recourse. If we view counterfactual explanations as potential paths to recourse, then we can conceptualize them as recommendations for algorithmic subjects. Specifically, we can view counterfactuals as recommendations for changes that a data subject can make to receive a desired decision at some point in the future. Conceptualizing counterfactuals as potential states of being forward in time naturally leads to considering longitudinal likelihoods. That is, when making recommendations for the future, we should consider prior observed changes over time. This perspective leads us to the primary goal of this paper: leveraging longitudinal data to assess and improve plausibility in counterfactual explanations.

We introduce a distance metric that compares prior observed changes to proposed changes in the form of counterfactual explanations. Let $A, B \in \mathbb{R}^{n \times d}$ be n observations of d features across two different points in time. Subsequently, let $D = B - A$, that is the change in the observed features over time. Now we define our distance metric

$$L(x, e; D, s) = \min_{|\mathcal{I}|=s} \frac{1}{s} \sum_{i \in \mathcal{I}} \|(e - x) - D_i\| \tag{1}$$

where $\mathcal{I}$ is an index set for prior observed changes, s is the desired size of the index set, x is the example to be explained, and e is the counterfactual explanation. In summary, we compare a proposed difference to the s closest differences and average them. By increasing s, we require that the proposed difference is similar to a larger number of observed differences.

We can justify and augment our approach in the following ways:

- Since there are likely a large variety in observed trajectories, we average the s most similar. This allows room for heterogeneity across trajectories without allowing a single rare trajectory to dominate our metric.
- In the likely chance that our data contains heterogeneous features, we can normalize our distance metric across features. Here, we consider dividing features by a metric for the dispersion of their observed differences. Our experiments use the median absolute deviation (MAD) or average absolute deviation (AAD), but other approaches can be implemented.
- For categorical features, several options can be exercised. For binary features, the average absolute deviation can be appropriate. For multi-class features, we normalize by the rate at which changes are observed in the longitudinal data.
- Normalization can empower discovery of implausible counterfactual explanations. If our feature has a high normalization value (e.g. 1 over the MAD), then changes to that feature are rarely seen.

Our proposed metric is flexible in its use; we can use it both during and after generating counterfactuals. Post-generation, the longitudinal distance metric can be used to evaluate and rank the plausibility of explanations. During generation, the distance metric can be used to further constrain the search space. Notice that in comparing proposed changes to those observed in longitudinal data, we not only re-weight distances on a per-feature basis, but we also incorporate dependencies between feature changes; ruling out a requirement to, for example, both change profession and increase the length of tenure in your current job.

A first approach to incorporating our longitudinal metric is to re-score a proposed collection of counterfactuals. Stochastic search algorithms used in [5] and [9] return a set of possible counterfactual explanations, usually incorporating a geometric distance metric; these can then be examined or prioritized by longitudinal distance. We generate counterfactuals using the methods in [5], and then score them by plausibility. This two-step approach allows us to use the more regular geometric distance for optimization, providing a more efficient search of feature space. In this paper, it also allows us to examine the baseline plausibility of counterfactuals generated with a geometric distance. In the appendix, we offer one simple way to generate counterfactuals using our longitudinal metric.

4. Ranking Explanations from DiCE and MIMIC-III

To ground our approach, we consider an example from healthcare. Suppose a predictive model is being used to assess patient risk for a disease. Counterfactuals may be useful to doctors by explaining individual predictions and showing potential paths to decreased risk. In this experiment, we use MIMIC-III, an electronic health records (EHR) dataset, to predict acute respiratory failure (ARF) within four hours of admission [11] [12]. Specifically, we use a version of MIMIC-III that has been preprocessed by FIDDLE, an EHR data processing pipeline [13] [14]. Our longitudinal data consists of measurements when the patient is admitted and repeat measurements four hours after admission. We train a random forests model to predict ARF using only the first time step of data, and use DiCE ([5]) to generate ten counterfactuals for individuals in the test set who are predicted to have ARF. Our dataset contains 1350 features to train our model, twenty of which are derivations from vital signs.

To rank counterfactuals, we use the longitudinal differences between the first and fourth hour of data. We normalize our metric using the AAD of the longitudinal differences, and we add a small tolerance (10^{-5}) to prevent division by zero. Finally, we conduct this experiment in two different settings: allowing all features to be changed (*ALL*) and allowing only vital signs to be changed (*VITAL*). These two settings should show us how our longitudinal distance metric can help assess plausibility, both when we know what features can be changed and when we may be unsure how to constrain the counterfactual search space. With this experiment, we seek to assess the discriminatory ability of our distance metric and subsequently evaluate plausibility for counterfactual explanations.

4.1. Results

We begin by looking at the relationship between the geometric distance and longitudinal distance. Figure 1 B and C plot the L1 distance compared to the longitudinal distance for

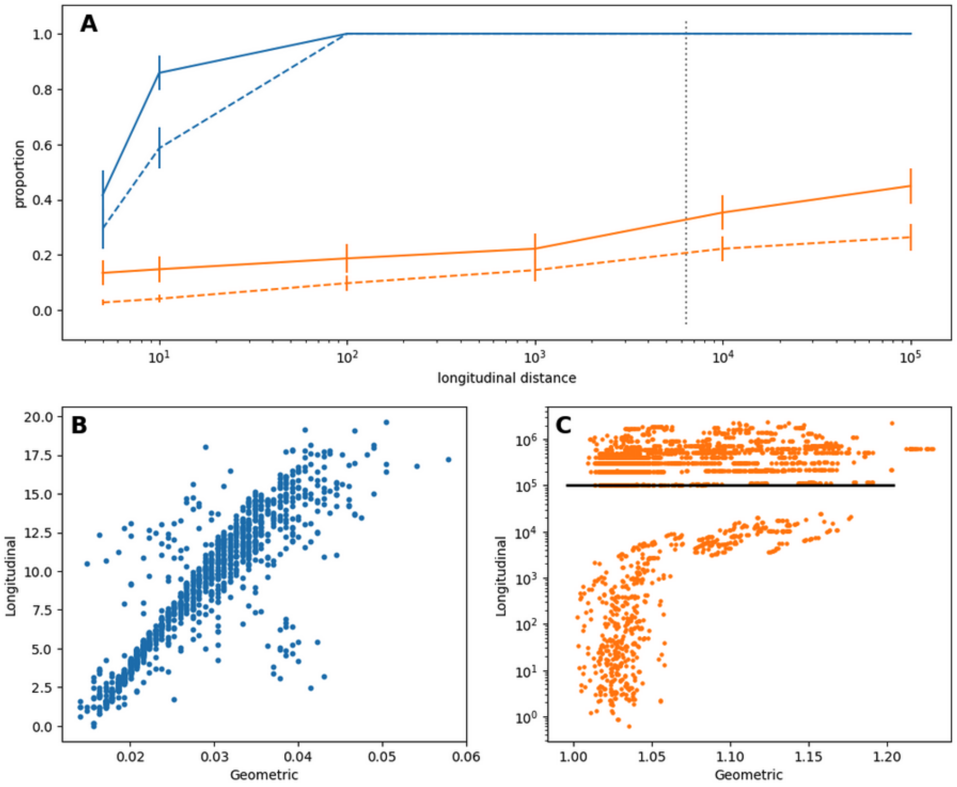

Figure 1: A shows the proportion of individuals with explanations below each threshold of longitudinal distance (x-axis). The solid lines represent the proportion of individuals with at least once explanation below the threshold, and the dotted line is the average proportion of explanations below the threshold. The vertical dotted line shows the distance value we would expect a feature to have if we observed it changing for one individual. **B** and **C** compare the Geometric and Longitudinal distances between an input and a counterfactual. In all plots, blue refers to *VITAL* and orange refers to *ALL*.

explanations using only vital signs and all features respectively. Overall, our metric is loosely correlated with the L1 distance, but there are significant deviations based on which features are changed. In the case of explanations that can only change vital signs, there is a noticeable linear relationship. Vital signs in our dataset generally change at a similar rate, so the cost of changing one or the other is similar. Therefore, the L1 distance is closely related to the longitudinal distance.

Looking at explanations that consider all features, the relationship is much more tenuous. Some features rarely change, leading to significant jumps in our distance metric even when only one feature is changed. Additionally, about half of the explanations change some immutable feature, such as Hospital Ward or Religion. Some explanations also change features that are mutable in theory, but not observed to have changed. Since the AAD is zero for some features, changing those features results in a large distance value. The resulting distance is above 10^5, that is $\frac{1}{AAD+tolerance} = 10^5$. Moreover, changing features without observed changes is not a rare occurrence. When we allowed any feature to be changed, 74 percent of the counterfactuals generated had a longitudinal distance value above 10^5. This is in stark contrast *VITAL*, where there are no explanations with a longitudinal distance value above 20.

Next, we consider plausibility at the individual level. Each individual receives ten counterfactuals, but our metric will help us see how many plausible counterfactuals an individual receives on average. Figure 1 **A** shows the proportion of explanations below a threshold, and the vertical line represents the distance value of a change that has occurred once in our train set. For the purposes of this experiment, we consider counterfactuals below that threshold to be plausible. We can see that not only are explanations less plausible on average in *ALL*, the proportion of individuals with plausible explanations is significantly lower. Consequently, the vast majority of individuals in our test set do not receive a plausible counterfactual if we consider changing all features. Constraining to vital signs, however, leads to only plausible counterfactuals.

Though we have seen that constraints can improve plausibility, it is important to consider the effect constraints have on validity. We consider a counterfactual 'valid' if it has the desired prediction (i.e. not at risk of ARF). Notably, constraints may prevent the generation of valid counterfactuals by not allowing changes to predictive features. While we were able to generate counterfactuals for all 229 individuals in the test set with *ALL*, we only generated counterfactuals for 120 individuals with *VITAL*. This disparity raises concerns around the tension between plausibility and validity; we can improve plausibility by constraining our search space, but we may constrain counterfactuals in a way that degrades validity.

In summary, our experiment elucidates the following:

- Longitudinal data allows us to detect and penalize unobserved changes in counterfactuals
- Constraints can improve plausibility
- Plausibility and validity are in tension with each other

4.1.1. Supplementary Results

In the supplement to this paper, we provide further results, specifically regarding the use of our metric to generate counterfactuals.

5. Future Work

In this paper, we only consider a longitudinal distance metric. Future work should explore modeling longitudinal data to further improve plausibility constraints. Future work should also consider implementing intermediate steps across time, and modeling plausiblity in terms of the subject's current features. Additionally, our approach is computationally complex due to row-wise comparison of matrices. Further work can investigate decreasing this complexity, potentially using prototypes or other clustering methods.

References

[1] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, Association for Computing Machinery,

New York, NY, USA, 2013, p. 623–631. URL: https://doi.org/10.1145/2487575.2487579. doi:10.1145/2487575.2487579.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.

[3] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, CoRR abs/1711.00399 (2017). URL: http://arxiv.org/abs/1711.00399. arXiv:1711.00399.

[4] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. D. Bie, P. Flach, FACE, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2020. URL: https://doi.org/10.1145%2F3375627.3375850. doi:10.1145/3375627.3375850.

[5] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, 2020. URL: https://doi.org/10.1145%2F3351095.3372850. doi:10.1145/3351095.3372850.

[6] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022). URL: http://dx.doi.org/10.1007/s10618-022-00831-6. doi:10.1007/s10618-022-00831-6.

[7] A. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, CoRR abs/2002.06278 (2020). URL: https://arxiv.org/abs/2002.06278. arXiv:2002.06278.

[8] M. Schleich, Z. Geng, Y. Zhang, D. Suciu, Geco: Quality counterfactual explanations in real time, 2021. arXiv:2101.01292.

[9] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: Parallel Problem Solving from Nature – PPSN XVI, Springer International Publishing, 2020, pp. 448–469. URL: https://doi.org/10.1007%2F978-3-030-58112-1_31. doi:10.1007/978-3-030-58112-1_31.

[10] M. Downs, J. Chu, Y. Y., D.-V. F., P. WeiWei, Cruds: Counterfactual recourse using disentangled subspaces, ICML Workshop on Human Interpretability in Machine Learning (2020) 1–23.

[11] A. E. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016). URL: https://doi.org/10.1038/sdata.2016.35. doi:10.1038/sdata.2016.35.

[12] A. E. Johnson, T. J. Pollard, R. G. Mark, Mimic-iii clinical database (version 1.4), 2016. URL: https://physionet.org/content/mimiciii/1.4/. doi:10.13026/C2XW26.

[13] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, J. Wiens, Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data, Journal of the American Medical Informatics Association 27 (2020) 1921–1934. URL: https://doi.org/10.1093/jamia/ocaa139. doi:10.1093/jamia/ocaa139. arXiv:https://academic.oup.com/jamia/article-pdf/27/12/1921/34838612/ocaa139.pdf.

[14] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. Sjoding, J. Wiens, Mimic-iii and eicu-crd: Feature representation by fiddle preprocessing, 2021. URL: https://physionet.org/content/mimic-eicu-fiddle-feature/1.0.0/. doi:10.13026/2QTG-K467.

[15] S. Flood, M. King, R. Rodgers, S. Ruggles, J. R. Warren, Integrated public use microdata

series, current population survey: Version 8.0, 2020. URL: https://www.ipums.org/projects/ipums-cps/d030.V8.0. doi:10.18128/D030.V8.0.

[16] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, Advances in Neural Information Processing Systems 34 (2021).

A. Genetic Longitudinal Counterfactuals

In addition to using our metric after generating counterfactuals, we present a method that leverages longitudinal data for generating counterfactuals using genetic algorithm. In the genetic algorithm, largely borrowed from [5] and [8], we begin by generating a random population of the desired class. Then, we assess the fitness of the population relative to our input and rank the population by fitness. The top half of the population is then mated (i.e. features are randomly chosen between two individuals). The next generation in the algorithm is made up of the top half of the current generation and their offspring. We repeat this cycle until the best fitness does not change substantially.

Algorithm 1 provides pseudocode for the above description. This algorithm is flexible enough to allow for a variety of fitness metrics, and in this paper we use our longitudinal metric to generate counterfactuals constrained by longitudinal distances. 2 and 3 show two metrics that are used in [5] to generate counterfactuals.

Algorithm 1 Genetic Counterfactuals

```
Input: (subject input $x$, desired outcome $z$, model $M$)
POP $\leftarrow$ IntialPopulation($x, z$)
currentBest $\leftarrow \infty$
repeat
    prevBest $\leftarrow$ currentBest
    mostFit $\leftarrow$ SelectFittest($x, z, M$)
    currnetBest $\leftarrow$ BestFitness(mostFit)
    POP $\leftarrow$ Mate(mostFit)
until currentBest $\approx$ prevBest
return POP
```

$$L_{prox}(x,e) = \left(\sum_{i \in Continuous} \frac{1}{MAD(X_i)} |x_i - e_i| \right) + \left(\sum_{i \in Categorical} (x_i \neq e_i) \right) \tag{2}$$

$$L_{sparse}(x,e) = \sum_i (x_i \neq e_i) \tag{3}$$

A.1. Experiment with *Adult-Income* Dataset

We performed an experiment to compare counterfactuals generated with and without our longitudinal metric. In the 'Default' algorithm, we optimize sparsity and proximity, and in the

’Longitudinal’ algorithm, we optimize proximity and longitudinal distance.

We use Adult-Income, a common dataset used in fairness and explainability research [15] [16]. The task is to use an individual’s demographic and economic information to predict if their income is above 50k. To augment our experiment, we also consider a threshold of 30k. Having two different thresholds should help us understand how plausibility interacts with the rarity of a desired decision. In Adult-Income, 24 percent of individuals have an income above 50k, compared to 44 percent who have an income above 30k. We expect that the lower threshold will lead to more plausible counterfactuals and higher validity for both the ’Default’ and ’Longitudinal’ methods.

Though the dataset does not contain any longitudinal data, it is simple enough to reason about what data subjects might look like over time. Therefore, we conduct a simple simulation to generate longitudinal data: we randomly allow some individuals to swap careers with someone else in their education class. We also allow some individuals to increase their level of education before moving to a new career. When swapping careers, all non-demographic variables are swapped (hours-per-week, occupation, and capital loss/gain). Finally, the simulation increases age randomly between one and ten years. This simulation shows some of the ways people can change their economic conditions without allowing any changes on immutable features, such as race or nationality. However, some features we do not include, such as marital status, can change in practice. We focus on allowing changes to features that could potentially be included in a recommendation.

Metrics for validity are presented in Table 1. Example counterfactuals are presented in Table 2. Overall, we find that the ’Longitudinal’ algorithms produces fewer valid counterfactuals, but also fewer counterfactuals with impossible changes.

	30k		50k	
Metric	Default	Longitudinal	Default	Longitudinal
Mean Validity	0.96	0.90	0.93	0.53
% validity=0	0	3	2	22
% validity=1	75	65	72	31
% immutable	73	0	84	0

Table 1

Metrics for ’Default’ and ’Longitudinal’ methods across both thresholds. *Validity* refers to whether or not a counterfactual is of the desired class. Since each individual in the test set receives ten counterfactuals, we calculate the mean validity across the those individuals. The last three rows refer to the percent of individuals who have no valid counterfactuals, the percent who have ten counterfactuals, and the percent of counterfactuals that change an immutable feature.

	Age	HrsWk	Workclass	Edu	Marital	Relationship	Race	Gender	Native-Cntry	Occupation	Prediction
Sub. A	33	40	Fed-gov	Some-col	Never	Not-in-fam	Black	Female	US	Exec	<50k
Def.	38	-	-	-	Married	Husband	White	-	-	-	>50k
	39	-	-	-	Married	Wife	-	-	Scotland	-	>50k
Long.	35	-	-	-	-	-	-	-	-	Transport	<50k
	37	50	-	-	-	-	-	-	-	Transport	<50k
Sub. B	37	55	State-gov	HS	Married	Husband	White	Male	US	Protect-serv	<50k
Def.	-	54	Self-emp	-	-	-	-	-	-	Sales	>50k
	-	54	-	-	-	-	-	-	Taiwan	Exec	>50k
Long.	53	48	-	Assoc-voc	-	-	-	-	-	-	>50k
	46	56	-	Bachelors	-	-	-	-	-	Armed-forces	>50k

Table 2: The top explanations for two individuals for both the 'Default' and 'Longitudinal' methods. For both individuals, we desire counterfactuals with a prediction of >50k.