Exploring counterfactual antecedents to crime analisys

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

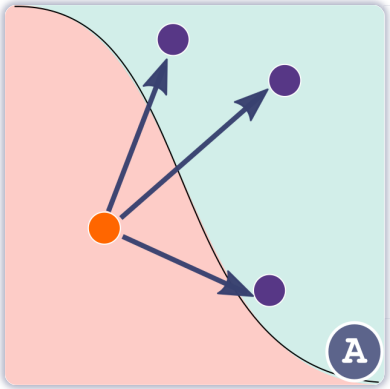
May 20th, 2021 - Rio de Janeiro - Brazil

Counterfactual antecedent

What is a counterfactual antecedent?

Readable antecedents - "If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5, you would have a score of 0.51."

Table:



	Features	Original Values	CFA Values
CF1	Passengers	4.11	1.33
CF2	Expansion Phase	1	5
CF3	Bus Stops	0.69	0.47
	HighIncomeHolder	0.0	0.15
CF4	Bus Stos	0.69	0.41
	HighIncomeHolder	0.0	0.07

B

Figure 1: Example of the concept of counterfactual antecedents.

The use in explainable AI

The usual approach to explanation: Focuses primarily on an explanation of the internal structure of the algorithms and how it led to the decisions.

Counterfactual approach to explanation: Describes dependency on the external facts that led to the decision.

Mathematical definition

Let's suppose a learning machine $f(\theta, \mathbf{x})$:

- $f(\bullet)$ - is the decision function.
- θ - is the parameter vector, already adjusted to a dataset.
- $\mathbf{x}$ - is a sample.

a conterfactual explanation consists in a synthetic sample $\mathbf{x}' = \mathbf{x} + \mathbf{a}$ that achieves a desired outcome y' in similarity $f(\theta, \mathbf{x}') \approx y'$ or constraint $f(\theta, \mathbf{x}') \geq y'$.

Important property: reduce the cost $c(\bullet)$ of changing an instance. So, $\min c(\mathbf{x}, \mathbf{x}')$.

Combinatorial optimization

Problem definition

Given that an action on any feature $\mathbf{a}_i$ can assume a set of values $\{\mathbf{a}_i^1 \ldots, \mathbf{a}_i^{k_i}\}$, k_i being the number of changes available for feature i. The decision of the set of features and the intensity of the change for each feature that results in changing the outcome with minimal cost.

This can be formulated as:

$$\begin{aligned} \min_{\mathbf{a}} \quad & cost(\mathbf{a}) \\ s.t. \quad & f(\mathbf{x} + \mathbf{a}) \geq y' \\ & a \in A(\mathbf{x}). \end{aligned} \tag{1}$$

$A(\mathbf{x})$ is the set of possible actions of $\mathbf{x}$,

$cost(\mathbf{a})$ have to increase with the increase of $\mathbf{a}$.

Change in a single direction, thus $\mathbf{a} \geq \mathbf{0}$ to simplify.

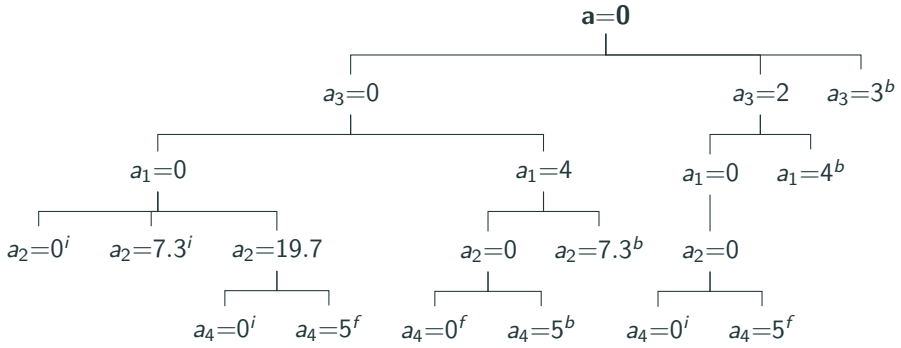
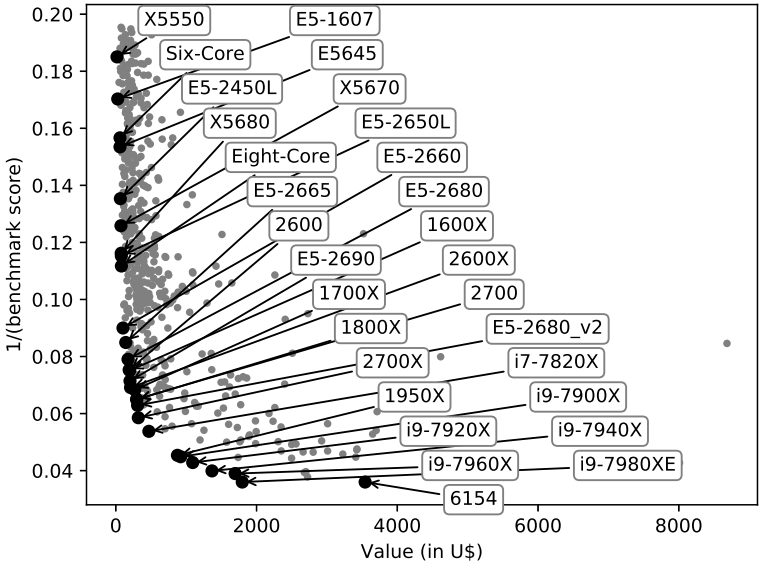


Figure 2: Tree of a branch and bound set of decisions.

Multi-objective optimization

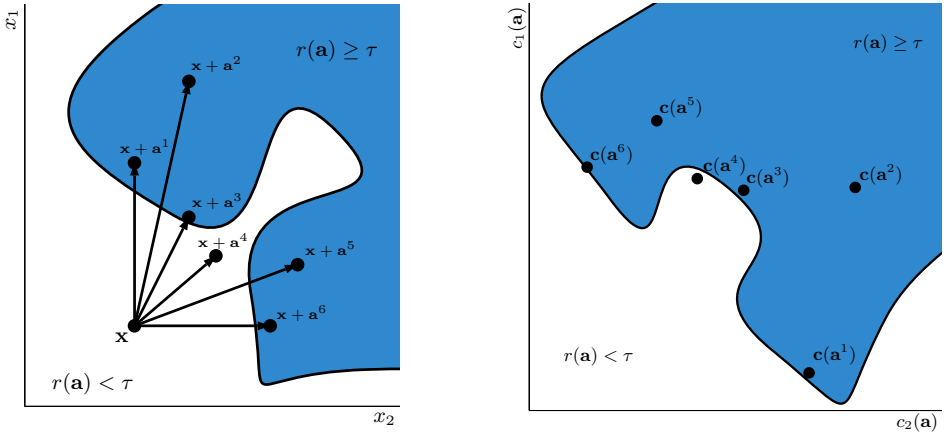
Example


X5550
E5-1607
Six-Core
E5645
E5-2450L
X5670
X5680
E5-2650L
Eight-Core
E5-2660
E5-2665
E5-2680
2600
1600X
E5-2690
2600X
1700X
2700
1800X
E5-2680_v2
2700X
i7-7820X
1950X
i9-7900X
i9-7920X
i9-7940X
i9-7960X
i9-7980XE
6154
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
1/(benchmark score)
0
2000
4000
6000
8000
Value (in U$)

Definition

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{f}(\mathbf{x}) \equiv \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\} \\ \text{sujeito a} \quad & \mathbf{x} \in \Omega, \Omega \in \mathbb{R}^n \\ & \mathbf{f}(\mathbf{x}) : \Omega \to \Psi \in \mathbb{R}^m \\ & f_i(\mathbf{x}) : \Omega \to \mathbb{R}, i = 1, 2, \ldots, m. \end{aligned} \tag{2}$$

Representation of a multi-objective problem



(a) Feature space

(b) Objective space

Figure 4: Representation of feature space (a) and objective space (b), taking two features and two objectives.

Definitions

Definition
- Feasible action.

An action $\mathbf{a} \in \mathbb{R}^d$ belongs to the feasible set of solutions $\mathcal{A}$ if and only if it achieves the desired outcome $r(\mathbf{x} + \mathbf{a}) \geq \tau$.

Definition
- Partial order.

Ordering relation on partially ordered sets occurs when all components are ordered in the same sense. We use the symbols $\preceq$ and $\succeq$ to describe the ordering relations on partially ordered sets; for example, $\mathbf{x} \preceq \mathbf{y}$ is equivalent to $x_i \leq y_i$, $\forall i \in \{1, \ldots, m\}$.

Definition
- Dominant action. A feasible action $\mathbf{a} : r(\mathbf{x} + \mathbf{a}) \geq \tau$ dominates $\mathbf{a}'$ if and only if $\mathbf{c}(\mathbf{a}) \preceq \mathbf{c}(\mathbf{a}')$.

Definition
- Pareto-optimal action.

Consider an objective function vector $\mathbf{c}(\bullet) : \mathbf{R}^d \Rightarrow \mathbf{R}^m$ that we want to minimize, and a feasible set of solutions $\mathcal{A}$. An action $\mathbf{a}^*$ is Pareto-optimal iff there is no action $\mathbf{a} \in \mathcal{A}$ that dominates $\mathbf{a}^*$.

Definition
- Monotonicity w.r.t. a partial order.

Given any two actions $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{a}' \in \mathbb{R}^d$ such that $a_i \geq a'_i, \forall\ i \in \{1, \ldots, d\}$, a function vector $\mathbf{f}(\bullet) : \mathbf{R}^d \Rightarrow \mathbf{R}^m$ is monotone if only if $f_j(\mathbf{a}) \geq f_j(\mathbf{a}'),\ \forall j \in \{1, \ldots, m\}$.

Algorithm

Algorithm 1 Model-Agnostic Pareto-optimal Counterfactual Antecedents Mining

Require: An sample $\mathbf{x}$, a objective function $c(\bullet)$, a decision rule $r(\bullet)$, a threshold τ, and number of allowed changes k.

procedure ENUMERATE($\mathbf{a}$, $\mathcal{D}$, $\mathcal{A}$)
  if $|i : a_i \neq 0 \ \forall i \in \mathcal{D}| > k$ or $\exists \mathbf{a}' \in \mathcal{A} : c(\mathbf{a}) \succeq c(\mathbf{a}')$ **then**
    return
  end if
  if $r(\bullet)$ is monotone and $r(\mathbf{x} + \overline{\mathbf{a}}^*) < \tau$ **then**
    return
  end if
  if $r(\mathbf{x} + \mathbf{a}) \geq \tau$ **then**
    $\mathcal{A} = \mathcal{A} \cup \{\mathbf{a}\}$
    return
  end if
  $i =$SELECT_FEATURE$(\forall i : i \notin \mathcal{D})$
  for $\forall \mathbf{a}' : a'_i \geq a_i$ **do**
    ENUMERATE$(\mathbf{a}', \mathcal{D} \cup \{i\}, \mathcal{A})$
  end for
end procedure
$\mathcal{A} = \{\}$, $\mathcal{D}^0 = \{\}$
$\mathbf{a}_i^0 = 0 \ \forall i \in \{1, \ldots, d\}$.
ENUMERATE$(\mathbf{a}^0, \mathcal{D}^0, \mathcal{A})$
return $\mathcal{A}$

*$\overline{\mathbf{a}}^*$ is the maximal achievable action $\overline{a}_i^* = \begin{cases} a_i, & \text{if } i \in \mathcal{D} \\ \max a_i, & \text{otherwise} \end{cases}$.

Properties

- MAPOCAM finds all Pareto-optimal solutions.
- Time complexity of MAPOCAM is $T(d, k) = \mathcal{O}((bd)^k)$, where d is the number of variables, b is the maximum number of possible states of each variable, and k is the maximal number of allowed changes.

Definition
Maximal percentile change (MPC).

$$l_j(\mathbf{x}) = \frac{|\{i|\hat{x}_j^i \geq x_j, \forall i \in \{1, \ldots, N\}|}{N} \times 100 \tag{3}$$

Metric $\max(|l_j(\mathbf{x} + \mathbf{a}) - l_j(\mathbf{x})|,\ j \in \{1, \ldots, d\})$.

Definition
Number of changes.

Consists on counting the number of changes (non-zero values) of an action $\mathbf{a}$:
$c(\mathbf{a}) = |\{\mathbf{a}_j|\mathbf{a}_j \neq 0, \forall j \in \{1, \ldots, d\}|$.

Definition
j-th feature change.

The feature change for the feature j consists on the magnitude of an action $\mathbf{a}$ for the feature j: $c_j(\mathbf{a}) = a_j$.

Details - Monotonicity of classifiers

Used to speed-up search.

- Linear classifiers.
- Enforced monotonicity.
- Trees.

We can also ignore the lack of monotonicity (and other properties) and do not speed up search.

Results

Two objectives – Logistic Regression

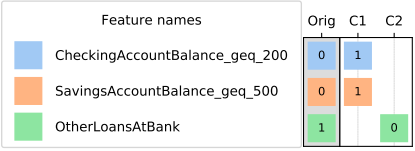


Figure 5: Enumeration for Logistic Regression of counterfactual antecedents with Pareto-optimal values when the MPC cost and the number of changes are the objectives. Orig column is the original sample and the columns C1 and C2 are counterfactual antecedents.

Multiple objective – Logistic Regression

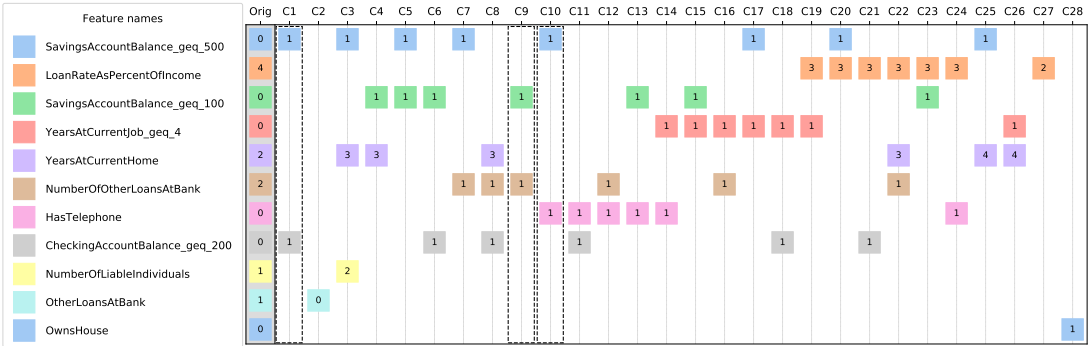
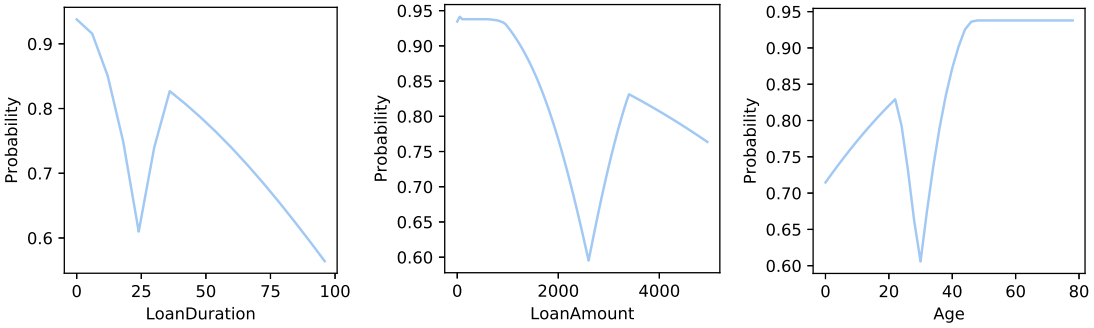


Figure 6: Enumeration for Logistic Regression of counterfactual antecedents with Pareto-optimal values when each feature is considered as an objective function. Orig column is the original sample and the columns C1 to C28 are counterfactual antecedents.

Non-Monotone – Neural Network



(a) `LoanDuration` **(b)** `LoanAmount` **(c)** `Age`

Figure 7: Representation of impact of increasing each variable in the the probability of having credit granted in a multilayer perceptron.

Multiple objective – Neural Networks

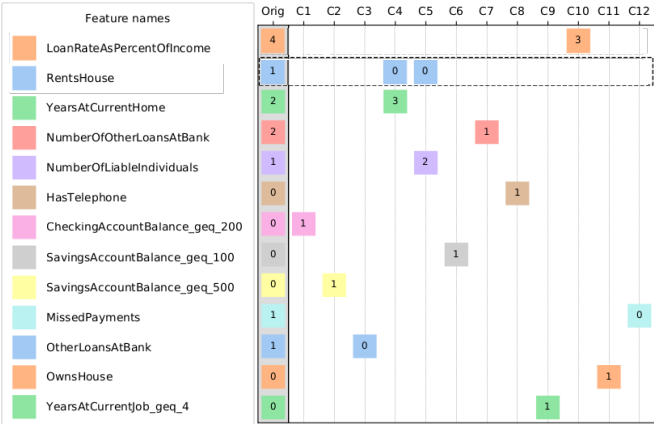


Figure 8: Enumerations for a multilayer perceptron (whose monotonicity is depicted in Fig. 7) with Pareto-optimal values when each feature is considered as an objective function. Orig column is the original sample and the columns C1 to C12 are counterfactual antecedents.

Time Performance and Overview

- In general, MAPOCAM achieves good performance when: (i) the classifiers are monotonic, or (ii) it explores the constraints of achievable leaves in decision trees.
- The execution time is also manageable without any information about the model — generally taking less than one minute.
- Using j-th feature change generate counterfactual antecedents that satisfy any preference (other objective functions) from the user.
- Any other method in the literature needs model-related procedures to work.
- MAPOCAM can create a diverse set of counterfactual antecedents using multi-objective optimization.
- MAPOCAM enumerates all proper counterfactual antecedents without resorting to norms and adjusting their parameters.

Crime in São Paulo - Dataset

- Data has 18953 regions, the number of crimes for each region is the sum of crime incidences from 2006 to 2017, and we have 21 urban and socioeconomic variables.
- Crime data has been provided by the Center for the Study of Violence in the University of São Paulo (NEV-USP)[1].
- Urban infrastructure data such as the location of schools, bus stops, and bars was provided by the Center for Metropolitan Studies (CEM-USP)[2].
- Housing, sanitary conditions, and populational profile was provided 2010's census from Brazilian Institute of Geography and Statistics (IBGE).
- Urban mobility was provided by São Paulo's subway system.

[1] `nevusp.org`

[2] `centrodametropole.fflch.usp.br/pt-br`

Crime in São Paulo

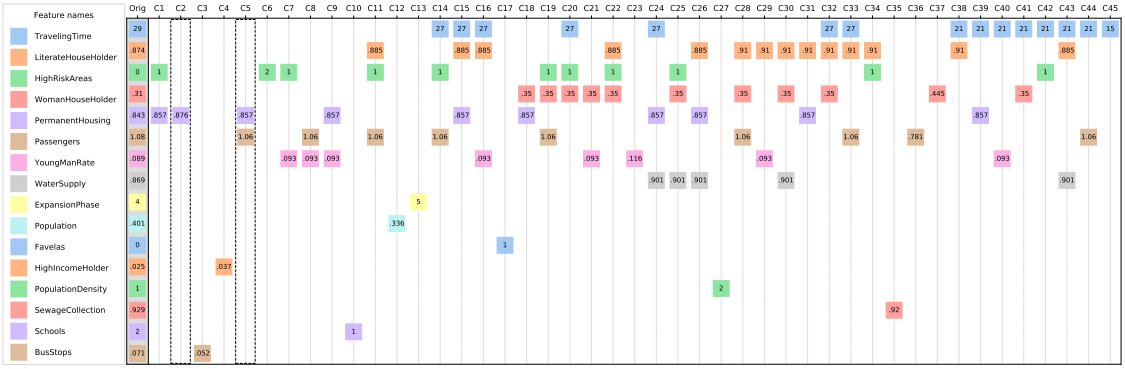


Figure 9: A multi-objective enumeration of actions that reduce the criminality rate of a region of São Paulo. Orig column is the original sample and the columns C1 to C45 are counterfactual antecedents.

CounterCrime

Motivation

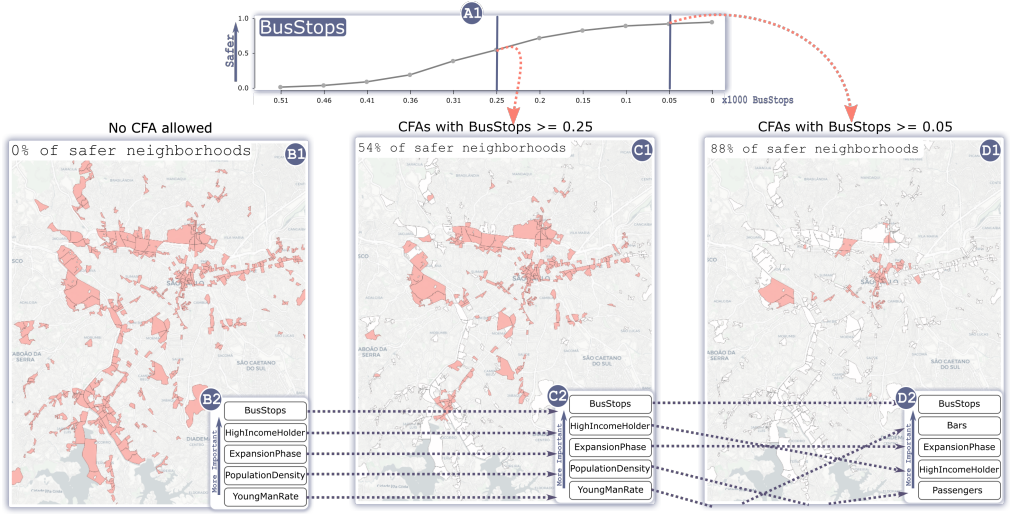


Figure 10: A case study investigating the impact of allowing counterfactual antecedents that reduce the number of BusStops in the criminality.

Identifying hotspots using urban and socioeconomic features

Each census region corresponds to an instance of data whose attributes are the urban and socioeconomic variables. We selected the 500 census regions with the highest number of crime events, labeling those regions as 0 and the remaining ones as 1.

Creating the model. We use a Logistic Regression as the classification model. The model was trained holding 20% of the data for testing, relying on 5-fold cross-validation to select the parameters and l_1 regularization. The performance of the model was 0.90 in AUC.

Defining hotspots. Logistic Regression was used to select the 500 regions with the lowest probability of being classified as safe, computing counterfactual antecedents for these regions.

Generating Counterfactuals. We use MAPOCAM to find CFAs for all regions labeled as unsafe by the classifier.

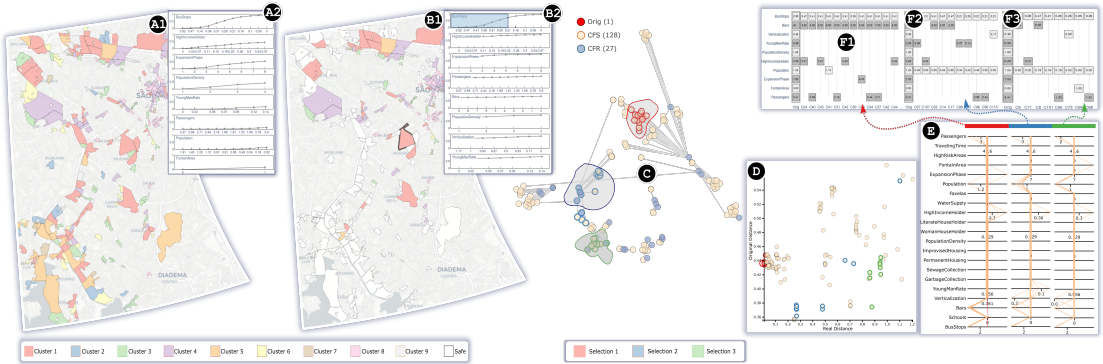


Figure 11: The proposed counterfactual antecedent (CFA) based crime analysis tool, called CounterCrime, that is composed in two main parts: Global analysis (A, and B) and local analysis (C, D, E and F).

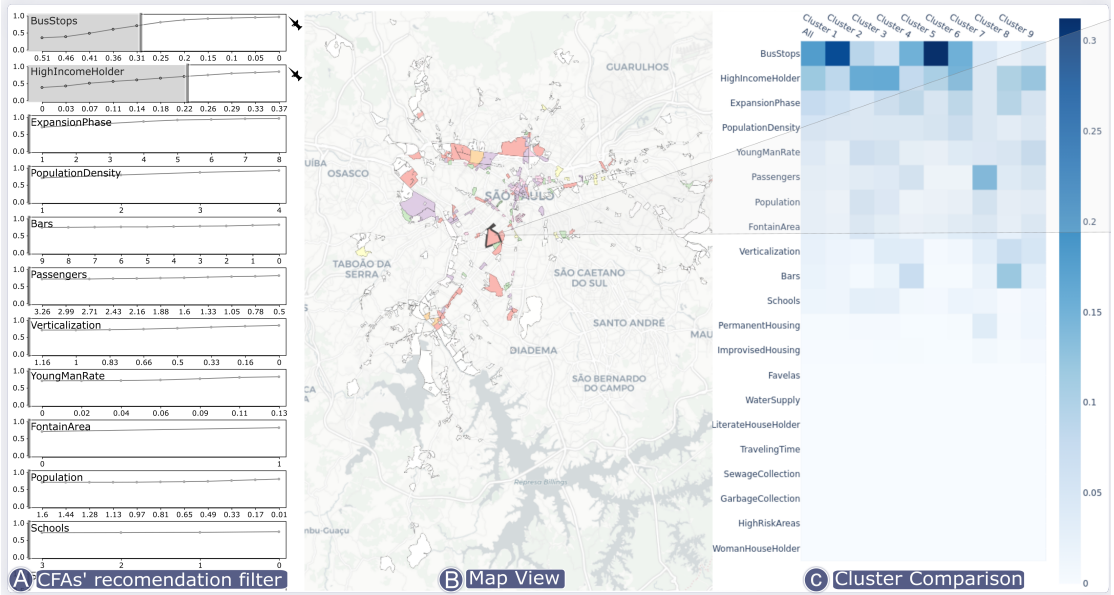
Ranking variables based on their counterfactual importance

- For each unsafe region $h \in \mathcal{H}$ we have a set of counterfactual antecedents $\mathcal{C}_h$.
- c_{ij} is the co-occurrence index indicating the number of times that $a_i \neq 0$ and $a_j \neq 0$.
- Build a stochastic matrix P^h (for each h) are given by $P^h_{ij} = \frac{c_{ij}}{\sum_{j=1}^{d} c_{ij}}$, what ensures that $\sum_{j=0}^{d} P^h_{ij} = 1$.
- Find the stationary eigenvector π^h of P^h satisfies $\pi^h P^h = \pi^h$ and each entry π^h_i indicates the importance of the variable i when computing CFAs for the region h.
- Sorting the entries of π^h we get a ranked list of variables that are more closely related in terms of the CFAs.
- To find π for a set of regions, we average the stochastic matrix belonging to that set.

Clustering regions based on their counterfactual antecedents

- We vectorize the stochastic matrices and use k-means to find clusters of similar regions.
- We also show the value of the stationary eigenvector (variable importance) for each cluster.

CounterCrime - Global



Inspecting counterfactual antecedents for a single region

For each found counterfactual antecedent we matched with the nearest safe region.

1. We projected the counterfactual antecedents and the equivalent safer regions using T-SNE.
2. We also calculate the distance from the counterfactual antecedents to the original values and the nearest safer regions.

CounterCrime - Local

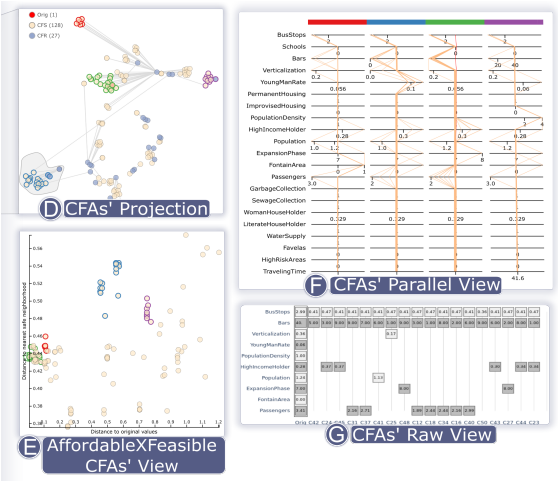


Figure 13: Local visualization of CounterCrime.

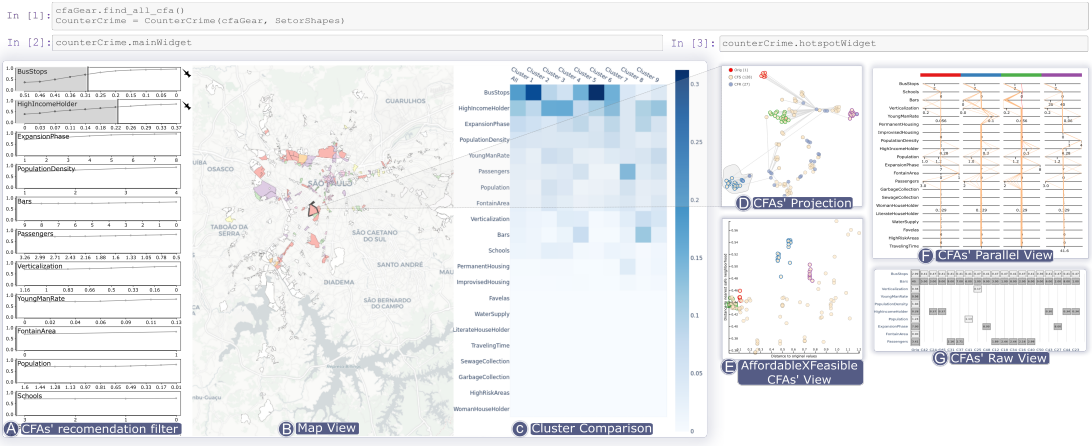


Figure 14: The proposed counterfactual antecedent (CFA) based crime analysis tool, called CounterCrime.

Case Study

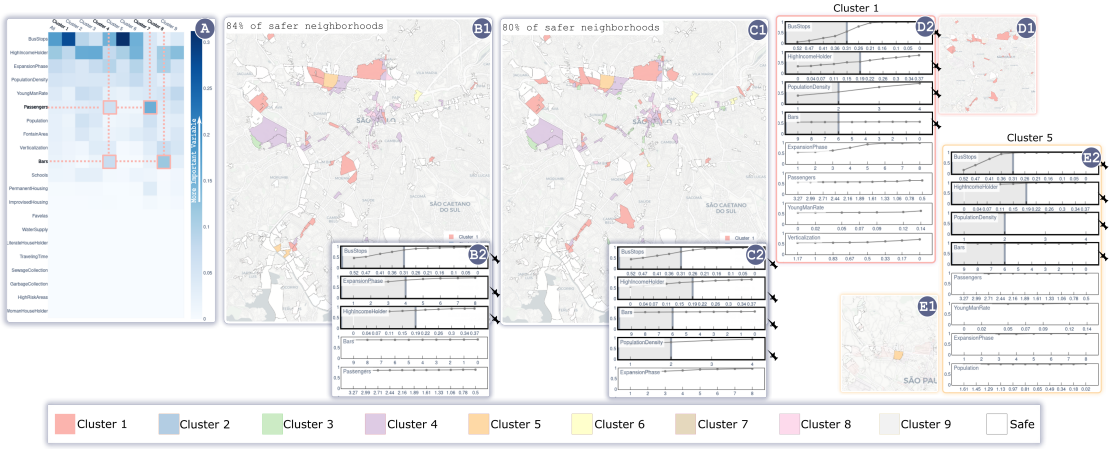


Figure 15: Case study investigating crime in the whole city.

Exploring counterfactual antecedents to crime analisys

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

May 20th, 2021 - Rio de Janeiro - Brazil