2022/08/06

理論談話会#16

理論合宿

# A statistical approach to small area synthetic population generation as a basis for carless evacuation planning

Mohammad Motalleb Nejad, Sevgi Erdogan, Cinzia Cirillo

*Journal of Transport Geography*, Vol. 90, 102902, (2021)

M1　近藤愛子

# 論文構成



1. Introduction
2. Literature review
3. Small area population estimation problem [課題の特定，問題設定]
4. Methodology
5. Anne Arundel county application [実証分析]
6. Conclusions

# Introduction



問題意識

- 低所得、車を所有しない住民が最も影響を受けやすい
  （ハリケーンKatrina, Rita, Maria)
- 彼等が避難できるよう、 (公共)交通サービスの提供が必要

効果的な避難輸送サービスの提供のため

目標　車を持っていない脆弱な層が
**どこに住んでいるのか**を知る


- 「車避難が前提」という点は日本とは違う？
- 日本では徒歩避難してほしいのに、の方が問題としては大きい
  津波や、豪雨などのハザード直前の避難で渋滞
- 豪雨・台風避難で、事前の広域避難とかがポピュラーになれば、大事になりそう
  交通弱者の特定など

# Introduction



目標

車を持っていない脆弱な層が
**どこに住んでいるのか**を知る

車保有と社会人口学的特性に相関があれば特定できるが…

課題

- 特性は標本調査のものしかわからない
- その標本の取得範囲(PUMA*)は、災害で被害を受ける範囲より大きい
  本研究の調査対象のAnne Arundel County(1070km²)は4つのPUMAに分けられる
  23区の面積が約630km²
- Census tractなら、災害リスクを考慮するのに十分な小ささだが、サンプルが少ない
  1000-8000人規模。PUMAの下位ネスト

*Public Use Microdata Area

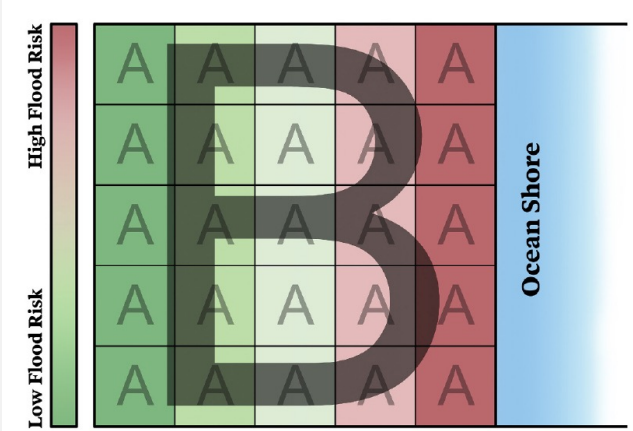



Fig. 1. Two layers of geographical area, large geographical area (B), and small area (A).

十分に説明している、と言える数の標本を得られるのはB
避難を考慮するにはAレベルでの情報が必要（緑の地域は避難必要ない）

▷**小地域内人口の特性**を適切に表す、**合成人口を生成する**手法の開発

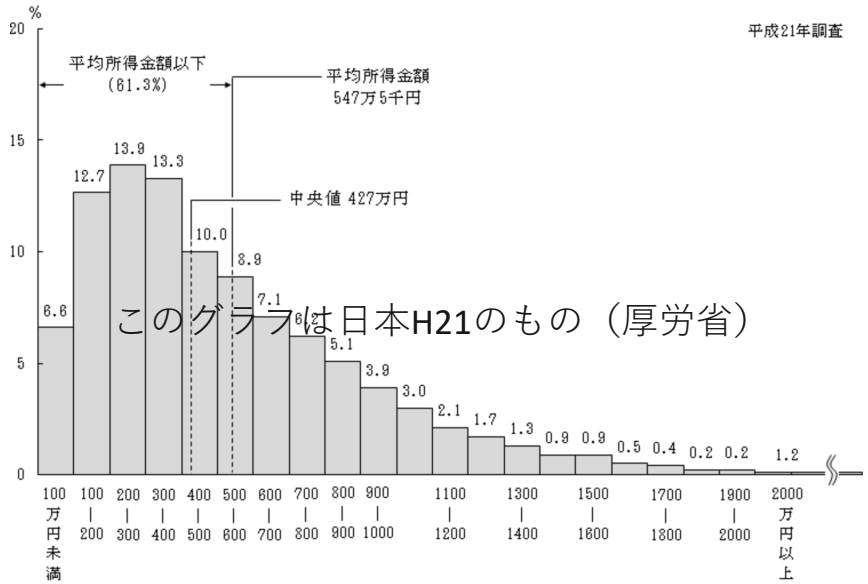
# 小地域人口推定-必要な条件



十分な信頼性を得るためには、合成人口は以下の2つの条件を満たす必要

1. 地域ごとの、要素間の従属関係が保存される
   - 地域内では、同じ社会人口学的特性（e.g. 教育、世帯収入、婚姻歴等）を持つ個人は、子供の人数や車の有無について、同じ傾向を持っている
   - ただし、地域間でこの傾向は異なる

2. 各要素の分布が実データと一致する
   - 例えば、世帯収入の分布




このグラフは日本H21のもの（厚労省）

# 小地域人口推定-方針



Census tract (=右図のAに相当)では、サンプルが少なすぎるため、1. の従属関係を導くのは困難

必要な条件
1. 要素間の従属関係の保存
2. 各要素の分布が実データと一致

仮定

PUMA (=右図のBに相当)内では、
1. の従属関係は同じ

▷ PUMAのデータより従属関係を推定し、
Census tractの各要素の分布を満たす人口を合成

従属関係の構造をcopula estimationにより捉える

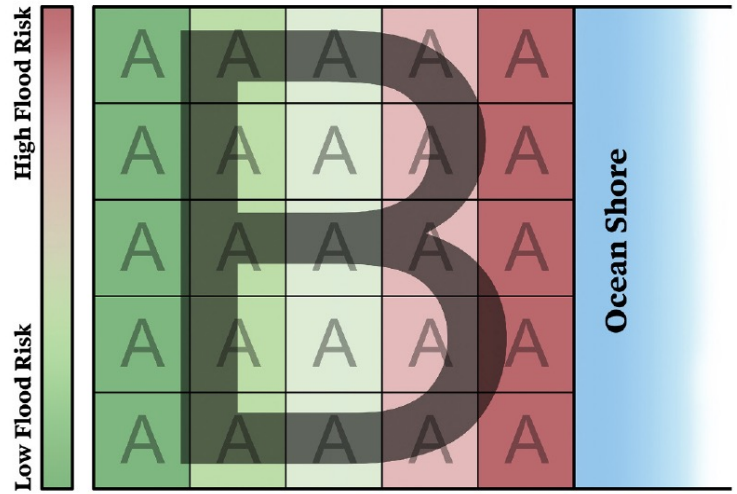



Fig. 1. Two layers of geographical area, large geographical area (B), and small area (A).

# Copulaとは



**Copulaとは?**　Sklar(1959)が提唱

$d$次元の周辺分布を、同時分布に写像する関数$C$をCopula関数と呼ぶ

$$C: [0,1]^d_{marginal\ distributions\ of\ d\ dimensional\ multivariate\ data} \rightarrow [0,1]^1 (1)$$

$H$を$d$次元の連続変数の同時分布とし、$F_k$を$k$番目の変数の周辺分布とすると、

$$H(x_1, x_2, \dots, x_d) = C\{F_1(c_1), F_2(x_2), \dots F_d(x_d)\}\ x_1, x_2, \dots, x_d \in \mathbb{R} (2)$$

$n$個の観測からなる多変量データがある時、$i$番目の観測、$k$次元目の変数に対して経験周辺分布$F_k^i$とランク$R_k^i$は

$$F_k^i = \frac{R_k^i}{n} (3)$$

$$R_k^i = \sum_{j=1}^{n} 1(x_k^j < x_k^i) (4)$$

# Copulaとは



式(1)-(4)を用いて、経験コピュラ(empirical copula) は

$$C_n\left(f_{1(x_1)}, \ldots, f_d(x_d)\right) = \frac{1}{n}\sum_{j=1}^{n} 1\left(F_1^j(x_1) < f_1, \ldots, F_d^j(x_d) < f_d\right) (5)$$

Sklarはすべての多変量連続変数群に対して、式(1)を満たす少なくとも１つのparametric関数が存在することを示した

経験コピュラ（観測されているデータ）に近いparametricコピュラを見つけたい


**本研究で用いられたコピュラ関数→**
Archimedean copulaというfamily
陽関数で、１つのパラメータ$\theta$で
記述できるという特徴がある

**Table 1**
Three most popular Archimedean copulas.

| Copula | Multivariate Copula $C_\theta$ | Range of the θ |
|---|---|---|
| Clayton | $[max\{f_1^{-\theta} + f_2^{-\theta} + \ldots f_d^{-\theta} - d + 1; 0\}]^{-1/\theta}$ | $\theta \in [-1, \infty]\backslash\{0\}$ |
| Frank | $-\frac{1}{\theta} loglog\left[1 + \frac{\prod_{i=1}^{d}(exp(-\theta f_1) - 1)}{exp(-\theta) - 1}\right]$ | $\theta \in R\backslash\{0\}$ |
| Gumbel | $exp[-(\sum_{i=1}^{d}(-loglog f_1)^{\theta})^{1/\theta}]$ | $\theta \in [1, \infty)$ |

# Copulaの選定



選んだcopula関数が従属関係をうまく表現できているか？

▷統計検定

**帰無仮説：**従属関係をあらわす$C_\theta$(=真のコピュラ?)が選定したcopula familyに含まれる

経験コピュラとパラメトリックコピュラの距離を用いた方法 Genest and Remillard(2008), Genest et al.(2009)

距離はFermanian(2005)による$C_n(u) = \sqrt{n}(C_n - C_{\theta_n})$

Cramér-Von-Mises statistics $S_n = \int_{[0,1]^d} C_n(U)^2 dC_n(U) = \sum_{i=1}^{n} \left(C_n(U_i) - C_{\theta_n}(U_i)\right)^2$ (8)

累積分布関数$F^*$と経験分布$F_n$を比べる指標

経験コピュラ

パラメトリックコピュラ

$n \to \infty$のときの$S_n$の極限分布はわからない、parametric bootstrapで近似する→P値の計算

P値が大きい→棄却することができない→copula familyにfitする

普通は、帰無仮説を棄却して逆を示す、、

# Bootstrap法で$P$値を求める



1. 元データ$X_n^d$
   i. データにターゲットコピュラを最尤法でfitさせる: $C_{\hat{\theta}}$
   ii. データの経験コピュラ: $C_n$
   iii. $S_n$を計算

2. データをサンプリング$X_n^{*d}$
   $X_n^{*d}$に対してi.-iii.を繰り返す

▷2.をK回繰り返す。$i$回目の$S_n^*$を$S_{n,i}^*$と書くと$P$値は

$$P = \frac{1}{K+1}\left[\sum_{i=1}^{K} 1\left(S_{n,i}^* \geq S_n\right) + 0.5\right] \quad (9)$$

# 合成人口の生成



**Copulaが統計検定をクリアしたら、それを用いて合成人口を生成する**

↓大きい範囲のデータ(PUMA)

Copulaは擬似観測値を与える i.e.依存関係は保たれているが、標準化?された値

累積分布関数の逆関数を用いて、合成人口に変換

↑小さい範囲のデータ(Census Tract)

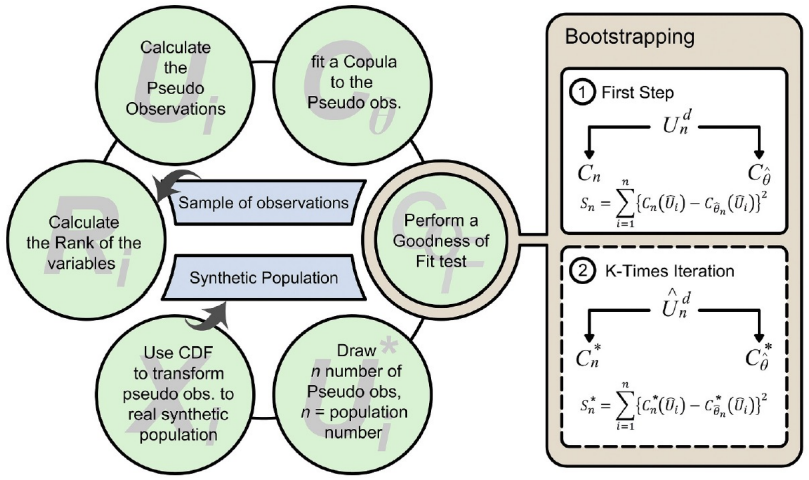



Fig. 3. Flowchart of the population-synthesizing algorithm.

データから合成人口を得るまでのフローチャート

累積分布関数$F$に従う乱数$X$を発生させたいとしたら
→[0,1]の一様分布の乱数$U$を用いて、
$X = F^{-1}(U)$とすればいい

▽

特定の依存関係をもつ乱数列$X_1, \dots, X_d$がほしいので
→コピュラ$C(u_1, \dots, u_d)$に従う乱数ベクトル$U_1, \dots, U_d$を発生させ、
$X_i = F^{-1}(Ui)$とすればよい

# 実証分析



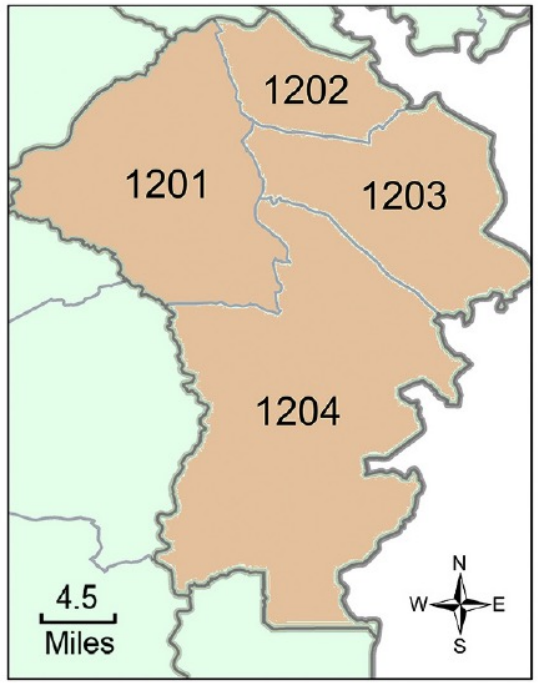
対象地域：Anne Arundel county

面積：1070km$^2$

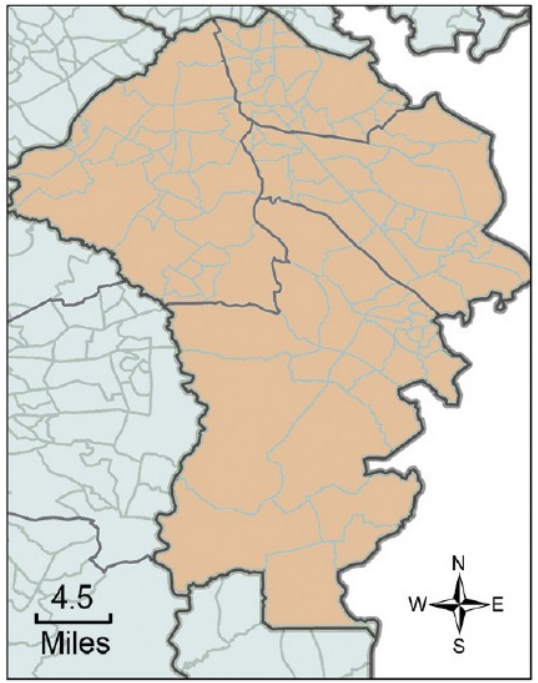
PUMAが4つ含まれる

使用データ：

1. PUMS (Public Use Micro Samples)
   - あるPUMAに属するサンプル
   - American Community Survey
2. 2010 Decennial Census Data
   - Census tractレベル
   - 10年ごと、国勢調査のようなもの
3. IRS Income Data
   - Zip codeごと
   - 各census tract内に、そのzip codeを持つ人の割合がわかる→Census tractレベルに変換




PUMA Level




Census Tract Level

# 合成人口の生成（実証）



Archimedean copula:

Clayton, Gumbel, Frank, Joe

で検定を実施

→Claytonを採用

**Table 1** 再掲: parametricコピュラ
Three most popular Archimedean copulas.

| Copula | Multivariate Copula $C_\theta$ | Range of the θ |
|---|---|---|
| Clayton | $[max\{f_1^{-\theta} + f_2^{-\theta} + \ldots f_d^{-\theta} - d + 1; 0\}]^{-1/\theta}$ | $\theta \in [-1, \infty] \backslash \{0\}$ |
| Frank | $-\frac{1}{\theta} loglog\left[1 + \frac{\prod_{i=1}^{d}(exp(-\theta f_1) - 1)}{exp(-\theta) - 1}\right]$ | $\theta \in R \backslash \{0\}$ |
| Gumbel | $exp[-(\sum_{i=1}^{d}(-loglog f_1)^{\theta})^{1/\theta}]$ | $\theta \in [1, \infty)$ |

各PUMAに対するコピュラのパラメータは以下の通り

**Table 2**
Estimated parameters for Clayton copula for census tracts within Anne Arundel county.

| PUMA | Copula | Statistics | P-value | Estimated parameter (θ) |
|---|---|---|---|---|
| 1201 | Clayton | 156.74 | 0.9995 | 0.30082 |
| 1202 | Clayton | 139.65 | 0.9995 | 0.33478 |
| 1203 | Clayton | 173.25 | 0.9995 | 0.28861 |
| 1204 | Clayton | 158.7 | 0.9995 | 0.29214 |

▷

各census tractの人口(人数)に対応する擬似観測値を生成

↓

Census data, IRS Income data から世帯人数, 世帯の種類, 収入, 子供の有無と年齢, 家庭の就労者の数, 年齢, 性別, 雇用状況, 人種の情報を持つ合成人口の生成

# 車を持たない人口の特定



車の所有有無を二項ロジットで推定 (Bierlaire, 2018)

データ：Maryland州の2017 National Household Travel Survey

$$U_{\text{no car}} = 0$$

$$U_{\text{car}} = \beta_0 + \beta_{\text{CH1}} + \beta_{\text{WIF}} + \beta_{\text{inc1}} + \beta_{\text{inc2}} + \beta_{\text{inc3}} + \beta_{\text{R1}} + \beta_{\text{R2}} \#(10)$$

定数項　世帯の就労者数　子供の数　収入below poverty　収入low middle　収入middle　アフリカ系　アジア系

**Table 3**
Results of Binomial Model used to estimate car ownership.

| Name | Value | Robust std. err | Robust *t*-test | P-value |
|---|---|---|---|---|
| Alternative Specific Constant - 0 car | 0.00 | Fixed | | |
| Alternative Specific Constant – at least one car | 0.583 | 0.180 | 3.24 | 0.00 |
| Number of children in the household (CH1) | 1.01 | 0.802 | 1.26 | 0.21 |
| Personal income < $25 k (inc2) | 1.55 | 0.369 | 4.20 | 0.00 |
| Household income $25 K - $50 k (inc3) | 3.65 | 0.468 | 7.81 | 0.00 |
| Household income $50 K - $75 k (inc4) | 3.93 | 0.588 | 6.68 | 0.00 |
| Race - African American (R1) | −1.35 | 0.562 | −2.40 | 0.02 |
| Race – Asian (R2) | −2.06 | 0.795 | −2.59 | 0.01 |
| Number of workers (WIF) | 0.745 | 0.262 | 2.85 | 0.00 |

# 配分結果



車所有の二項ロジットを、
p14の合成人口に対して配分し可視化

Low income: 世帯収入が $25,000未満と設定

- 色が暗いほど低収入世帯が多い
- 円が大きいほど車を持たない世帯が多い

▽

示唆

- 低収入世帯が多い地域ほど、車を持たない世帯が多い。
- 車を持たない世帯は北部に集中している

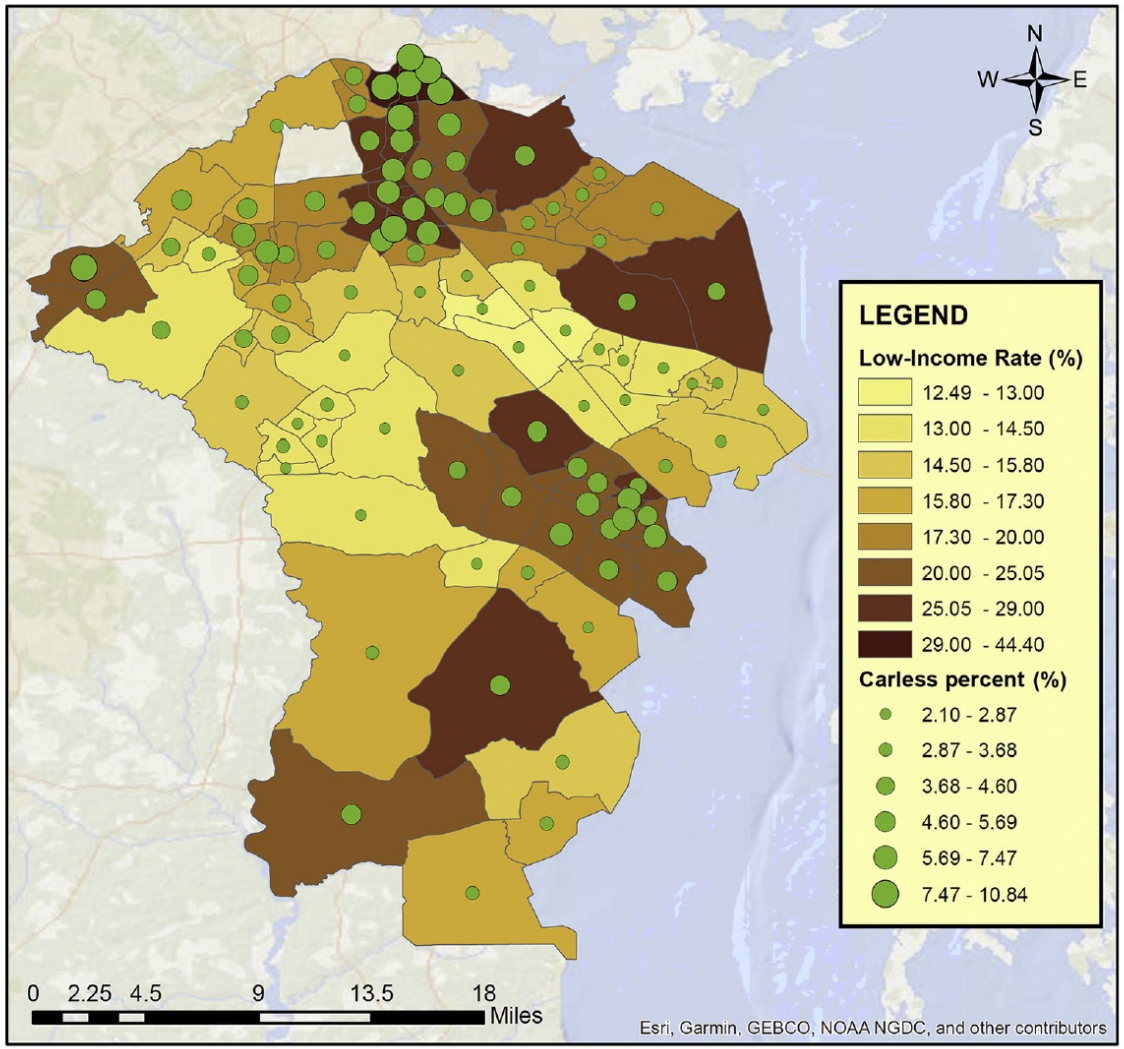



**Fig. 8.** Car ownership and income level for Census Tracts within Anne Arundel County.

# 結論



成果

Copulaをつかった合成人口の生成手法

- 要素間の従属関係、隠された相関関係の保持
- より広い範囲のデータと、狭い範囲のデータを併せて使うことで地域間の差異を表現

課題

- 適切なコピュラ選定のための検定手法の特定
- 細かい範囲でのデータがない/非公開、などのデータの限界

応用例

- 低所得者や障害者、車を持たない世帯の居住部のアクセス性
- アクティビティモデルのより詳細なインプット

▷必要な人に必要なものが届く、（避難）政策の立案

# 参考文献




- Sklar, A., 1959. Fonctions de répartition `a n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231.
- Genest, C., Rémillard, B., 2008. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. Institut Henri Poincaré 44 (6), 1096–1127.
- Genest, C., Rémillard, B., Beaudoin, D., 2009. Goodness-of-fit tests for copulas: a review and a power study. Insurance: Math. Econ. 44 (2), 199–213.
- Fermanian, J.-D., 2005. Goodness-of-fit tests for copulas. J. Multivar. Anal. 95 (1), 119–152.
- Bierlaire, M., 2018. PandasBiogeme: A short introduction. Technical Report TRANSP-OR 181219, Transport and Mobility Laboratory, ENAC, EPFL, Lausanne, Switzerland.