

3D TRAD と nCAR

慶應義塾大学

柴田 里程

TRAD(Textileplot, R and DandD)

- 目標
 - **データを**的確に**理解**する
 - モデルの適合度を的確に判断する
- ヒューマンインタフェース
 - GUI
 - 痒いところに手が届くデザイン（直感に合ったインタフェース）
 - 面倒な操作を覚えたりせず，最小限の手間で理解が進む

AIにできないこと：的確な理解，柔軟な判断，説明，責任

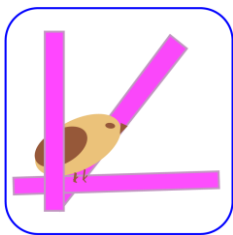
A. Narayanan and S. Kapoor, AI snake oil, Princeton University Press, 2024
(的場知之訳：AI 過大評価社会，草思社，2026)

REFORMS: Reporting Standards for Machine Learning Based Science (<https://arxiv.org/abs/2308.07832>)

- 機械学習はしばしば 的確性, 再現性, 一般性に欠ける
- それが科学研究の現場でも広く使われるようになってきた
- 誤った認識を広め科学の信頼性を損なうことにもなりかねない
- 対応策を検討するためのミーティングの報告
 - 32のチェックリスト

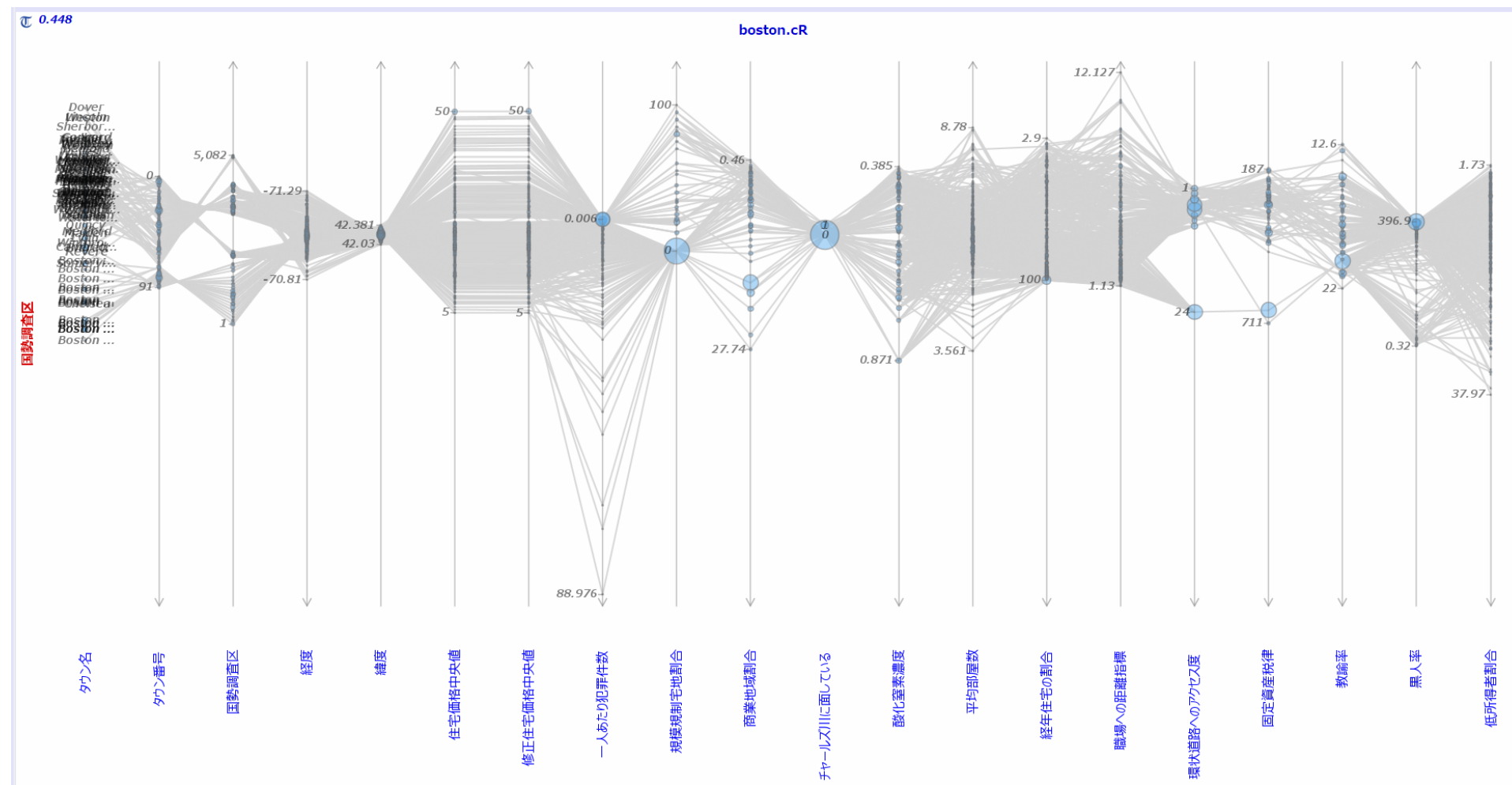
AI（機械学習）の効能を否定するわけではないが
学術研究でさえ「ガマの油売り」に踊らされてしまっている
現状に対する危機感.

Section	Item
Study goals	1a. Population or distribution about which the scientific claim is made
	1b. Motivation for choosing this population or distribution (<i>1a.</i>)
	1c. Motivation for the use of ML methods in the study
Computational reproducibility	2a. Dataset used for training and evaluating the model along with link or DOI to uniquely identify the dataset
	2b. Code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used
	2c. Description of the computing infrastructure used
	2d. README file which contains instructions for generating the results using the provided dataset and code
	2e. Reproduction script to produce all results reported in the paper
Data quality	3a. Source(s) of data, separately for the training and evaluation datasets (if applicable), along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation
	3b. Distribution or set from which the dataset is sampled (i.e., the sampling frame)
	3c. Justification for why the dataset is useful for the modeling task at hand
	3d. The outcome variable of the model, along with descriptive statistics (split by class for a categorical outcome variable) and its definition.
	3e. Sample size and outcome frequencies
	3f. Percentage of missing data, split by class for a categorical outcome variable
	3g. Justification for why the distribution or set from which the dataset is drawn (<i>3b.</i>) is representative of the one about which the scientific claim is being made (<i>1a.</i>)
Data preprocessing	4a. Identification of whether any samples are excluded with a rationale for why they are excluded
	4b. How impossible or corrupt samples are dealt with
	4c. All transformations of the dataset from its raw form (<i>3a.</i>) to the form used in the model, for instance, treatment of missing data and normalization. Preferably through a flow chart
Modeling	5a. Detailed descriptions of all models trained
	5b. Justification for the choice of model types implemented
	5c. Method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds
	5d. Method for selecting the model(s) reported in the paper
	5e. For the model(s) reported in the paper, specify details about the hyperparameter tuning
	5f. Justification that model comparisons are against appropriate baselines
Data leakage	6a. Justification that pre-processing (Section 4) and modeling (Section 5) steps only use information from the training dataset (and not the test dataset)
	6b. Methods to address dependencies or duplicates between the training and test datasets (e.g. different samples from the same patients are kept in the same dataset partition)
	6c. Justification that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage
Metrics and uncertainty	7a. All metrics used to assess and compare model performance (e.g., accuracy, AUROC etc.). Justify that the metric used to select the final model is suitable for the task
	7b. Uncertainty estimates (e.g., confidence intervals, standard deviations), details of how these are calculated
	7c. Justification for the choice of statistical tests (if used) and a check for the assumptions of the statistical test
Generalizability and limitations	8a. Evidence of external validity
	8b. Contexts in which the authors do not expect the study's findings to hold



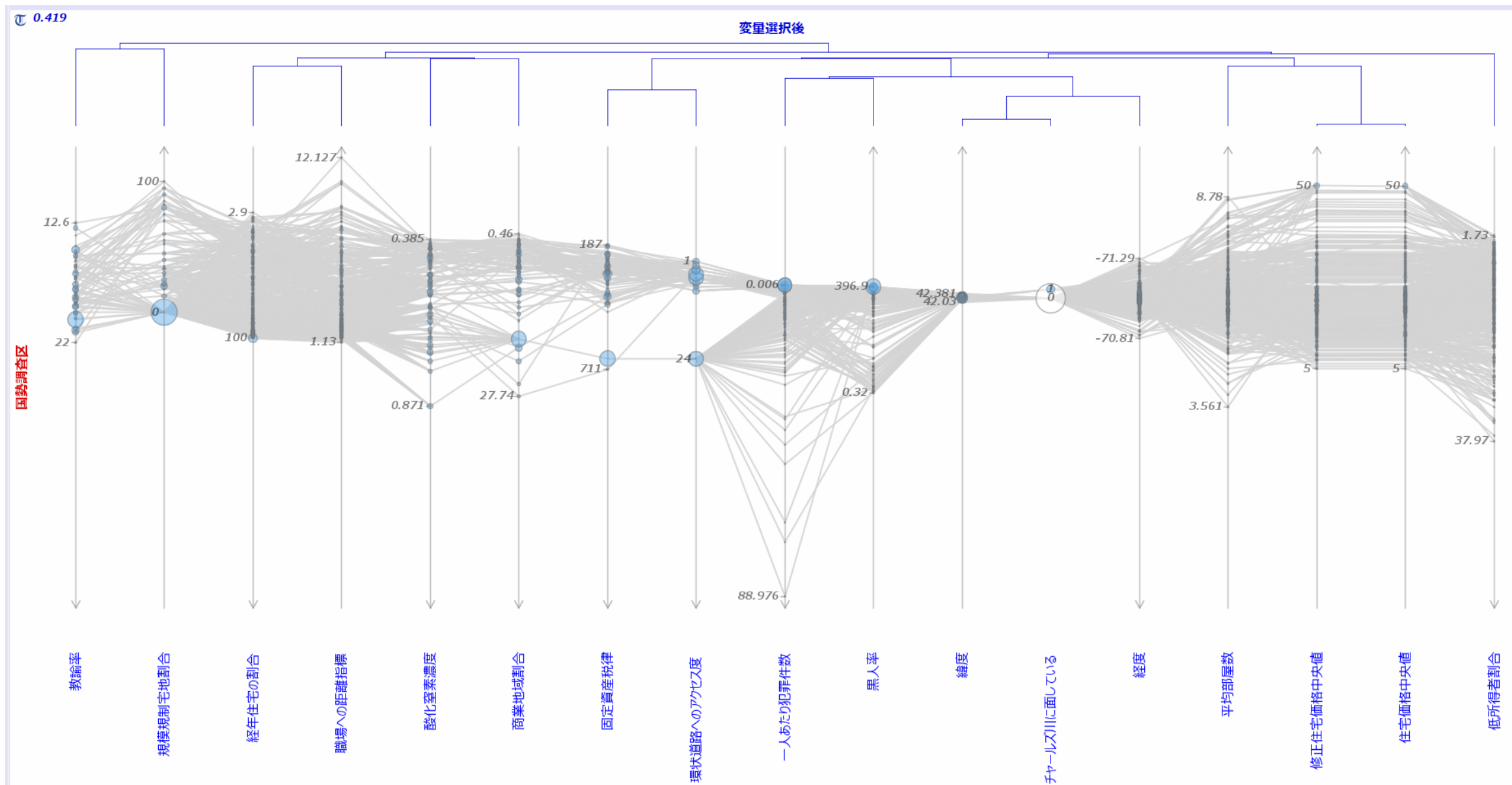
TRAD (<https://datascience.jp/TRAD.html>)

Boston データ（記録ID は自動的に隠されている）



- 水平線基準
（記録を示す線分がなるべく水平になるように各変量の位置, 尺度を調整）
- つぶれた変量は他の変量と直交

多変量データの変量間の関係（タウン名、タウン番号、国勢調査区を隠し、変量のクラスタリング）



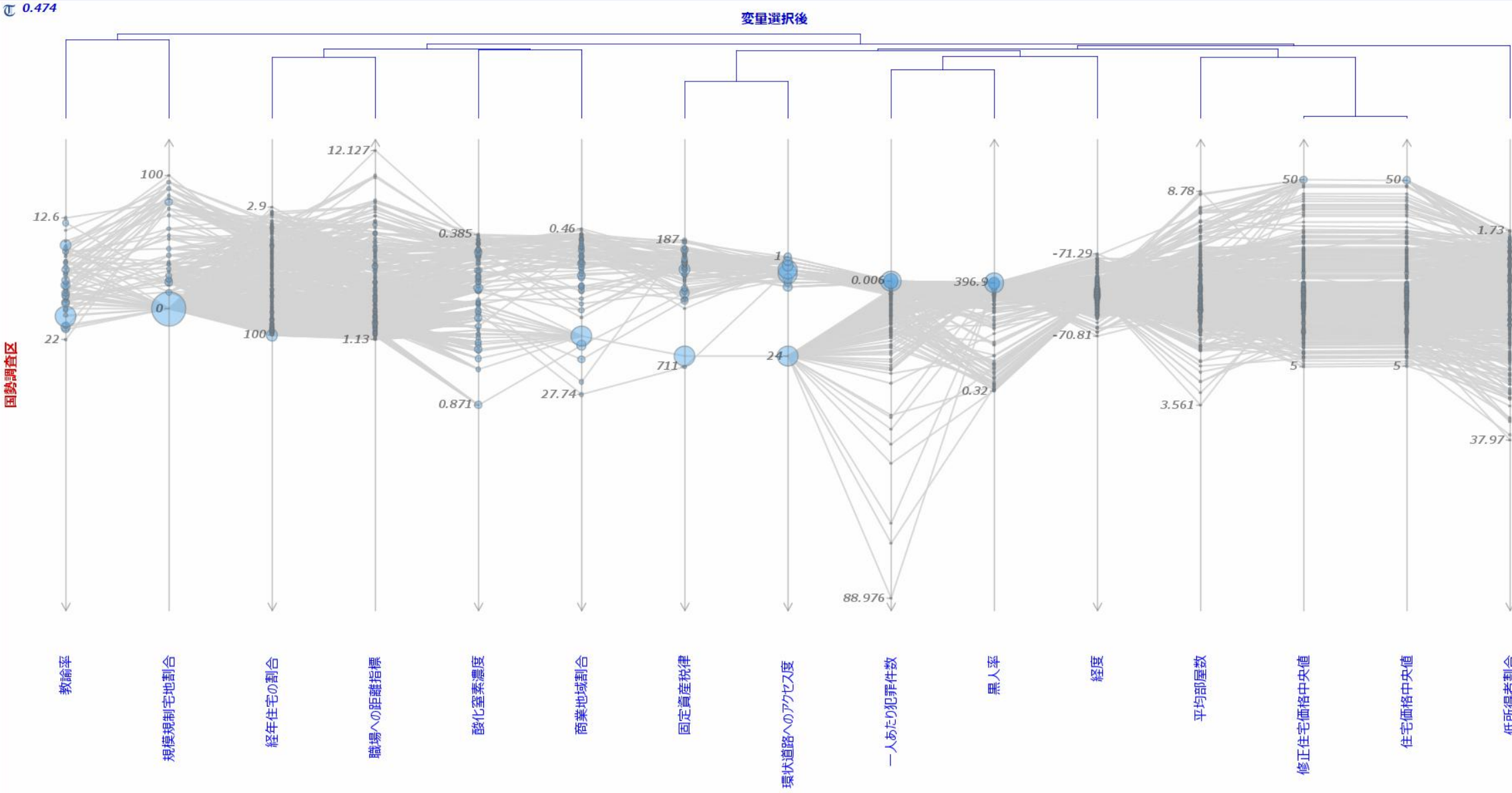
すぐわかること

- 緯度，チャールズ川に面しているかどうかは他の変量と直交(独立)している.
- 住宅価格と密接に関係しているのは，経度と平均部屋数のみ
 - 住宅価格が部屋数と密接に関連するのは当然
 - 住宅が安いか高いかを考察するなら，単位面積当たりの価格のほうが適當では？
 - 経度が関係し西側のほうが高いのは地理的な条件によるものであろう
- 黒人率と犯罪率は全体的には相反，ただし黒人率がかなり低い地域では犯罪率が低い.
- 残りの変量は国勢調査区(tract)のプロファイルで住宅価格とはほとんど関係ない
 - 商業地域割合が高いかどうかが特徴的な指標

このデータに対する**基本的な理解**

これなしに先へ進むと偏見に基づいた無責任な結果を導くことになりかねない

さらに、緯度、チャールズ川に面しているかを隠す



Boston データの歴史

- David Harrison Jr. and Daniel L Rubinfeld, Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management*, vol. 5, 81-102, 1978
 - Q: ボストン都市圏で、大気汚染が住宅への支払意思額をどれだけ左右するか？
 - A: 大気汚染がひどい、また、また所得が高いほど、支払意思額への影響は大きい (単純な回帰モデルOLSによる結論)
 - C: 大気汚染といっても Nox だけで、これは商業地域の割合つまり交通量が多いかどうかに関連し、直接住宅価格に影響している訳ではないであろう。また所得と言っても低所得者率で代用していて、全体的にはそれと住宅価格は逆相関。
- 数多くの学術論文 (次の Bivand で survey されている)
- Roger Bivand, Revisiting the Boston data set - Changing the units of observation affects estimated willingness to pay for clean air, REGION, vol.4-1, 2017
 - Q: Hariison 達は、NOX については国勢調査ではなく、TASSIM という基本的に交通量や風向などに基づくシミュレーション結果を利用している、その結果は国勢調査区単位ではなく独自の122区域のうち該当する 96 区域の値を 506の国勢調査区に割り当てている。これが問題ではないか
 - A: 国勢調査区を TASSIM の区域に合わせて合区したほうが、支払い意思額への OLS や SEM の適合度は上がる
 - C: すでに見てきたように NOX は交通量と地形に大きく依存し、住宅価格は基本的に西高東低。このことを踏まえた上で空間データ解析でないと本質をとらえているとは言えない。また支払い意思額を被説明変量にすること自体が問題。 **2 階から目薬。**

ちなみに、このBivand が元データの誤りを修正し R パッケージ spData 中で公開している。ここではこの中の boston.c を利用している。

3D TRAD

- 空間的な広がりを持つデータを**理解**するには，動的な 3D 表示が欠かせない.
- SEM, SAR, CAR, sCAR, nCAR など様々な空間データモデルが提案され使われているが，その妥当性の検証をするにも 3D 動的表示が欠かせない
- 近傍系を視覚的に設定するためにも必要
- 3次元空間といっても，物理的な3次元に限らず平面グラフ上の値として一般化しておく必要がある.
- これまでの TRAD の TextilePlot 表示との連携
 - カテゴリカルな値の数値化に利用できる
 - 型に応じた表示

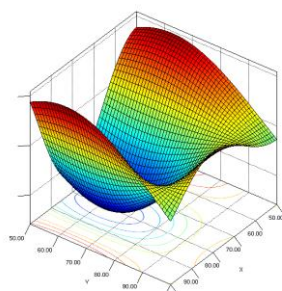
Javaから利用できる 3D animation package

- JGL (GPUへの直接アクセスなし、古い)
- OpenGL (GPUへの直接アクセス, C/C++)
 - JOGL
 - LWJGL, libGDX (low level のみ)
 - Java 3D (JOGL の上位だが, すでにサポートされていない)
 - Java3D (Oracle)
 - JavaFx (high level だが重い)
 - Jzy3d (<http://www.jzy3d.org>, low level から high level まで)

Red Book

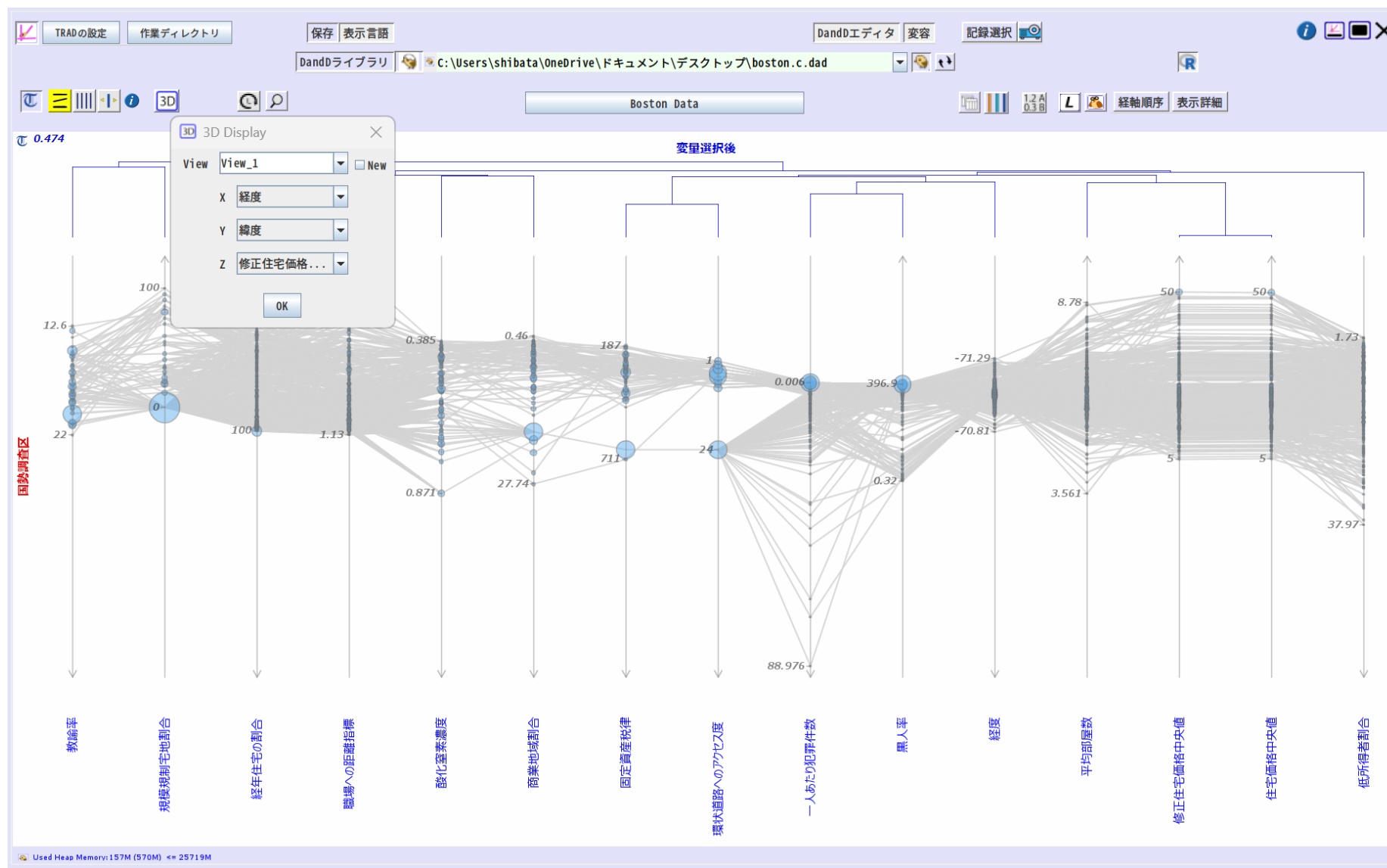
J. Kessenich et al., OpenGL
Programming Guide 9th edition,
Addison Wesley, 929 pages, 2017

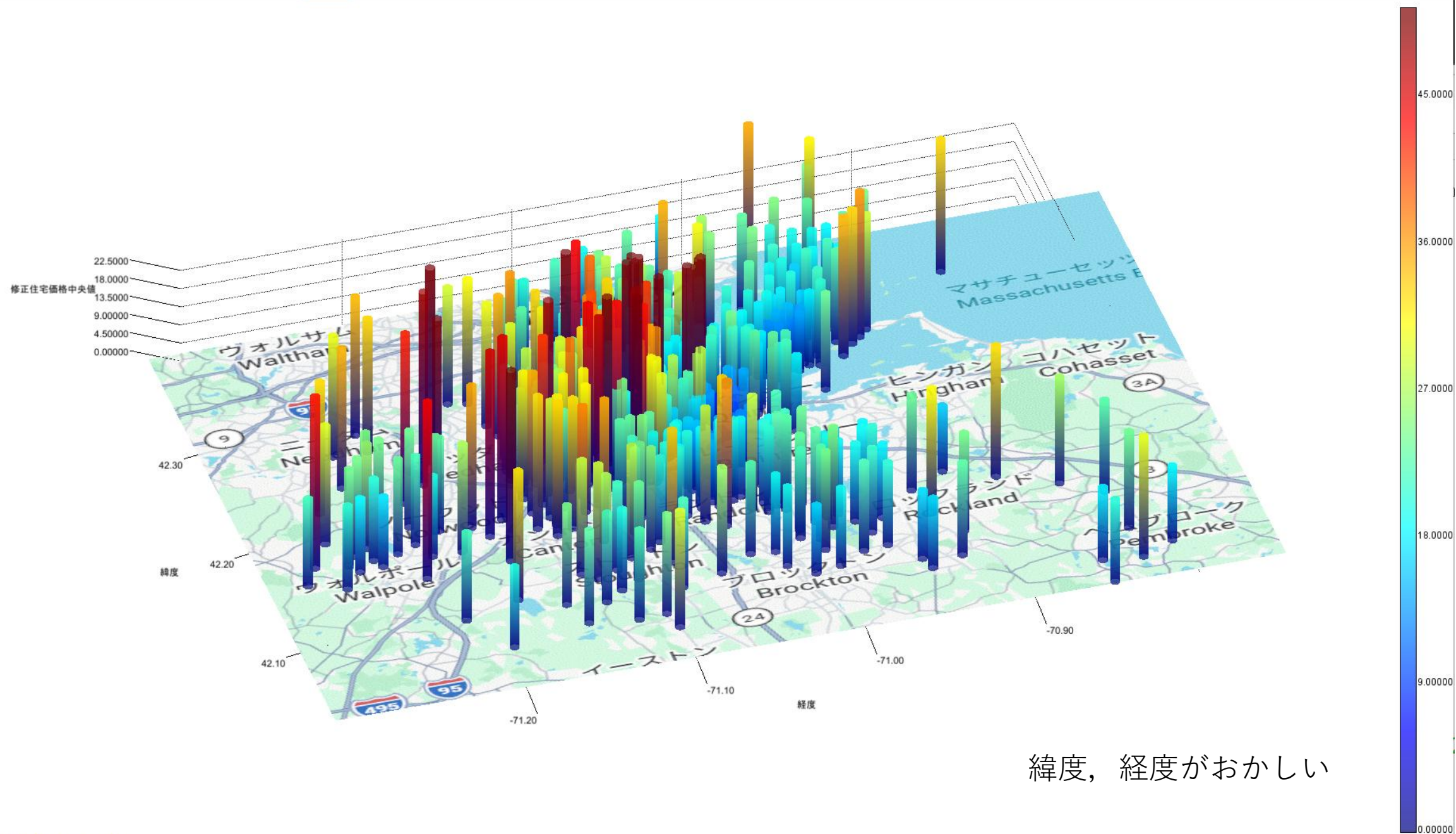
Scientific 3d plotting



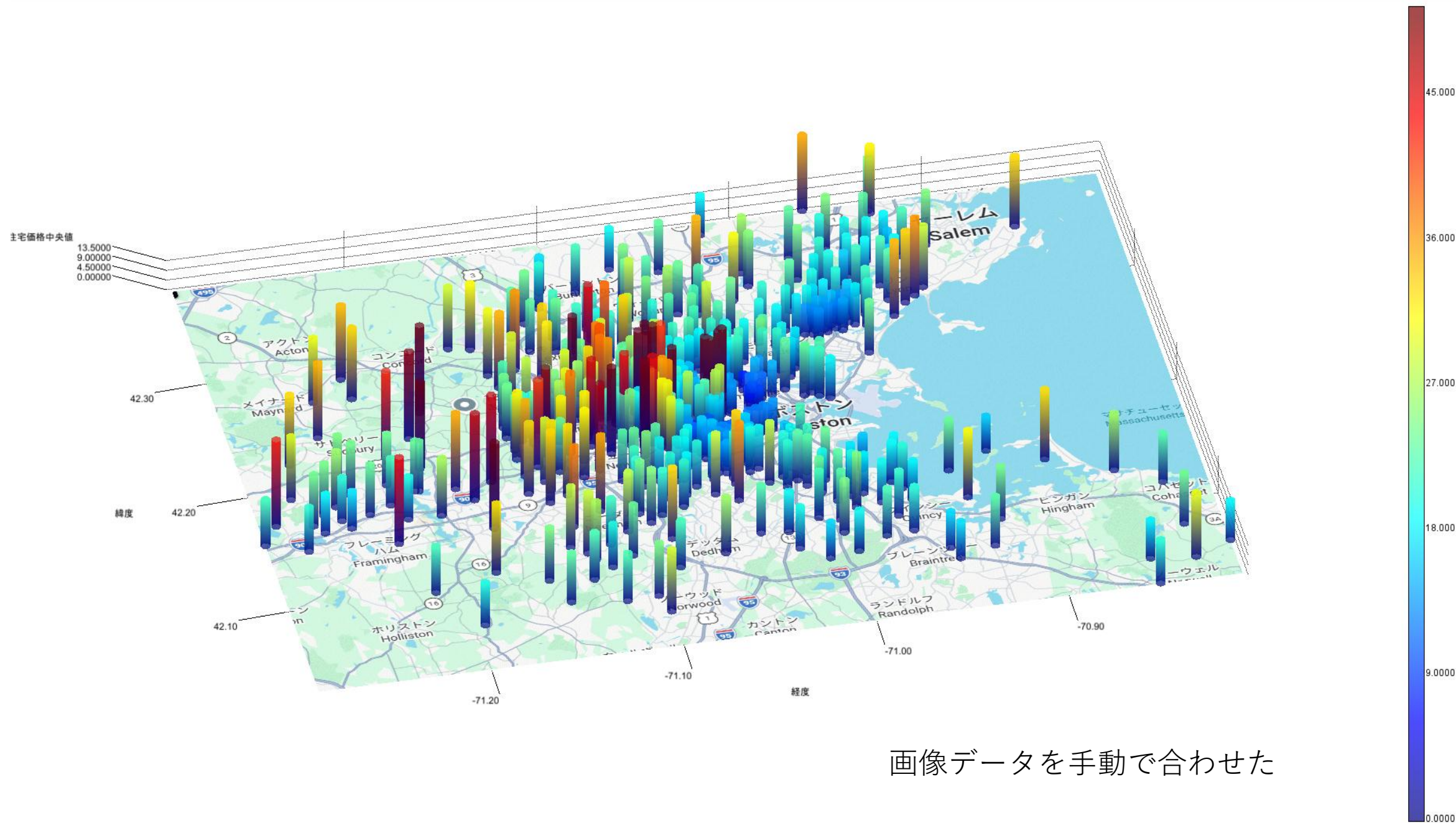
よくできたパッケージではあるが
数々のバグがあり, 詳細なマニ
ュアルもないので, ソースプログラ
ムを修正する必要がある

Bostonデータの 3D 表示

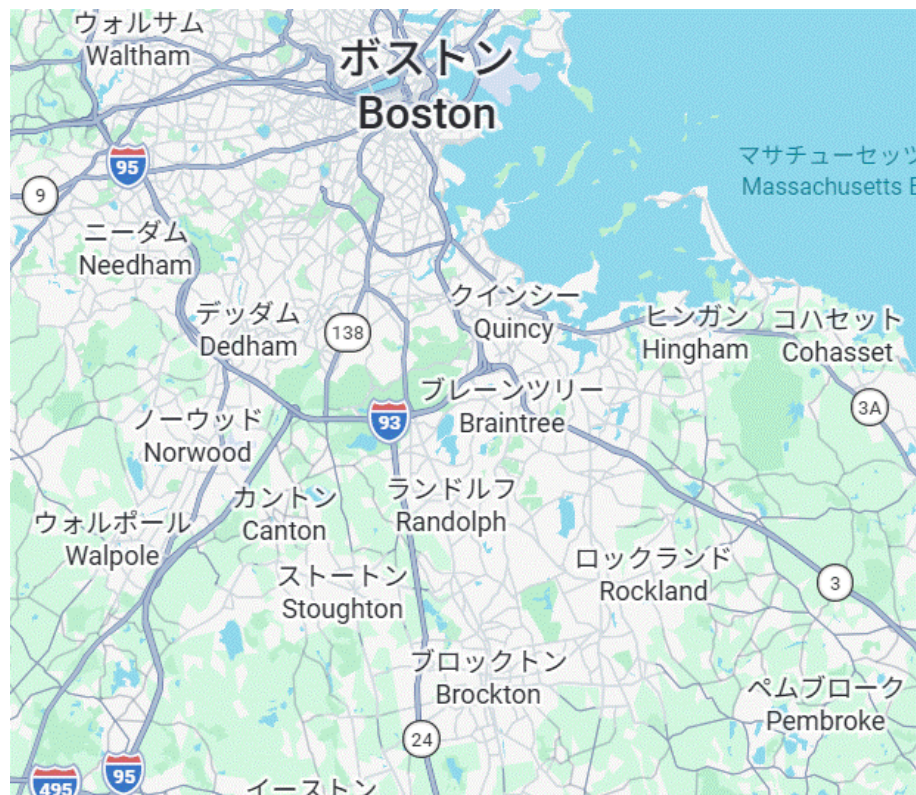




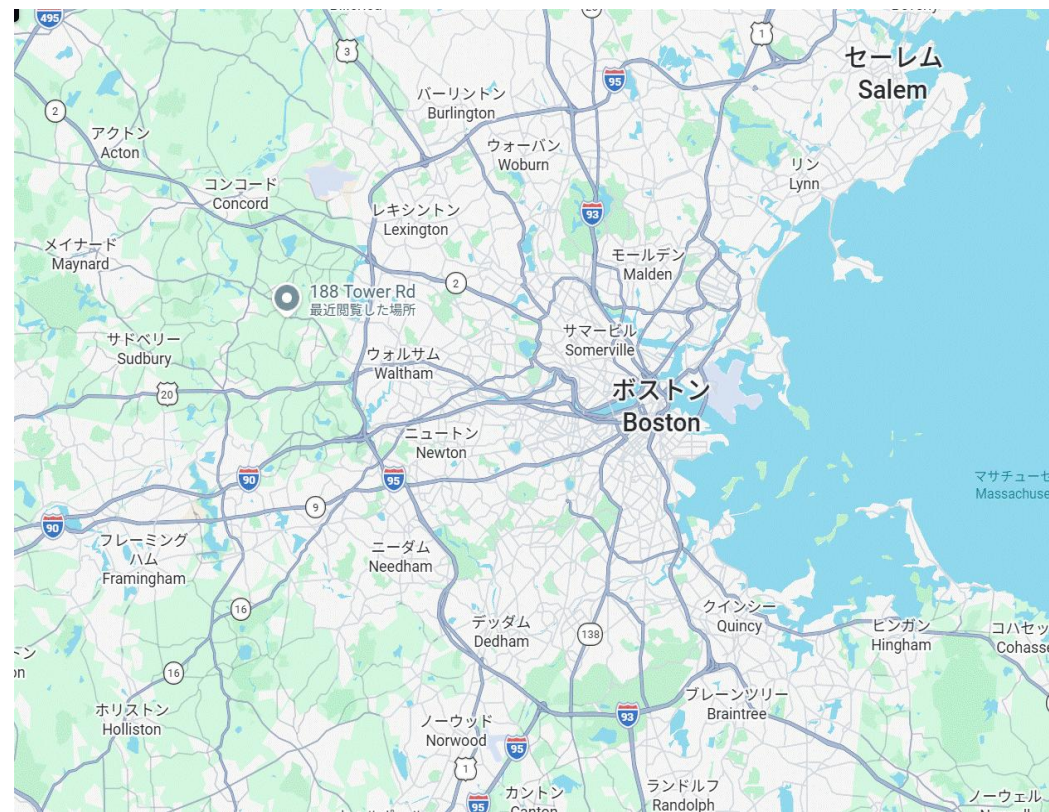
緯度，経度がおかしい



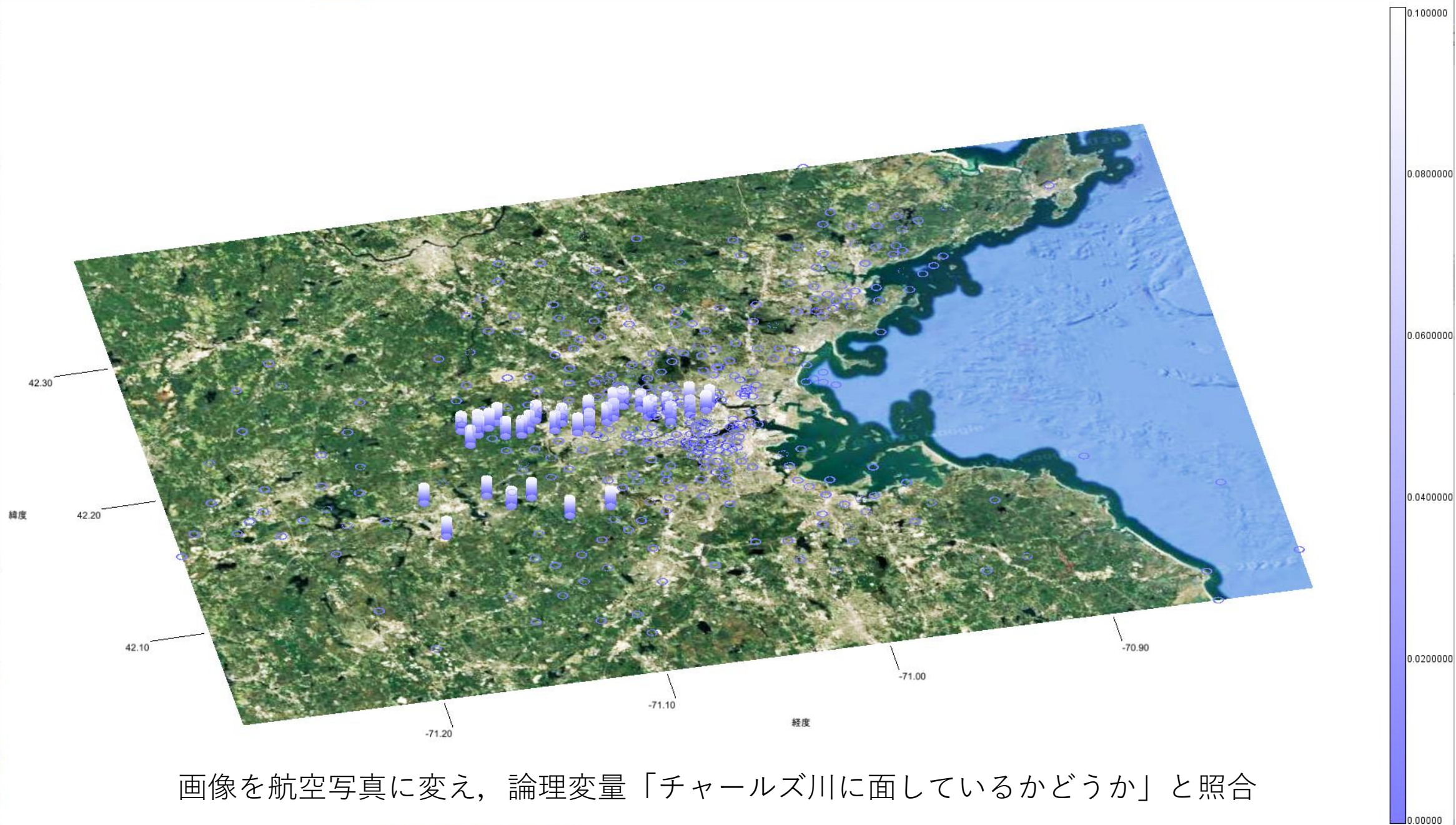
画像データを手動で合わせた



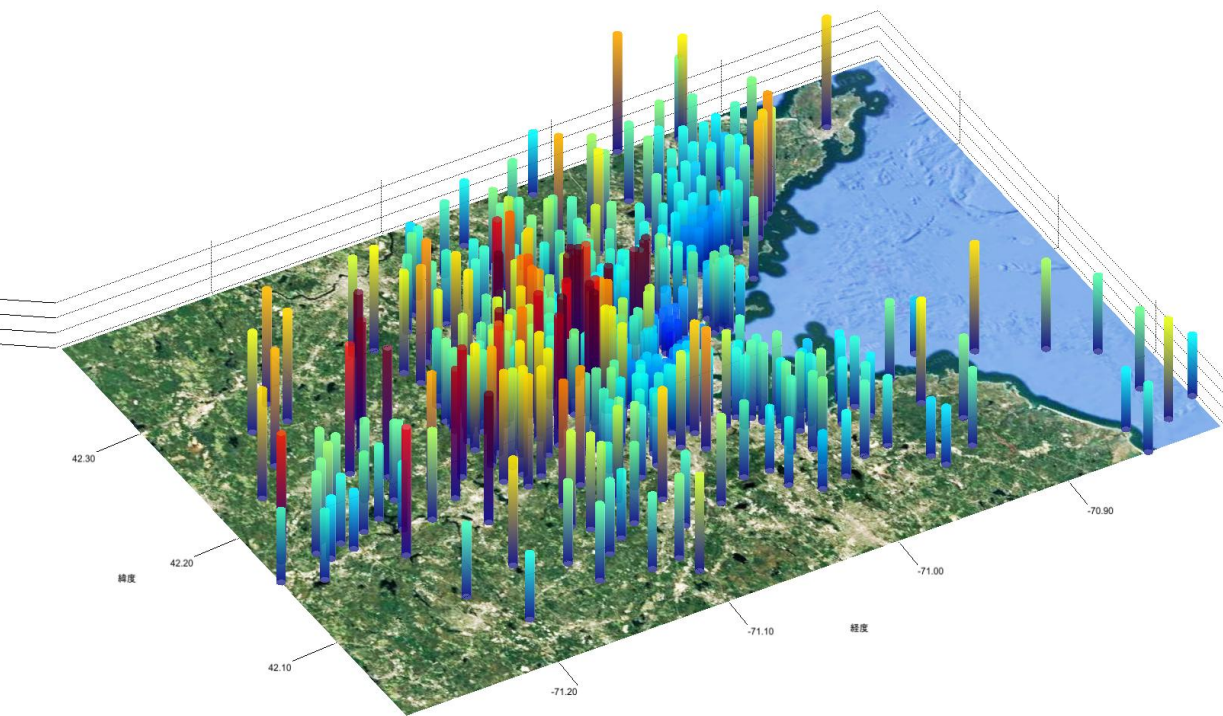
座標データの示す領域



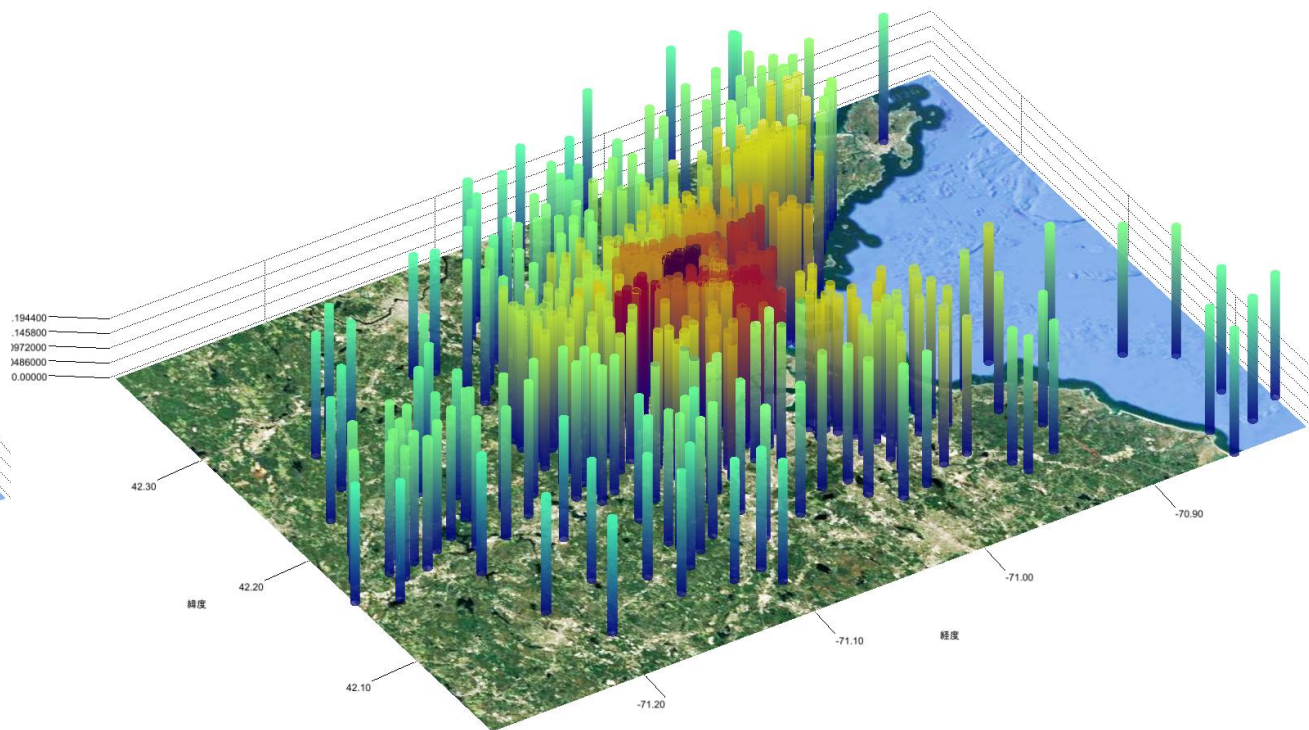
タウン名から手動で切り出した領域



画像を航空写真に変え、論理変量「チャールズ川に面しているかどうか」と照合



住宅価格



NO x

住宅価格もNOxも地形の影響を大きく受けている。
 ただし、中心部の海岸寄りや西部など、部分的にはNox
 が高く住宅価格が低いところも見受けられる。
 今後、地域を絞り nCAR をうまく使いこなすことでこのような
 特徴を抽出できると期待。

複数の3D 同
 時表示も可能

Boston Data で 3D TRAD を試してみたい方へ

- <https://datascience.jp/TRAD.html> から [TRAD20260908win.exe](#) のようなインストーラをダウンロードしクリックインストール
 - Mac(Intel x64, AArch64) 版は準備中
- URLパネルで
<https://datascience.jp/DandD/Examples/Boston.dad>
を選択. XY画像も含まれている.
- まだ 3D 部分はベータ版ですので, 随時, 左上隅の  を右クリックしてアップデートしてください.

今後の計画

- 型を反映した3D表示
 - 計量, 計数, 序数, 日時, 時間, マーク, 順マーク, 論理
- 明度, 彩度など表示の詳細の調整機能
- nCAR の実装と検証
 - 3D 表示を活かした近傍系の設定
- 3D 表示でのインタラクティブな理解支援
 - 表示範囲の変更
 - さまざまな属性の確認
 - 部分的な隠蔽
- 物理的な3Dではない「グラフ上の値の3D表示」
 - XY 平面にグラフを置く 3D 表示