データサイエンス実験 A

馬場教授第 2 回レポート

中央大学理工学部ビジネスデータサイエンス学科
23D7104001I 高木悠人

A. フィールドワークのデータを利用した分析

1. 利用データの説明

本実験では、本郷地区を調査エリアとした駐車場のポイントデータとその経路を記録したトラッキングデータを利用した。本データの収集は、2024 年 11 月 13 日に行われた授業内に約 1 時間散策し、歩いた経路上に見つけられたものを Geo Tracker で記録した。

2. 可視化手法の説明

当該データの可視化手法はヒートマップとした。ヒートマップは他の分析ツールと比べて専門的な知識を必要とせず、容易に特徴を可視化できるというメリットがある。特に、区画法と比較すると、計測範囲に対してプロット量の大小が極端に変化しない場合は適当な区画設定を行い難くヒートマップの方が適しているといえる。本データの収集範囲は約 500m 四方であり区画の設定距離が短くなってしまうため、ヒートマップを利用した可視化手法を選択した。

3. 収集したデータの可視化結果

(1) ポイントデータ

図 1 に収集した駐車場のポイントデータをヒートマップで可視化した。

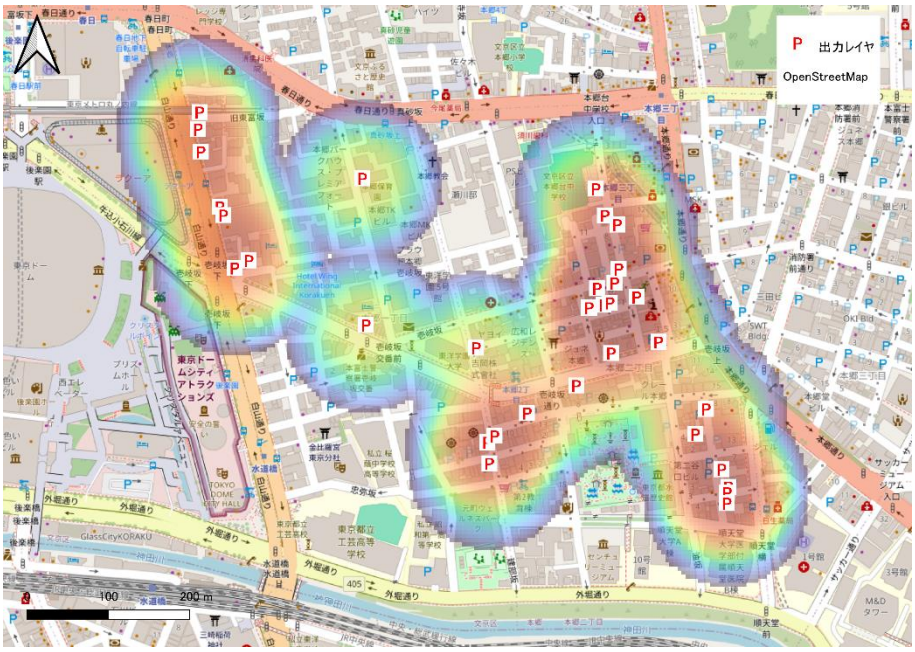

図 1 駐車場のピントデータのヒートマップ

上記よりわかることを箇条に構造化して示す。

1. 図の中央には分布が少なく大通り付近に多く分布している。
2. 白山通りよりも本郷通りの方が多く分布している。
3. ヒートマップを用いることでポイントデータの分布の特徴を容易に可視化できている。

(2) トラッキングデータ

図 2 に収集したトラッキングデータのポイントデータをヒートマップで可視化した。

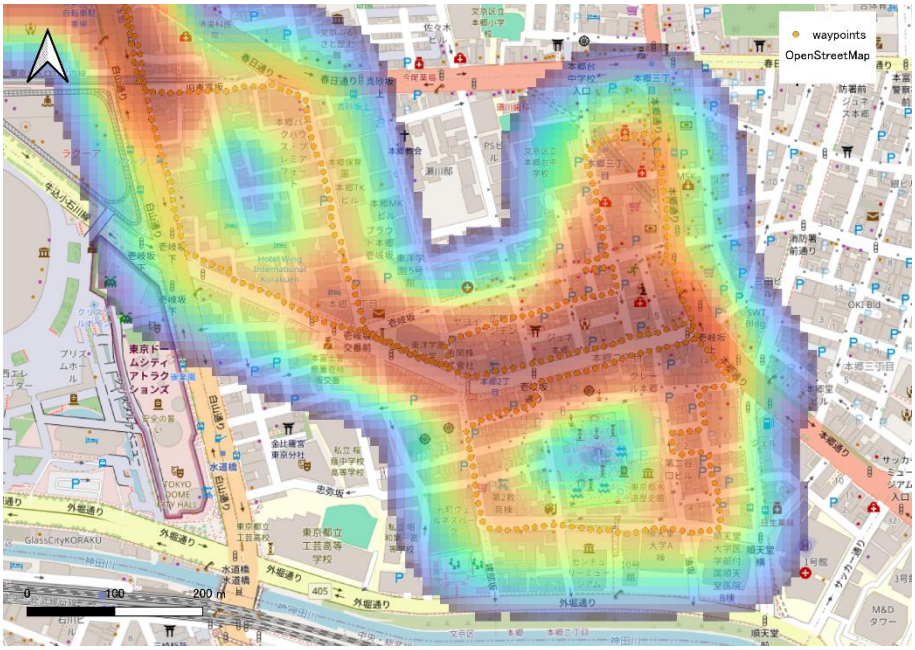

図 2 トラッキングデータのヒートマップ

上記よりわかることを箇条に構造化して示す。

1. 図中央付近に多く分布している。
2. 左下や右上には分布していない。
3. 色に差が生じていることからトラキング調査ルートに偏りがあることがわかる。

(3) 全班のトラッキングデータ

図 3 に全班のトラッキングデータのヒートマップを示す。

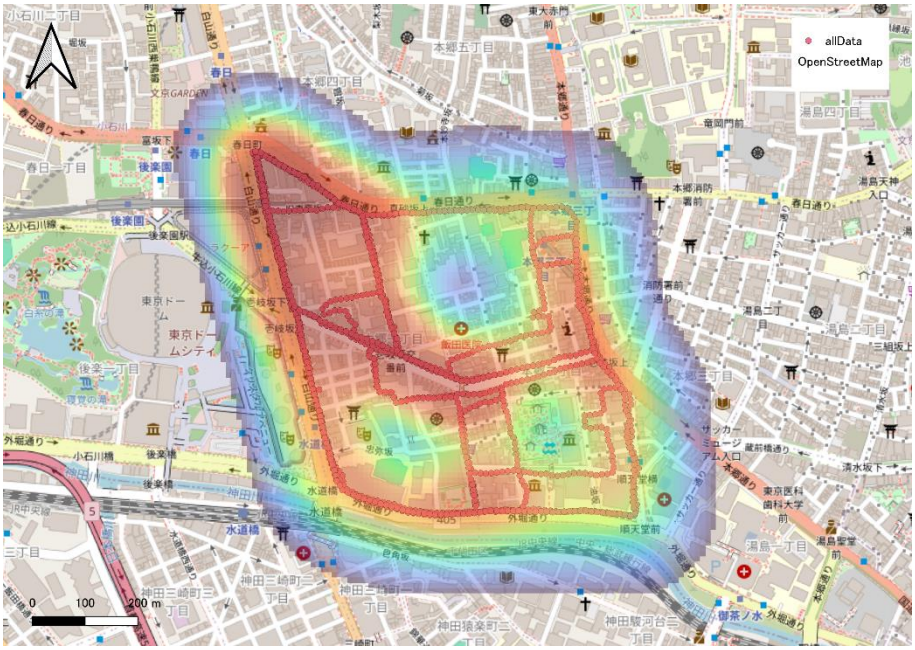

図 3 全班のトラッキングデータのヒートマップ

上記の図からわかることを箇条に構造化して示す。

1. スタート付近（左上）や中央部は色が暖色となっており密集していることがわかる。
2. 右上側が寒色となっていることから密度が小さく、調査されていない傾向にあるといえる。
3. 外周は常に暖色となっていることから、多くの班に調査されていることがわかる。

B. オープンデータを利用した分析

4. 統計解析手法の説明

本実験では平均最近隣距離を求めてから最近隣距離法で解析する。最近隣距離法とは、点分布が集中しているのか分散しているのかを知る手法である。まず、各点から最近隣の点までの距離の平均を平均最近隣距離(W)を以下の式から求める。

$$W = \frac{1}{n}\sum_{i=1}^{n} d_i$$

そして、その平均最近隣距離 W の期待値を以下のように近似して算出する。

$$E[W] \approx \frac{1}{2\sqrt{\frac{n}{S}}}$$

ただし、W を E[W]で割った値を 1 と比較して評価することも可能である。本解析では、上記のWの期待値を利用するこの時、Wが平均最近隣距離の期待値よりも十分に小さいとき点は集中しているといえ、逆に十分に大きい場合は分散しているといえる。一方、ほぼ等しいといえる場合はランダムな分布であると示せる。その後、統計的仮説検定を用いて、上記の評価の有意性を示す。仮説検定では、点分布が一様ランダムに分布すると仮定し近似的に以下の正規分布の式を利用する。

$$N\left(\frac{1}{2\sqrt{\frac{n}{S}}}, \frac{4-\pi}{\frac{4\pi n^2}{S}}\right)$$

この分布に基づいて W を標準化すれば、

$$Z = \frac{W - \frac{1}{2\sqrt{\frac{n}{S}}}}{\sqrt{\frac{4-\pi}{\frac{4\pi n^2}{S}}}}$$

となり、標準正規分布による検定が可能となった。そのため、点分布が一様ランダムに分布する場合Wは近位的に以下の正規分布に従うことを利用する。

$$N(0.5\sqrt{\frac{S}{n}}+\frac{0.051L}{n}+\frac{L}{n\sqrt{n}},\frac{0.070S}{n^2}+0.037\sqrt{\frac{S}{n^5}})$$

ただし、S は対象領域の面積、L は周長である。この分布に基づき W を標準化して標準正規分布による検定を行った。

5. 収集したオープンデータの説明

本項では、文京区内にある学習塾を調査対象として分析する。私は現在塾講師としており、塾の立地による顧客となる塾生徒数の影響は存在すると考えている。例えば、小中学生対象の塾である場合は治安等の問題から大通りに面した立地や駅付近など見通しの良い場所を選ぶ場合が多いと入会面談の際に耳にする。さらには、都内の塾の場合は、集合型難関校向け授業など広い地域から集まってくる場合も想定できる。したがって学習塾の立地には偏りがあるのか知りたいと考えたためオープンデータとして取得するに至った。本実験では、Google Maps で「文京区 学習塾」と検索し、その結果を QGIS で記録した。図 4 に収集した学習塾のポイントデータを示す。

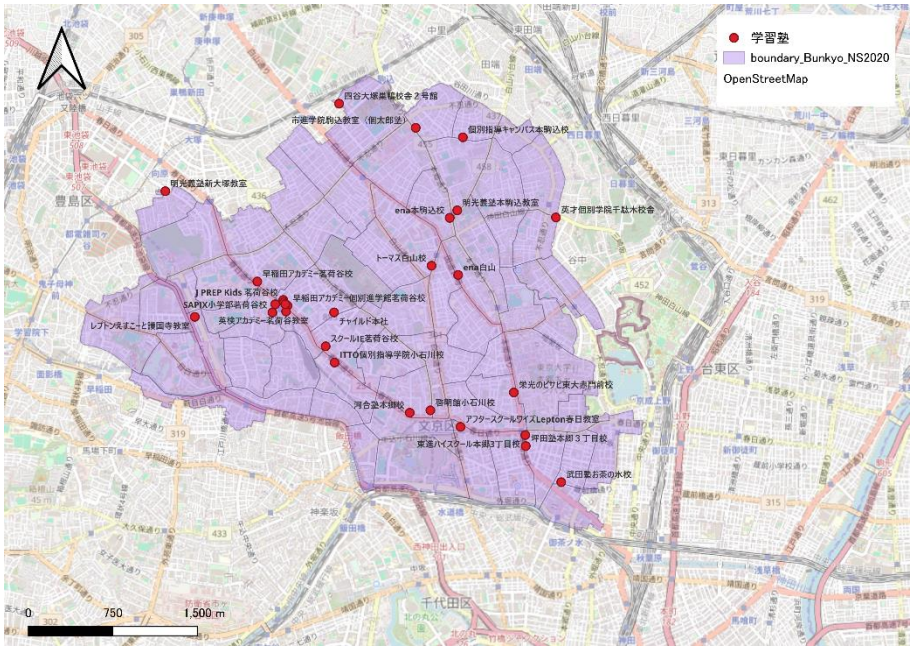

図 4 収集した学習塾のポイントデータ

上記よりわかることを箇条に構造化して示す。

(1) 図左側の茗荷谷駅付近に密集していることがわかる。

(2) 大通り沿いに分布していることがわかる。

6. 統計解析結果

本データをもとに距離行列を作成し解析を行った結果、以下のような数値が得られた。

$$\frac{W}{E(W)} = 2.780483$$

$$Z_2 = 1.658484$$

1 つ目の値は、平均最近隣距離を期待値で割ったものであるから 1 との離れ具合から密集か分散などの分布状態を知ることができる。今回の値は 2.78 となり 1 より離れているといえるため、分散しているといえる。そして Z_2 は絶対値が 1.96 よりも大きい場合有意といえるのだが、本解析では下回っているため有意と断定できない結果となった。

C. 考察とまとめ

今回、フィールドワークで収集したデータとオープンデータの可視化手法と解析手法について行った。初めに、可視化手法について考察する。

フィールドワークによって収集したデータをヒートマップで可視化したところ、ポイントデータのプロット密度を色で示すことができた。そのため、ヒートマップを利用しないプロットのみの場合に比べてその傾向を把握しやすいことが分かった。特に、大規模な分析の場合プロットのみの場合理解しにくいが、ヒートマップや区画法を用いることで人間が認識しやすい形式に変換できる。そして、分析と理解が容易であるためハザードマップなどのように、統計的な知識を有していない大衆向けの資料などにも有用であると考察できた。

次にオープンデータの解析について考察する。本解析では文京区内の学習塾の地理データを収集した。その結果をプロットしたところ、大通り沿いや、特に茗荷谷駅付近に集中していることが分かった。それを踏まえて統計的解析手法および仮説検定を行った結果、有意であるとは断定できないが分散しているという結果を得られた。本来、先に述べた駅付近の密集が真であるならば、平均最近隣距離が期待値よりも小さくなるのが妥当である。特にプロットによる可視化により茗荷谷駅付近に集まっていることは解釈可能であるため、本解析では、標本数が少なかったことによる対立仮設の採択となったのだと考えた。よって、今後の展望としては、塾の定義を定め文京区という広範囲地域ではなく文京区西部のように抽出し厳正に調べることが必要であると考えた。そして、先の収集データの説明の項でも述べた通り学習塾の分布は大通りに面していると推測したため、大通りと学習塾の距離を求めその有意性についての解析も研究材料になりうると考えた。

D. 参考文献

（ア）データサイエンス実験 A, 中央大学理工学部ビジネスデータサイエンス実験 A,2024.

（イ）中原宏,太田賓,北海道大学建築工学科教授,
「市街地における施設用途の立地連関に関する考察」,
（https://www.jstage.jst.go.jp/article/journalcpij/15/0/15_355/_pdf#:~:text=%E6%9C%80%E8%BF%91%E9%9A%A3%E8%B7%9D%E9%9B%A2%E6%B3%95%E3%81%A8,%E8%BF%91%E9%9A%A3%E8%B7%9D%E9%9B%A2)%E3%82%92%E6%B8%AC%E5%AE%9A%E3%81%99%E3%82%8B%E3%80%82）,
2024/11/25 参照

E. 付録

本解析で用いたRコードを以下に示す。

```
#最近隣距離法
d001 = read.csv("学習塾データ 2024.csv" ,header=T) #データの読み込み
d002 = d001[,-1] #文字列になっている列を除外
n = length(d002)
dist = NULL
for(i in 1:n){ #各行について 0 より大きい最小距離を抽出
  a = min(d002[i,d002[i,]>0])
  dist = rbind(dist,a)
}
W = (1/length(d002))*sum(dist) #平均最近隣距離
#文京区の周長は約 18,900m
#文京区の面積は約 1,127,300m2
L = 18900
S = 1127300
E_W = 1/(2*sqrt(length(d002)/S))
w = W/E_W #点が集中または分散しているかの指標
w
#仮説検定
a = 0.5*sqrt(S/n)+0.051*(L/n)+L/(n*sqrt(n))
a
b = 0.070*(S/n^2)+0.037*sqrt(S/n^5)
b
Z_2 = (W-a)/sqrt(b)
Z_2
#Z_2 は標準正規分布に近似的に従うため、 #値が-1.96 より小さいまたは 1.96 よりも大きければ #有意水準 5%で統計的に有意である
```