地理情報科学実験　2回目：地理情報システムを用いた空間分析

鈴木　朝陽

2024 年 12 月 11 日

1.　本課題の目的

地理情報システムを用いることで目的に応じた空間データの解析を行うことが出来る。特に、複数のデータを重ね合わせる操作、空間的に欠損している箇所を補間などが容易に行える。本実験では、履修者が収集した空間ポイントデータおよびトラッキングデータを利用して、その空間的傾向を統計的に明らかにする。履修者間のデータを共有することで、異なるデータ間の関係性についても考察する。

本実験の目的は、収集した空間データを利用して、空間重ね合わせやデータ間の空間的距離や統計的仮設検定を用いることで、空間データの集積関係について考察することである。併せて、地理情報システム、具体的には QGIS を用いて空間データの初歩的な解析が出来るように学習を進める。必要に応じて外部データを用いることで、様々な空間データを扱えるようになることを目指す。

2.　フィールドワークのデータを利用した分析

2．1　利用データの説明

利用したフィールドワークのデータは、主に 3 つあった。1 つ目は、文京区本郷 1、2 丁目の外周にある飲食店のポイントデータとトラッキングデータだ。自分の班の飲食店のデータを図 1 に示した。

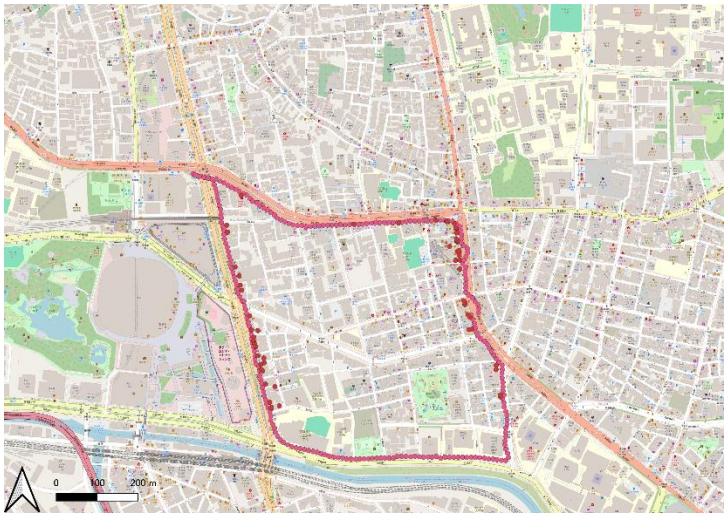

図 1　飲食店のデータ

図 1 から飲食店のポイントデータは外周の東側と西側でかなりの偏りが見られた。

2 つ目は、かなり似た計画でフィールドワークを行っていたため、5 班の自販機のポイントデータも利用した。自販機のポイントデータを図 2 に示した。

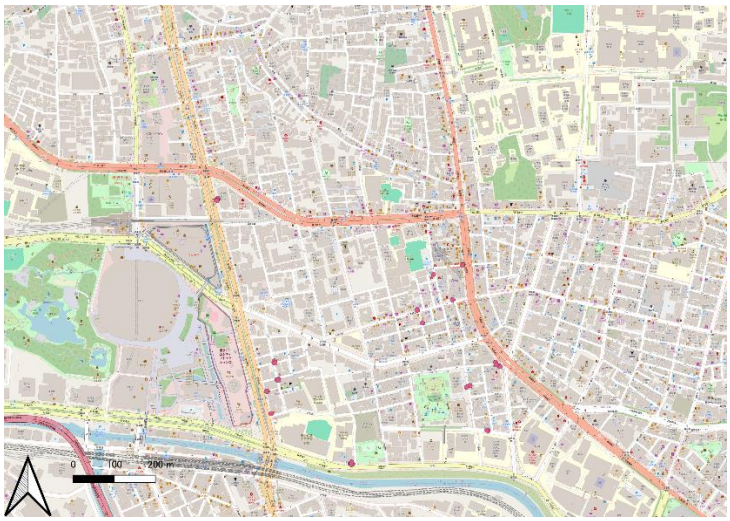

図２　自販機のポイントデータ

図２より、自販機のポイントデータも同様に東側と西側で偏りが見られた。

３つ目は、自班との比較を行うため、全班のトラッキングデータを用いた。全班のトラッキングデータを図３に示した。

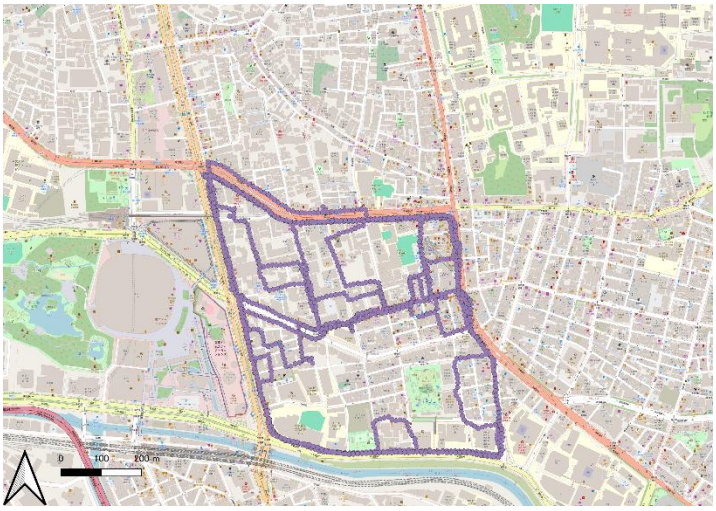

図３　全班のトラッキングデータ

図３より、南側に比べて北側にフィールドワークが集中していることが分かった。

２．２　可視手法の説明

本実験で用いた可視手法は２つであった。

１つ目はヒートマップであった。ヒートマップはポイントデータの相対密度を可視化密度の計算にはカーネル密度関数が用いられ、各点のカーネル密度を足し合わせることでポイントデータの空間的密度を表現することが出来る。ヒートマップは区画法によるカウントに比べ連続的に点の存在の分布を把握することが出来、直観的に理解がしやすいという利点がある。

２つ目は標準偏差距離であった。標準偏差距離は、点分布の重心から標準偏差分の半径の円を描画することが出来る。空間上の統計量を算出し、定量的な議論が出来る。

２．３　収集データの可視化結果

まず初めに飲食店のポイントデータに関する可視化結果を示した。飲食店のポイントデータのヒートマップを図４に、標準偏差距離を図５に示した。

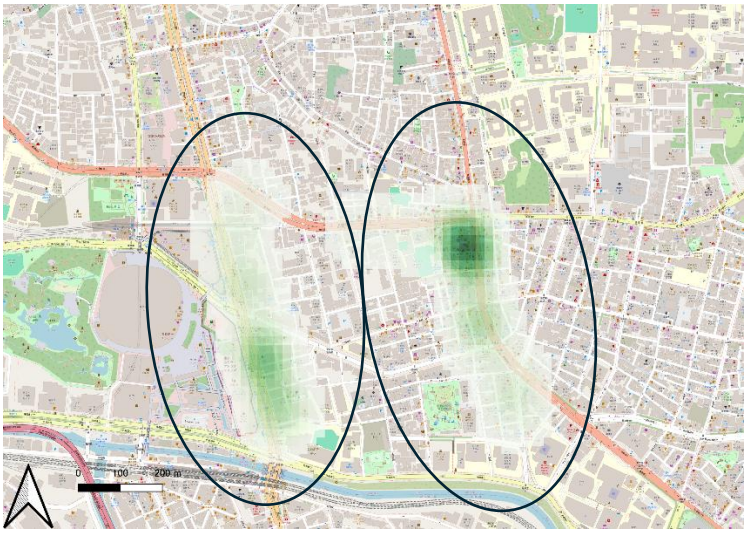

図４　飲食店のポイントデータのヒートマップ

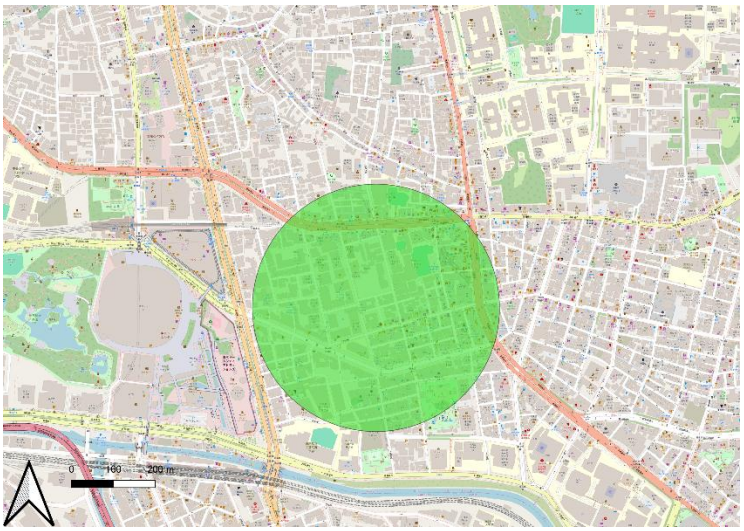

図５　飲食店のポイントデータの標準偏差距離

図４、５より飲食店のポイントデータは、西側と東側にデータが密集していることが分かった。西側は南に行くほどデータが密集し、東側は北に行くほどデータが密集していることが分かり、東側は西側よりデータが密集していることが分かった。また分散は大きいことが分かった。

次に自販機のポイントデータに関する可視化結果を示した。自販機のポイントデータのヒートマップを図６に、標準偏差距離を図７に示した。

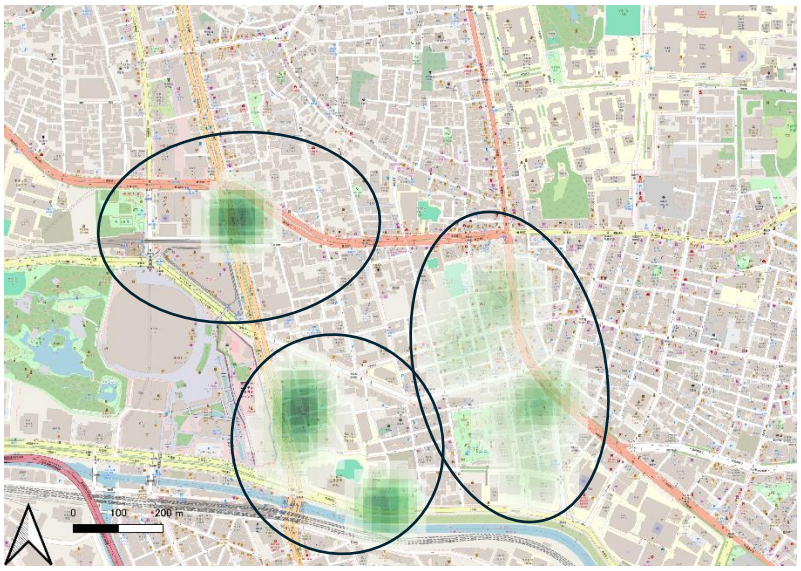

図６　自販機のポイントデータのヒートマップ

図７　自販機のポイントデータの標準偏差距離（黄色）

図６、７より自販機のポイントデータは、北西側と南西側と東側の３つのグループに分かれて密集していることが分かった。図４と異なり、自販機は西側のほうが東側よりデータが密集していることが分かった。さらに図５と比べると、自販機のほうがより分散が大きいことが分かった。

続いて飲食店のトラッキングデータに関する可視化結果を示した。飲食店のトラッキングデータのヒートマップを図８に、標準偏差距離を図９に示した。

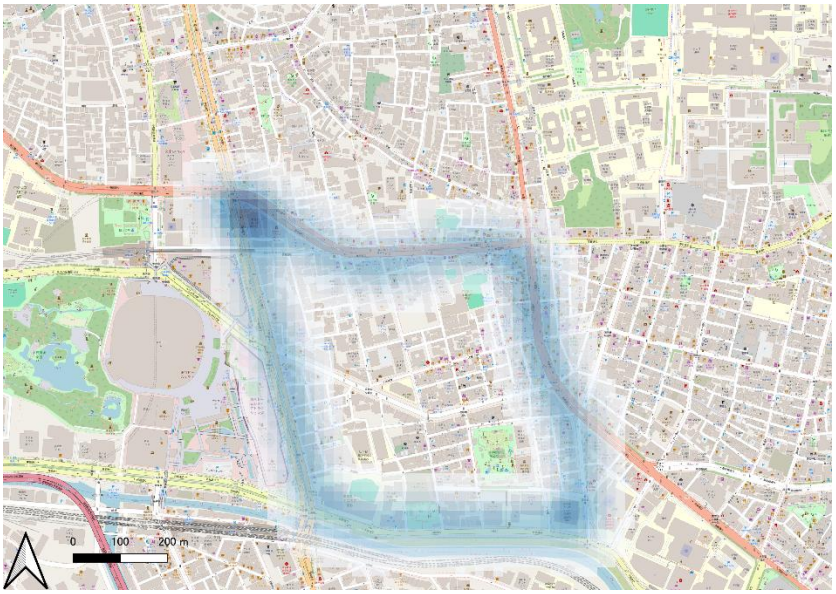

図８　飲食店のトラッキングデータのヒートマップ

図９　飲食店のトラッキングデータの標準偏差距離

図８、９より飲食店のトラッキングデータは偏りが少なく、まんべんなくフィールドワークをすることが出来ていた。また、分散が大きいことが分かった。

さらに全班のトラッキングデータに関する可視化結果を示した。全班のトラッキングデータのヒートマップを図１０に、標準偏差距離を図１１に示した。

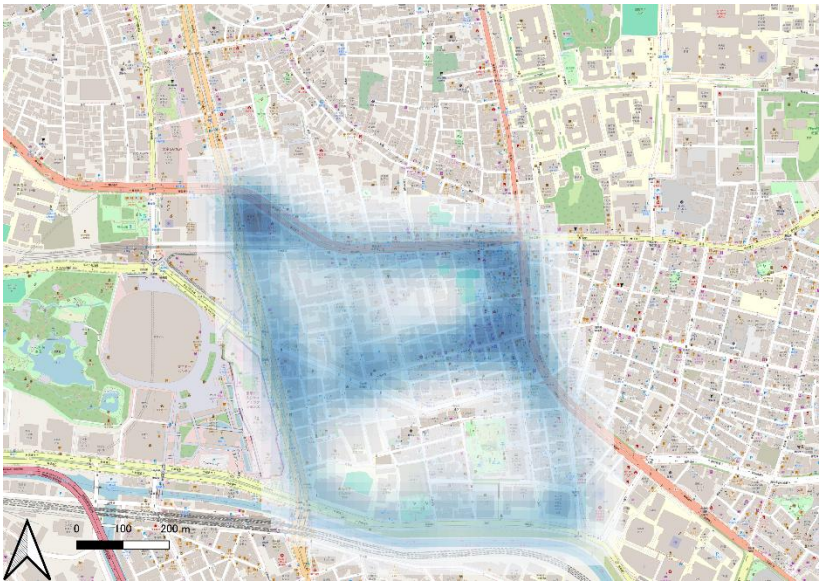

図１０　全班のトラッキングデータのヒートマップ

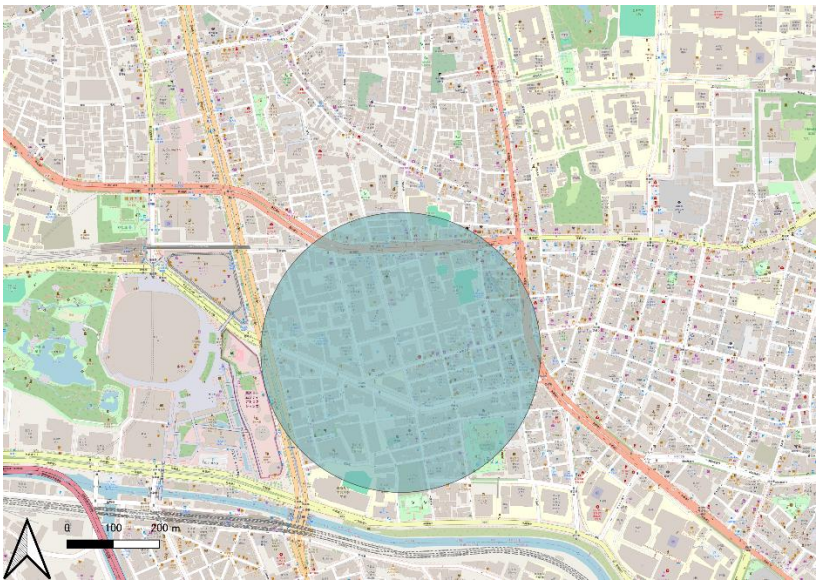

図１１　全班のトラッキングデータの標準偏差距離

図１０、１１より全班のトラッキングデータは、北側に密集しており、南側には密集が少ないことが分かった。また、図９と比べると分散が小さいことが分かった。

３．オープンデータを利用した分析

３．１　統計解析手法の説明

点分布が集中しているのか、分散しているのかを知るため、各点の際近隣の点までの距離の平均を求める平均最近隣距離法を統計解析の手法として用いた。平均最近隣距離 W は、点 i から際近隣の点までの距離をd_i、点の個数を n として以下で求められた。

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

点が面積 S の平面上でランダムに分布していると仮定すると、この時、平均最近隣距離 W の期待値は以下で求められた。

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

点が集中または分散しているかの指標 w は以下で求められた。

$$w = \frac{W}{E[W]}$$

w≪ 1：点は集中
w≒ 1：点はランダム
w≫ 1：点は分散

本実験では点の個数が多くない場合を採用した。点分布が一様ランダムに分布する場合、対象領域の面積を S、周長を L として、W は近似的に以下の正規分布に従うことを利用した。

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

この分布に基づき W を標準化して、標準正規分布による Z 検定が可能となった。このときの帰無仮説 H_0を以下で設定した。

帰無仮説H_0*: 点はランダムに分布している。*

3．2　収集したオープンデータの説明

収集したオープンデータは、文京区全域におけるイタリアンの店舗であった。イタリアンのポイントデータを図１２に示した。

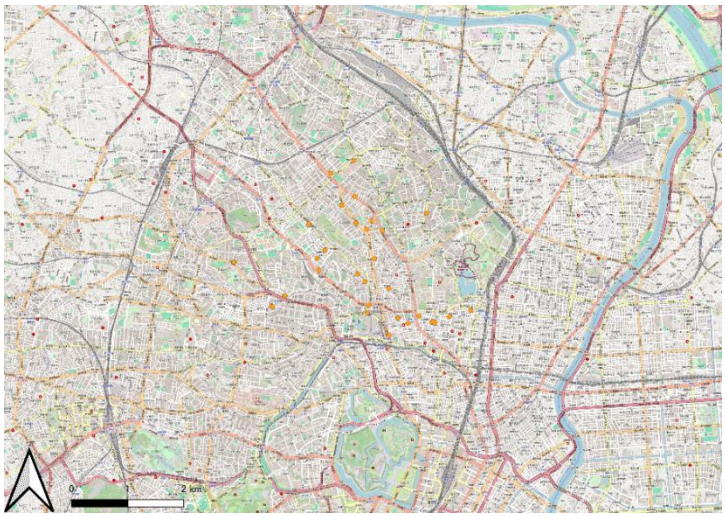

図１２　イタリアンのポイントデータ

図１２より、文京区全域におけるイタリアンの店舗数は３２店舗であった。

イタリアンのポイントデータのヒートマップを図１３、標準偏差距離を図１４に示した。

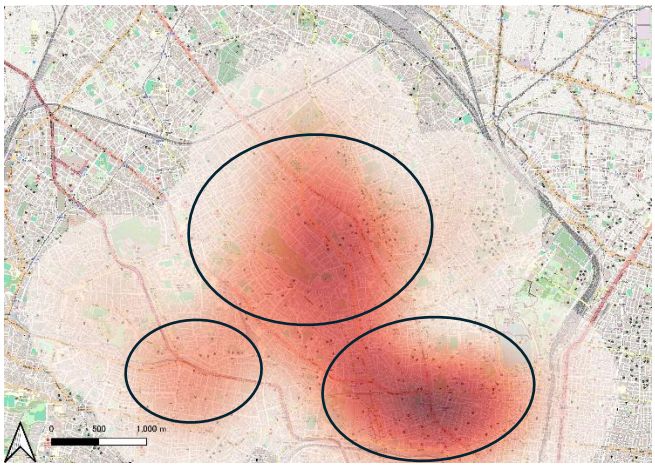

図１３　イタリアンのポイントデータのヒートマップ

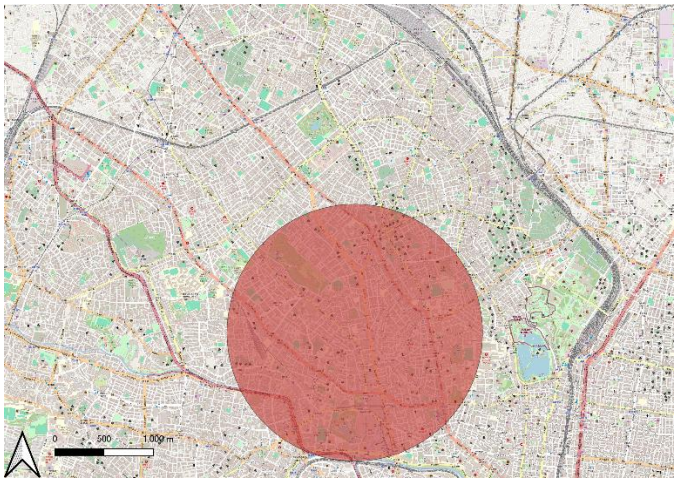

図１４　イタリアンのポイントデータの標準偏差距離

図１３、１４よりイタリアンのポイントデータは、東南、東北に密集していることが分かった。また分散は大きかった。

３．３ 統計解析結果

以下に R のコードを添付した。

[R コード]

```
#最近隣距離法
d001 = read.csv("C:/朝陽/朝陽データサイエンス実験 A/地理情報 2 /distance2.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))
E_W
w = W/E_W #点が集中または分散しているかの指標
w

#最近隣距離法による仮説検定
#点が少ない場合
a = 0.5*sqrt(S/n)+0.051*(L/n)+L/(n*sqrt(n))
b = 0.070*(S/n^2)+0.037*sqrt(S/n^5)
Z_2 = (W-a)/sqrt(b)
#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
Z_2
```

以上のコードを R で実行した結果、平均最近隣距離 W は約 274.9 で、平均最近隣距離の

期待値 E[W]は 95.35、点が集中または分散しているかの指標 w は 2.883 であった。したがって、W>>E[W]であり w>>1 より点は分散していることが分かった。Z_2≒4.299>1.96 であったため、帰無仮説 H_0は棄却され、有意水準 5 %において有意であると言えた。すなわち、点はランダムに分布していなさそうであった。

4. 考察とまとめ

2.3 の収集データの可視化結果から、飲食店と自販機には似た分布が見受けられた。飲食店も自販機も駅を意識した配置であった。特に自販機は水道橋駅と本郷三丁目駅に加えて春日駅も視野に入れている様子が見受けられた。しかし、西側より東側の本郷三丁目駅のほうが飲食店の店舗数が多かったことに対し、自販機は東側より若干西側、特に水道橋駅付近における台数が多いことが分かった。これは飲食店と自販機が互いに需要と供給を補完し合う関係にあるがゆえの配置ではないかと考察した。飲食店の店舗数が少ないエリアに自販機を設置することで、飲料をどのエリアにおいても補完することが出来る。

また、自班のトラッキングデータと全班のトラッキングデータを比較すると、どちらも本郷三丁目駅あたりにデータが密集していることが分かり、どの班にとっても観測したい事物が多くあったことがうかがえた。全班のトラッキングデータから南側の密集が少ないことが分かり、全体を通して南側のデータが不足していると考察した。

3.3 より統計解析結果から、イタリアンのポイントデータは分散が大きく、ランダムに分布していない可能性が高い、すなわち配置に意味がありそうなことがわかった。さらに 3.2 の収集したオープンデータの説明から、イタリアンのポイントデータは特に東南、東北に密集していることが分かった。またこの密集は大通りに付随していると考察した。国道を青色に、その他道の幅が広い一般道を水色に塗り、大通り沿いにあった店舗を黄色のポイントでプロットした図を図１５に示した。

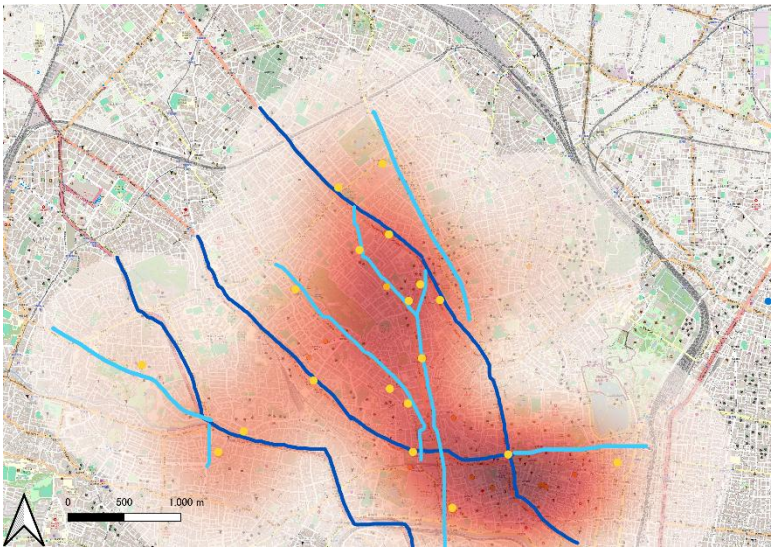

図１５　大通りに付随しているイタリアンのポイントデータ

図１５より、イタリアンは３２店舗中２０店舗が大通りに付随していた。その中で国道に付随していたイタリアンは7店舗であり、その他大通りに付随していたイタリアンは１３店舗であった。したがって、イタリアンは大通りに付随している可能性が高く、特に一般道に付随している可能性が高いのではないかと考察した。

参考文献

[1] 中央大学理工学部ビジネスデータサイエンス学科、「データサイエンス実験 A」、p.49~51（2024）.

[2] スライド「DS 実験 A：実験関連資料 2 回目」