商圈(Retail Trade Area)引力模型分析

# 简介

# 相关工作

商圈吸引力模型

人群区域划分

基于地理的可视分析

统计分析和机器学期

# 商圈理论与数据

## 零售商圈引力模型

零售商圈是零售交易区域的辐射范围，但是商圈的概念并没有很明确的定义，本文中，我们认为零售商圈的商业企业聚集所形成的空间范围。同样，在商圈级别划分标准下，本文中所研究的商圈是指由核心商业圈和次级商业圈组成的空间范围。

商圈理论中应用最广的是Reilly Rule 和Huff Model 以及其的演化模型。

由于大多数情况下，企业很难获取详细的商业信息，那么如何选择投资地成为了一个难题，而Reilly Rule 最早为企业提供了容易实现的决策指导。Reilly Rule认为商业也具有相互吸引的特性，它以万有引力定律为核心，来确定商圈吸引力临界范围。但是Reilly Rule是以商圈为中心的研究，并且需要有较严格的前提才能使结果有效。在我们对上海十九个大型商圈进行研究之后，我们发现使用Reilly Rule 是很难确定商圈范围的，由于商圈差异和人群行为等因素。

与Reilly Rule 有所不同，Huff Model 是从顾客的角度进行研究，通过模型计算出一个概率值，这个概率值能够代表当前用户去往某个商圈的概率，但是模型计算概率的时候用到的阻力和魅力因素仅仅包含距离和商店规模。在大都市商圈引力研究中，这两个因素依旧很重要，但是其他因素，例如地理位置，商品档次等，对概率的计算同样占据很重要的位置。

除此之外，还有很多演化模型，我们在5.1中进行了详细的对比研究。

## Reilly Rule and Huff Model

Reilly Rule 是W. J. Reilly 最早在1931年提出，它的核心观点是：具有零售中心地机能的两个都市，对位于其中间的一个都市或城镇的零售交易的吸引力与两都市的人口成正比，与两都市与中间地都市或城镇的距离成反比（查原句）。模型(康帕斯—Reilly变形)如下：

其中是A城市的辐射范围（与B相比），为城市A和B之间的距离，和分别是两个城市的人口。在计算过程中，为了更加符合实际以及方便计算，我们用商圈代替城市，时间成本代替距离，商圈所在行政区人口总数进行计算（考虑是否需要说明）。

Reilly Rule 在一定程度上能够确立商圈辐射范围，但是由于它考虑的因素过少，导致误差十分大，我们使用康帕斯法则得到了商圈辐射范围，但是和实际有很大误差，它的误差如表所示。同样，应用Reilly Rule 还需要具备几个前提：（1）两个城市（商圈）交通情况类似；（2）两个城市（商圈）属性类似；（3）两个城市（商圈）人口（人群类型）类似。而在实际的研究中，这些因素很难测定，并且具有很大的差异性，我们对满足这些前提的商圈进行研究，与实际的误差如表所示。

Huff Model 是由David. L. Huff 于1963年提出，它认为：从事购物行为的消费者对商店的心理认同是影响商店[商圈](https://baike.baidu.com/item/%E5%95%86%E5%9C%88)大小的根本原因，商店商圈的大小规模与消费者是否选择该商店进行购物有关（查原句）。我们对商圈吸引力的研究以及对大都市零售商圈引力模型的建立主要是基于Huff Model，模型如下：

其中 为 地区顾客到商圈 消费的概率， 是商圈 的魅力， 地区顾客到商圈 的阻力， 和 是以经验为基础所估计的修正值， 是互相竞争的商圈数。

Huff Model 从某种程度上得到和实际值很相近的商圈吸引力概率值，但是如果相互影响的商圈数目过多，那么它的计算精度会有一定程度的下降（如图所示），同时在当前经济背景下，更多的因素会对商圈魅力产生影响，而阻力的测定也不仅仅是基于空间距离。在本文中，我们根据文献并且与企业市场经理多次交流，总结了十几个可能对商圈魅力和阻力造成影响的因素，并进行了相关性分析和效应量计算（4.2和4.3），之后对商圈吸引力模型进行了设计（4.5）。

## Data

本文所进行的研究使用了大量数据，极少包含主观因素的影响，因此在一定程度上说具有很高的客观性，所用数据集如下：

交通卡刷卡数据：我们使用了2015年4月上海市轨道交通刷卡数据进行分析研究，数据集包含30天共计4亿多行数据，每一行数据包含刷卡时间，刷卡地点，消费金额，进出站标记等属性。

商圈数据：本文中使用了上海市所有核心商圈和部分大型次核心商圈进行研究，共计19个商圈与142个大型商场的信息。

实际吸引力概率：我们根据交通卡刷卡数据区分出工作人群和购物人群，其中我们认为在五天工作日出现四次以上并且路径完全一致，时间大致相同的cardid为工作人群。而周末两天这些工作人群出现并和工作日出行路径一致，时间大致相同的人认为这些人在周末也在工作。这样，我们就得到了购物人群数，并对这些人群进行分析，得到了不同人群聚集地以及公司聚集地信息。之后我们通过对一个月人流的统计分析，得到了不同地区顾客到不同商圈购物的概率值，这个值我们认为能够代表商圈对其的吸引力程度。

时间成本：由于上海市的交通发达，传统意义上的空间距离对计算结果会产生一定误差，因此我们引入时间成本这个概念来代替空间距离作为阻力因素的一部分。最初我们对上海地铁官方网站数据进行爬取。在初步计算之后我们发现有很大误差，在实际测算之后我们认为这是由于在交通系统中，总是会以最坏情况（即最大时间）进行测算，因此我们对刷卡数据进行统计分析，把任意站和商圈间的乘客以时间排序，然后取前80%进行平均值计算，因为会有部分乘客由于某种原因长时间停留在轨道交通系统中。

商业面积：在商圈吸引力研究中，商业面积也具有很高的地位。但是对商业面积的测定十分困难，我们对每个商圈商场进行统计，并对每个商场的官方主页中得到其公布的商场总面积和营业面积两个数据，之后进行计算，得到了每个商圈的商业面积。但是由于只能得到大型商场的营业面积，无法得到步行街等小型商场的营业面积，在对南京东路、豫园等小型商场集中的商圈的计算会出现误差，因此在0.0中，我们进行了对比研究。

人口总数、人口密度、GDP、人均GDP：这些可能的影响因素我们从政府的年度报表中得到，因此具有很高的可信度。

房屋均价：在与企业经理讨论后，我们认为房屋均价可以一定程度上代表这个地区的经济程度，而经济程度可能会对此地区顾客选择商圈有所影响。这部分数据来自互联网房地产企业公布的信息。

商圈等级，商品档次，枢纽等级：其中商圈等级和枢纽等级来自于上海市政府公布数据，商品档次通过我们对商圈内商场店铺类型和数量进行统计并分类得到。

商圈的活力指数：这部分数据来自于复旦大学城市发展研究院发布的《上海市商圈活力指数月报》(查原句)，是由人流数据、手机信令数据、Pos机刷卡数据等计算得到。

# 人群与聚集地分析

每天生成的上千万条刷卡记录中包含着不同的群体, 如上班族, 老人和游客等. 这部分主要分析上班族群体的移动行为特征与聚集地划分。如果一个人的刷卡记录满足：

其中, *W*表示所有满足条件的上班族集合；*Wi*表示对第*i*个乘客若满足至少在工作日连续四天存在刷卡记录并出现在相同的始发站和目的站，则认为该乘客属于上班族群体。

此外，为了进一步分析上班族的地铁移动行为特征，这部分定义了常规的上班族和非常规的上班族。常规的上班族是指那些一周最多连续工作五天的人；非常规的上班族是指连续工作大于五天的人，通常会在周六或者周日加班。

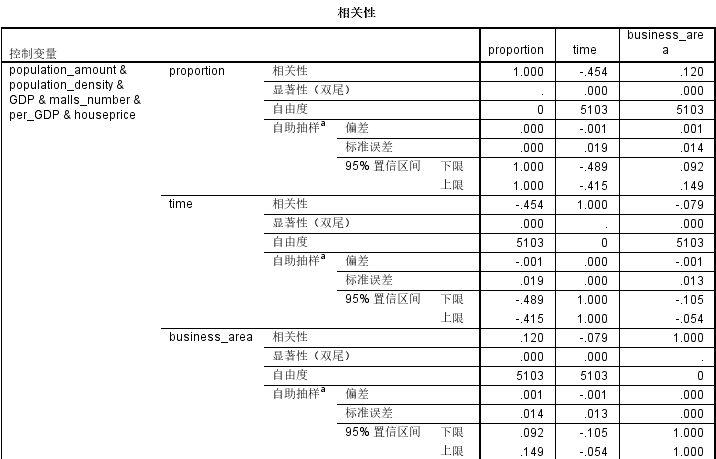
之后，通过可视化的形式，我们得到了上班族的居住地点和工作地点（如图）和购物人群的居住地点（如图），有助于进一步分析该群体的出行行为特征。

商业行为和居民成分有着密切的关系，通过研究我们发现，商圈附近必然会有工作区存在，但是工作区却不一定会依托商圈存在。本文中对具有相似居民组成的区域进行对比研究，验证了模型（如图）。

# 引力模型研究

## 大都市零售商圈现状

## 相关性与相关系数



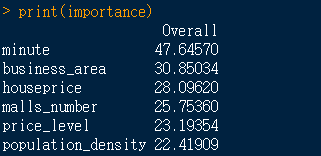
我们计算了Huff Model中两种主要变量对概率值的显著性与相关性，如上图，我们可以看到商业面积与时间成本对商圈吸引力来说都具有显著地相关性，同时时间成本的相关系数为-0.454，商业面积的相关系数为0.120。

由公式aa可以看出，在Huff Model中， 和 是模型调节指数，由于在商业方法中，这两个指数是由相关领域专家通过经验得到，为了对商圈吸引力的研究更加深入，我们邀请了相关领域专家，帮助我们给出两个调节指数值，作为主观指数值，同样，我们通过大样本相关分析，得到了相关系数，把相关系数作为一组调节指数，作为客观指数值。我们加入了一个约束条件， + =2，经过归一化处理和放大处理之后，我们得到了两组调节指数值。我们使用具有两种指数值得模型进行了商圈吸引力概率的计算，得到了商圈的辐射范围。两种调节指数值如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 主观调节指数 | | 客观调节指数 | |
| 调节因子 | (t) | (s) | (t) | (s) |
| 原始数据 | 1.5 | 1.2 | 0.454 | 0.120 |
| 归一化 | 0.556 | 0.444 | 0.791 | 0.209 |
| 放大 | 1.112 | 0.888 | 1.582 | 0.418 |

我们通过加入调节指数，使用Huff Model进行计算，结果如图所示，其中我们使用的是进行归一化与放大操作之后的指数值。

通过模型计算结果的可视化对比之后，我们可以清晰地看出，经过指数调节后的模型精度有了明显的提高，但是两种指数调节方法并没有很明显的优劣性，经过讨论后，我们认为这是由于Huff Model仅仅使用商业面积和距离来进行计算的原因，而实际中，魅力和阻力的确定更加复杂。为了能够得到更准确地吸引力值，我们使用机器学习的方式对数据进行了训练，得到了一组影响因子的值，可能影响因子与训练结果如图所示：



## 商圈引力模型设计

商圈是具有吸引力的，我们模型设计的基础同样是万有引力定律。面积越大，商品种类越多的商圈自然而然的吸引更多的消费者，但是在本文的研究中，商圈选取的都是大城市的核心商圈，商圈属性的差异比较小，在我们的相关性分析中，也能得到同样的结论：商业面积，商圈等级等因素对吸引力结果的影响因子都没有很高。但是一些因素，例如时间成本，换线次数等，对顾客选择商圈产生的影响要高得多。

在经过多次验证和分析之后，我们提出了一个适用于大城市大型商圈的商圈吸引力模型：

我们认为影响顾客选择商圈的阻力因素主要为时间成本，尽管其受到一定的主观情绪影响（地铁线路图中的距离等因素），但是这是影响顾客选择商圈的一个很重要因素。其次，商圈商品档次，商圈内商场数，商圈内商业面积，商圈知名度为主要魅力因素。

而某地顾客到各大商圈的概率值则为：

因为在我们现阶段的工作中，无法验证调节因子的值是否适用于所有相似的大城市商圈。因此，我们通过多种分析方式所做的调节因子研究暂时只适用于上海市。

本文中，我们使用了两种方式对调节因子进行研究与确立，统计学相关性分析的方法与机器学习训练因子的方法。我们分别求出了不同因子的对应值，并对其进行对比研究。结果如图所示：

但是我们发现在某些地点，求出的值会有较大的误差，我们经过讨论认为，可能产生这些误差的因素主要有两个，一个是公交车对数据统计的影响，另外一个是换线次数对模型精度的购物阻力的影响。为了能够更加准确地预测出商圈的吸引力与辐射范围，我们对实验样本进行了分割，并对阻力值进行了优化。得到公式：

我们使用优化模型进行计算，得出的结果对比如图：

## 计算误差研究

我们使用原始模型进行计算，得出了吸引力值，然后计算其与实际值得平均误差，如图所示：

我们可以发现，在整体情况下，时间成本和误差正负与大小没有必然的联系，之后我们缩小范围，以期待能够得到有意思的联系，

很可惜的是，我们并没有得到很明确的联系，这有可能是因为因素分析不够明确的原因，在为来的工作中将会对这个问题进行深入研究。

同时，我们队具有不同调节指数的Huff Model计算值进行比较，如图所示：

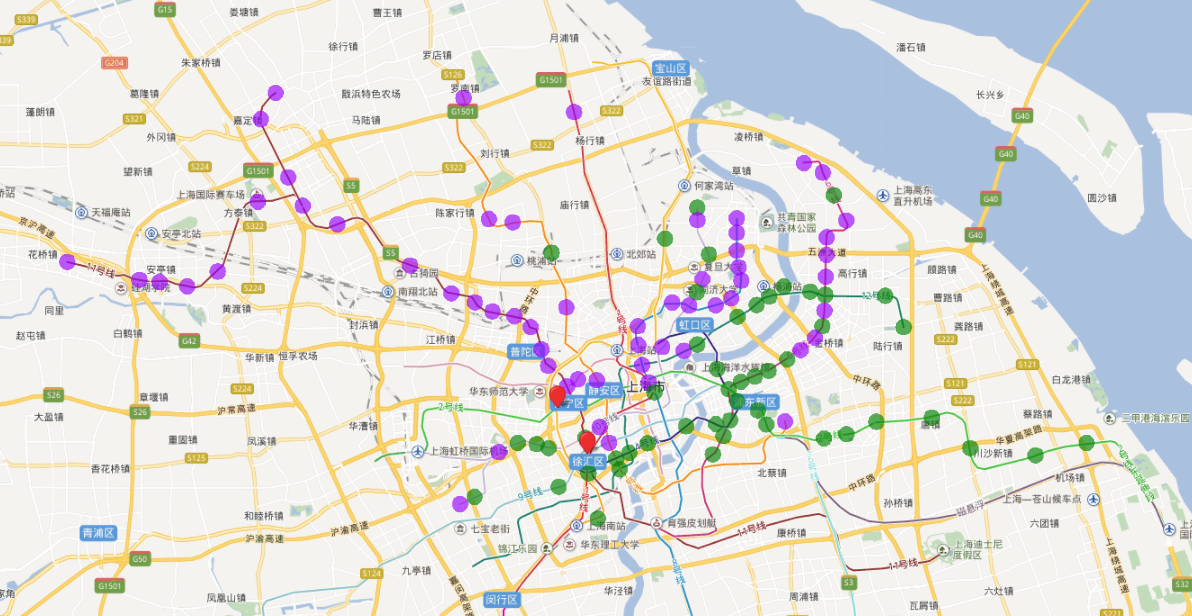
其中客观指数调节，很明显具有更好的准确度，因为其计算结果中误差值小于0.025的较其他两种要高很多，而主观指数调节得到的值能够得到更多的误差小于0.01的数据。

在对较大误差地点进行单独分析时，我们发现，这些站点大多去某一商圈的时间成本小于10 min，经过讨论，我们认为，这是由于数据所产生的误差，因为我们使用的是地铁刷卡数据，对没有加入同样占有公共交通很大比例的公交数据，而商圈附近居民更偏好于乘坐公交车到最近的商圈购物，这就导致了我们测量的实际概率值有误差，这种误差主要体现在，过小的估计了最近商圈对居民的吸引力，导致在之后的计算中产生了较大的误差，但是在现阶段的工作中，我们暂时无法解决这个问题，为了再次提高模型的适用程度，并更好的进行优化和改进，我们去除了这些时间成本小于10 min的地点，再次进行模型计算，实验结果如下：

其中共有201个位置，共3819组数据。

和上图对比，我们可以看到，在去除这些小时间成本的位置之后，我们得到的结果具有更小误差的位置更多，因此，我们认为数据误差在本文的研究中是客观存在的，但是如果我们暂时剔除这些点，我们能够得到相对误差更小的结果。

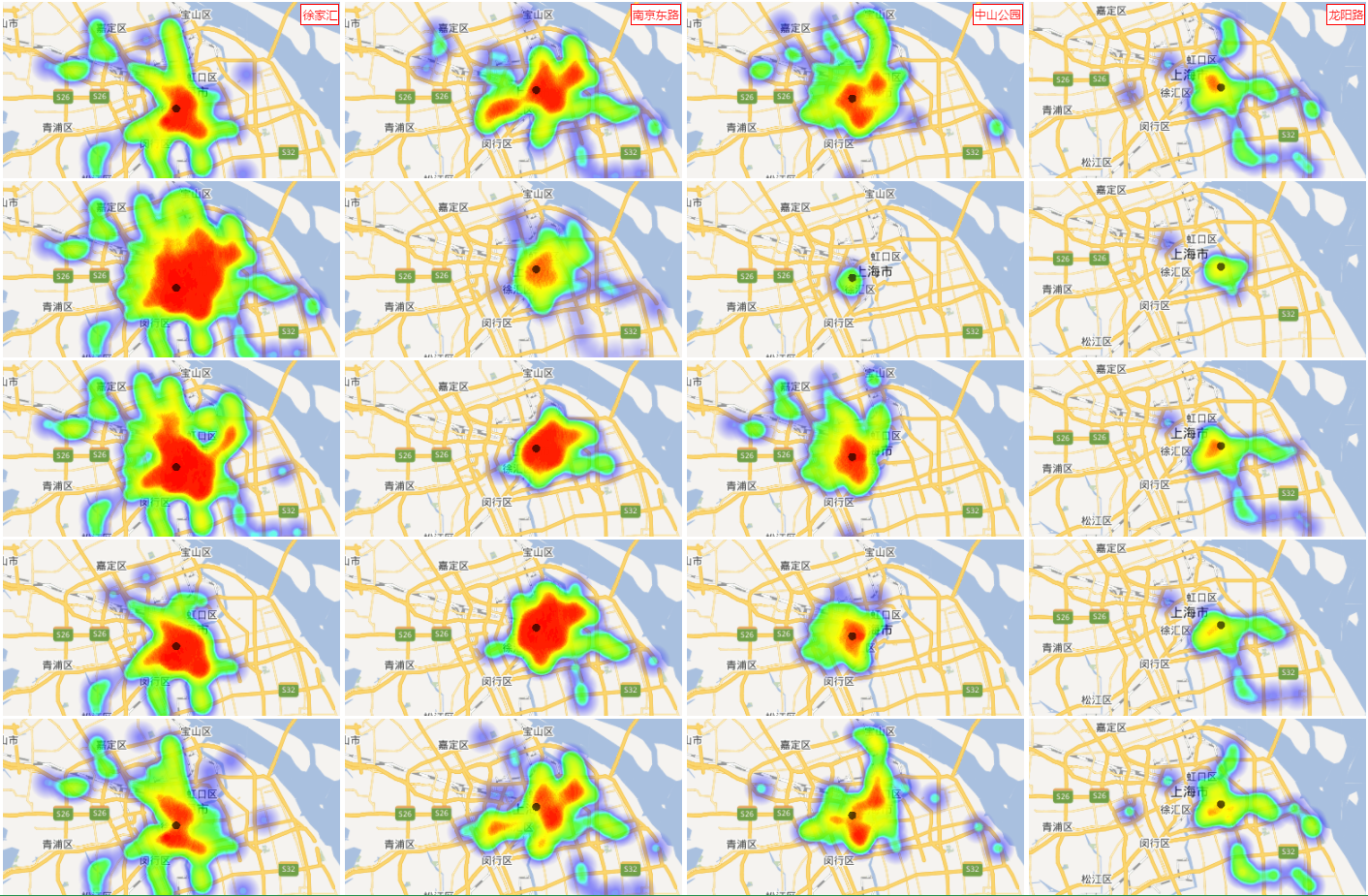
在我们通过实际值绘制辐射范围图时，发现了一个很明显的特征，这是在我们之前的研究中没有重视的，那就是在多种影响因素中，换线次数得多少对结果有很大的影响，我们对统计数据进行分析，发现如果两个商圈对某地的吸引力和时间成本大致相等，那么通过更少换线次数能够到达的商圈更具有吸引力，同时，他们的差异是很明显的，我们对一些具有上述特征的位置和商圈进行深入分析，结果如下（选择商圈—中山公园，徐家汇；地点—时间成本差值小于5min，认为时间成本一致）：



其中中山公园为2,3,4号线交汇，徐家汇站为1,9,11号线交汇。通过可视分析，我们可以清晰地看出具有相同时间成本的位置与两个相似的商圈之间的联系，其中紫色为更加偏好徐家汇的人，绿色为更加偏好中山公园的人。图中，具有更近地理位置并不意味着具有更好的吸引力，可以很明显的看出，如果没有换线的话，那么顾客会更加偏好这个商圈，尽管成本相同，同样的，换线次数越多，那么这个商圈对顾客的吸引力越差（或者说阻力越大），这样，在我们的研究中需要加入换线次数这一个因素，将会很好地提高准确度。

# 案例研究

## 模型对比研究



## 模型有效性验证

本文的工作是基于2015年数据进行的，为了能够验证我们模型的有效性，我们使用2017年数据集进行了验证研究。我们选取了2017年上海市最新发展起来的两个大型商圈，通过模型进行了商圈吸引力分析，并计算出了对应的商圈的影响范围。由于在2015年这两个商圈并没有确立和发展，因此很好地避免了我们研究过程中可能产生对结果造成影响的误差。之后我们把模型计算出的辐射范围与实际辐射范围进行对比，我们发现，我们的模型能够很好地预测出商圈的辐射范围。结果如图所示

我们的模型能够大致规划处商圈的辐射范围，通过辐射范围以及辐射范围内的人群类型研究，可以有效的通过建设针对性商铺来扩大效益，降低零售行业的成本。但是在对商圈周围顾客的吸引力分析依旧有一定的问题，这将是我们下一阶段研究的课题。

# 讨论

# 总结与未来工作