Greggs-Pret 指数：运用机器学习分析消费习惯，作为衡量英格兰南北社会经济差异的指标

Robin Smith[1], **Kristian C. Z. Haverson**[1]

[1]系　　工程与数学系，谢菲尔德哈勒姆大学，霍华德街，谢菲尔德，S1 1WB，英国　　K

电子邮箱: obin.smith@shu.ac.uk

r

2023年4月1日

摘要

在英格兰，人们常戏称，一个城镇中 Greggs 烘焙店的数量，是衡量该地区“北方性”的可靠指标。相反，Greggs 在烘焙食品和三明治市场上的商业竞争者 Pret-a-Manger，据称在英格兰较为“南部”的地区更受欢迎。借助支持向量机和人工神经网络（ANN）回归模型，我们首次利用 Greggs 和 Pret 的相对地理分布，对英格兰的南北分界进行了量化。随后，将计算出的分界线分别与另一条基于家庭国内总收入（GDHI）的分界线进行比较。两者高度吻合。我们认为，这可能是因为英格兰的大部分财富集中在伦敦，而英格兰大多数 Pret-a-Manger 门店也集中于此。我们还根据 Morrisons 和 Waitrose 这两家同样被认为具有南北地域关联的大型超市的相对地理分布开展了进一步研究，但分析结果有所不同。对于所有指标，南北分界线都经过 M1 沃特福德峡服务区附近。作为一个常见的英国习语，人们经常将这一地点称为英格兰南北分界线上的一个位置；值得注意的是，本研究得出了相同结论。这项带有调侃意味的分析旨在强调凸显南北分界的更严肃因素，例如预期寿命、教育和贫困。

关键词： 机器学习、社会、文化、人工神经网络、支持向量机、烘焙食品

1. 引言

英格兰南北分歧的观念贯穿英格兰历史：一条将国家一分为二的界线，使两边在社会、经济和文化方面各不相同。一些历史学家认为，这种分歧最早可追溯到1069年。当时，“如同一头狂怒的狮子”[1], 征服者威廉向北进军，试图控制他新王国中较为桀骜不驯的地区 [2]。此后，正如俗语所说，英格兰北部比南部遭受了更多苦难；这种说法不可避免地催生了许多刻板印象，但本文不作讨论。

Greggs是一家英国烘焙连锁店，专营咸味食品，例如酥皮烘焙食品、香肠卷和三明治，以及甜甜圈、曲奇等甜食。Greggs由John Gregg于1939年在泰恩赛德创立，最初是一家小型面包店。1951年，它在泰恩河畔纽卡斯尔的戈斯福斯开设了第一家门店。部分由于其地理起源，它是一个与英格兰北部相关联的品牌英格兰。从这样不起眼的起点发展至今，Greggs 已成为英国最大的零售烘焙连锁企业，门店遍布全国。近期的快速增长，部分归功于它在 2019 [3] 年初成功推出了著名的纯素香肠卷。



图 1. 南北分界线长期以来一直是英格兰备受争议的问题。

相比之下，Pret-a-Manger（“Pret”）是一家英国三明治连锁店，由 Jeffrey Hyman 于 1983年创立。首家门店于 1984年在汉普斯特德开业——这里是伦敦的一个街区，也是英格兰最富裕的地区之一。公司于 1986年进入清算程序，随后 Sinclair Beecham 和 Julian Metcalfe 从公司清算人手中购得其名称和视觉品牌形象，并将业务发展至 2021年的 4.615 亿英镑营收。

鉴于 Greggs 和 Pret 分别起源于英格兰北部和南部，因此人们常常将“北方性”和“南方性”分别与这两个品牌联系起来。2017年，学生报纸 The Tab 发布了一篇诙谐文章，声称通过在地图上标出英格兰所有 Greggs 门店的位置，界定了英格兰南北分界 [4]。他们的计算基于这样一种观点：一个城市中每家 Greggs 门店对应的人口越少，该地就越偏北。他们选定的临界点是每家门店对应25,000人。任何超过这一数字的城镇或城市都被划分为“南部”。他们的研究颇具争议地将苏格兰和威尔士的城市也纳入计算，而我们的研究将其排除在英格兰南北分界的讨论之外。

本文在 [4] 的研究基础上，尝试通过 Greggs 和 Pret 门店的相对地理分布来界定英格兰南北分界。研究采用了监督式机器学习算法，包括支持向量机和人工神经网络（ANN）回归模型，并通过 Matlab 统计与机器学习工具箱实现。这些方法将在下一节中作更详细的介绍。研究结果还与另一条分界线进行了比较，该分界线由人工神经网络根据英格兰各国际领土层级（第2 [5]级）地区的总家庭可支配收入（GDHI）进行训练后得出。此外，本文还纳入了一种基于 Morrisons 和 Waitrose 门店相对地理分布的南北分界替代定义。

2. 方法论

2.1. 支持向量机

支持向量机（SVM）是一种监督式计算机算法，通过示例学习为对象分配标签 [6,7]。支持向量机已成功应用于研究和社会领域中日益广泛的各种应用。它们在 20 世纪 90 年代发明后不久便声名鹊起，时至今日仍是高准确率分类算法的首选方法。

该算法会获取一组输入数据特征。在本例中，这些特征是英格兰境内每一家 Greggs 和 Pret 门店的纬度和经度坐标。每个数据点都被标记为一个类别；在本例中，类别分别为“Greggs，北部（+1）”和“Pret，南部（–1）”。支持向量机的目标是在训练数据中寻找一个能够最大程度分隔不同类别的超平面。对于我们的二维纬度和经度输入而言，这实际上就是一条直线。该算法通过最大化间隔来寻找能够分隔数据的最优超平面，其中间隔定义为超平面与每个类别中距离其最近的数据点之间的距离。由于 Greggs 门店的数量远多于 Pret 门店，因此会产生一个困难：支持向量机的设计前提是每个类别中的数据量相等。为解决这一问题，我们会反复从 Greggs 数据中抽取随机样本集，使其样本量与 Pret 门店的数量相匹配。随后，针对每个样本集确定最优超平面。我们使用了 1000 个随机样本集，并计算出一个平均超平面，以界定南北分界。

当待分类数据具有许多特征，或数据中存在明显的分隔间隔时，支持向量机尤其有用。对于 Greggs-Pret 数据，结果是一条穿过英格兰的直线，能够最佳地区分 Greggs 和 Pret 门店的地理位置。本研究采用标准的 Matlab 支持向量机模型 [8]。

2.2. 神经网络回归模型

人工神经网络回归模型 [9] 根据一组输入来预测输出变量。分类神经网络的输入特征（自变量）可以是类别型的，但回归模型的输入必须是数值型的。在本应用中，输入变量只有两个：英格兰境内每家 Greggs 和 Pret 门店的纬度和经度坐标。输出设定为：Pret 门店位置为 1，Greggs 门店位置为 0。所使用的人工神经网络是标准的 Matlab 模型 [10]。

神经网络由一组相互连接的层组成，每层包含若干节点。输入变量（坐标）通过数学公式组合，在神经网络中逐层传递，并乘以特定权重，再加上特定偏置。在训练阶段，会调整这些权重和偏置，使网络输出与训练数据输出之间的差异最小化。这里同样存在一个难题：Greggs 门店的数量远多于 Pret，而神经网络在各类别数据量相等时效果最佳。为解决这一问题，我们反复抽取 Greggs 数据的随机样本，使其样本量与 Pret 门店数量相匹配。随后，对每个样本，为了确定英格兰的“北方性”或“南方性”，将英格兰各地均匀分布的一组网格坐标作为输入，提供给训练好的神经网络。

对于每对坐标，都会计算一个数值。接近 1 的数值表示南部，接近 0 的数值则表示北部。此过程使用 1000 个随机样本重复进行，并据此计算出一张平均地图。地图上数值为 0.5 的位置用于确定分隔北部与南部的等高线。

该方法同样应用于GDHI数据。在此，将英格兰所有地区的国际领土层级第2级（ITL2）[5] 经纬度坐标选作神经网络的输入，并将该地区的平均GDHI设为输出。GDHI值经过线性缩放，使最低GDHI为0、最高GDHI为1。随后，再将一组贯穿英格兰、间距均匀的坐标网格输入训练好的神经网络。对于每一对坐标，都会计算出一个数值。接近1的数值表示预测财富水平较高，接近0的数值则表示财富水平较低。之后，以2021年英格兰的平均GDHI（21,962英镑）定义一条等高线，将北部与南部分隔开来。

3. 结果

3.1. Greggs与Pret比较

最初，80%的数据用于训练模型，20%的数据留作模型测试。训练好的支持向量机平均准确率为78%，这意味着在应用于剩余测试数据时，有78%的数据被正确归入Greggs（北部）和Pret（南部）类别。Greggs 和 Pret 门店的分布以及最佳分界线如图2所示。

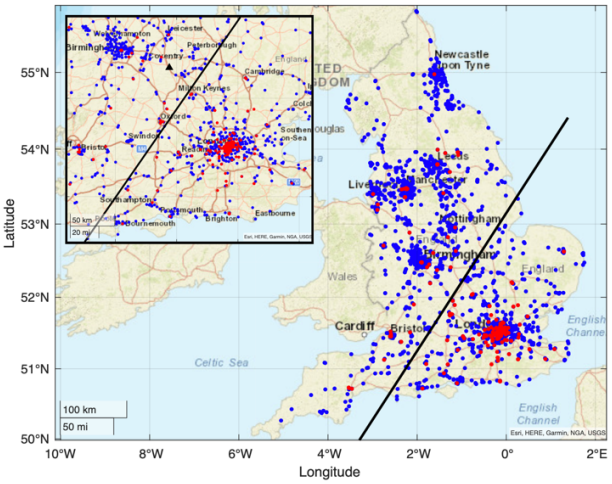


图2。通过使用支持向量机分析 Greggs 和 Pret 门店确定的南北分界线。Greggs 门店以蓝色标示，Pret 门店以红色标示。

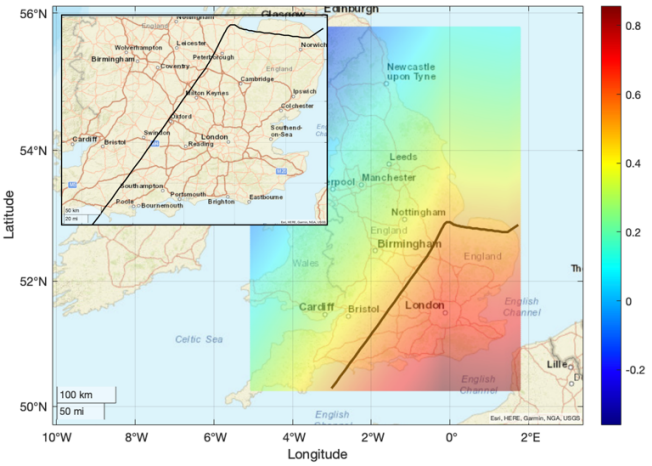


图3。通过对Greggs和Pret门店位置进行人工神经网络回归分析所确定的南北分界线。热力图表示根据位置计算得出的“北方性”和“南方性”。黑线为分隔等高线。

人工神经网络的表现同样出色，在应用于测试数据时取得了81%的准确率。该模型预测的南北分界线如图3所示。在其线性区域内，这一预测与支持向量机的预测相似。然而，随着轨迹接近终点，它向南急转。该模型将诺福克北端，包括金斯林和诺福克海岸杰出自然美景区，划为北部。诺里奇避开了这一划分，仍被归类为南部。

值得注意的是，这两种计算得出的预测线都经过M1高速公路上的沃特福德峡服务区附近。在英国人的谈话中，这里长期以来一直被认为是位于英格兰南北分界线上的地点。

3.2. 家庭国内总收入

与前面的分析一样，80% 的GDHI数据用于训练人工神经网络，20% 留作测试。训练后的人工神经网络平均准确率为 95%，这意味着神经网络能够根据英格兰某地的坐标，以 95% 的准确率预测该地的总家庭可支配收入（在测试数据的 20% 样本范围内）。然而，国际领土级别数据集中只有 43 个英格兰地点，因此统计数据有限。

与第 3.1 节相同，我们生成了覆盖英国陆地范围的经纬度坐标网格。随后使用训练好的人工神经网络预测这些地点的总家庭可支配收入。接着确定了与英格兰平均总家庭可支配收入（21,962 英镑）对应的等高线，如图 4 所示。图中还显示了 Greggs-Pret 分界线，以作比较。

从定性角度来看，这三条线的走势相似。这可能是因为大部分财富集中在伦敦，大多数 Pret 门店也集中于此。

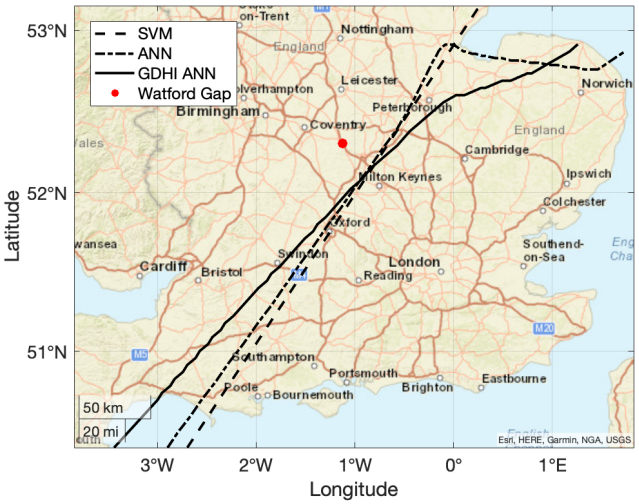


图 4。通过对英格兰43个地区的 GDHI数据进行人工神经网络回归分析确定的南北分界线，并与基于 Greggs 和 Pret 门店位置的预测结果进行比较。

3.3. Morrisons-Waitrose 对比

Greggs-Pret 的地理分布指标存在多方面的缺陷。首先，Pret 是一个高度以伦敦为中心的品牌，这意味着它或许反映的是“伦敦性”，而非“南方性”。此外，这两个品牌在英格兰西南部都不常见。康沃尔是英格兰最偏西南的郡，以康沃尔馅饼闻名。因此，在该地区，Greggs 并未垄断烘焙食品行业* [11]。所以，研究转而考察 Morrisons 和 Waitrose 超市的分布情况。这两个品牌都不与某一座特定城市相关联，而是遍布英格兰各地；然而，Morrisons 与“北方性”相关联，Waitrose 则与“南方性”相关联。

与之前一样，80% 的数据用于训练模型，20% 的数据留作测试。经过训练的支持向量机平均准确率为 64%，这意味着 64% 的测试数据被正确分配到 Morrisons（北部）和 Waitrose（南部）类别中。由于这两个品牌的门店在地理位置上的重叠程度高于 Greggs 和 Pret，因此较低的准确率是合理的。人工神经网络的准确率为 79%。这些模型推导出的南北分界线如图 5 所示。使用人工神经网络推导出的分界线与沃特福德峡相距不到 2 英里。

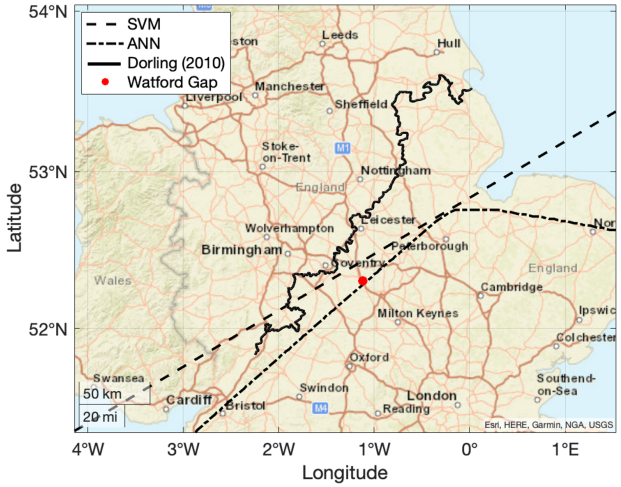


图 5。根据 Morrisons 和 Waitrose 超市的分布确定的南北分界线。Dorling 预测的分界线 [12] 也绘制在图中。

4. 讨论

这项轻松有趣的研究旨在凸显英格兰北部与南部之间真实且重要的文化和社会经济差异，并展示这些差异在某些方面如何反映在消费习惯中。我们尝试将 Greggs-Pret 分界线与家庭国内总收入进行对应，并取得了一定成功。然而，这更有可能反映了英格兰的财富和 Pret-a-Manger 门店都以伦敦为中心这一事实。此外，这项 Greggs与Pret比较还确定，即使包括英格兰最南端，康沃尔实际上也属于北部。基于这些原因，Greggs-Pret 指数似乎存在显著局限性。因此，我们扩大了分析范围，纳入 Morrisons 和 Waitrose 门店的分布，并提取了其他分界线。

在图 5 中，将 Morrisons 和 Waitrose 的结果与 Dorling 于 2010 发布的一条线进行了比较。Dorling 对 [12]

*截至2019年，康沃尔没有 Greggs 门店

英格兰进行了地理分析，考察的因素包括预期寿命、教育、失业、贫困水平和房价。其结果是一条详细的分界线，按照英格兰的每个政治选区对英格兰进行了剖析。我们的 Morrisons-Waitrose 线与 Dorling 经过仔细考量的线有部分重叠。自 Dorling 于 2010 年开展这项工作以来，机器学习和人工智能在众多研究领域中的应用呈爆发式增长。我们认为，使用现代方法对 Dorling 所采用类别中的数据进行最新的重新分析，可能会产生有意义的结果。

5. 致谢

R.S. 感谢 L. 在获取地理数据方面提供的帮助，并感谢 V.A.S.D. 进行校对。

参考文献

[1] McGrath, K. 和 McGrath, K.，2019年。《可耻的 Royal Rage》。Royal Rage 与盎格鲁-诺曼权威的建构，约1000年至1250年，第147-172页。[2] Palliser, D.M.，1993年。《末日审判书与“北部的蹂躏”》。《北方史》，29(1)，第1-23页。[3] Jewell, J.，2019年。Greggs、其纯素香肠卷与有效公关的艺术。《西邮报》，2019年1月11日。[4] https://theteab.com/uk/2017/08/02/weve-figured-ex-actly-north-plotting-every-single-greggs-store-map-44385[5] 地区家庭可支配总收入：城市区域。英国国家统计局。2021年。[6] Noble, W.S.，2006年。什么是支持向量机? 《自然生物技术》，24(12)，第1565-1567页。[7] Boser, B.E.、Guyon, I.M. 和 Vapnik, V.N.，1992年7月。最优间隔分类器的训练算法。载于《第五届年度计算学习理论研讨会论文集》（第144-152页）。[8] 支持向量机分类。MathWorks。https://uk.mathworks.com/help/stats/support-vector-ma-chine-classification.html，访问于2023年3月。[9] Krogh, A.，2008年。什么是人工神经网络? 《自然生物技术》，26(2)，第195-197页。[10] 回归神经网络模型。MathWorks。https://uk.mathworks.com/help/stats/regressionneuralnet-work.html，访问于2023年3月。[11] https://www.timeout.com/uk/food-and-drink/how-cornwall-went-to-war-with-greggs，访问于2023年3月。[12] Dorling, D.，2010年。持续存在的南北分歧。《英国的经济地理学》，第12-28页。