基于机器学习的房地产市场划分方法——以武汉市为例

滕定康[1]　陈思宇[1]　骆杨[1]（房地产估价师证号4220190055）　张静[1]（房地产估价师证号4220150021）

（1.永业行土地房地产资产评估有限公司，湖北 武汉 430062）

摘　要：本文以网络抓取的2024年3月武汉市6154条住宅房地产成交数据为样本，研究如何运用机器学习技术，结合大数据分析手段对房地产市场进行划分。本文选择市场均价作为特征变量进行房地产市场划分，探讨神经网络、K均值模型在房地产市场划分中的应用，并进一步探索如何改进K均值模型至基于K均值的价格与坐标分离模型的空间划分模型，最终确定基于K均值的价格与坐标分离模型是一种较好的房地产市场划分模型。通过利用归一化互信息将基于K均值的价格与坐标分离模型结果与武汉市现行的《房地产市场区域板块划分》和传统克里金插值法进行对比，验证模型的有效性和实用性。研究表明，基于机器学习的“基于K均值的价格与坐标分离模型”在房地产市场分析与管理中具有重要意义。

关键词：房地产市场；大数据；机器学习；K均值

一、引言

自2008年美国《自然》杂志推出“大数据”封面专栏以来，“大数据”逐渐成为互联网技术的热门词汇，其在房地产行业的应用也开始火热起来。事实上，从全球范围来看，房地产行业早在20世纪末就将“大数据”应用于房地产市场管理，如联邦德国于20世纪80年代投入使用的“自动化交易案例收集系统”（AKS）[1]，但该系统为市场信息的收集系统，而缺乏挖掘数据信息的能力。在我国，房地产行业对“大数据”的利用更偏向于利用其挖掘市场信息并辅助于营销活动，例如花样年控股集团利用用户手机APP数据构建“社区电子商务”平台、万科集团通过对其所掌握的业主数据构建“城市配套服务商”等[2]。

总体而言，房地产行业对“大数据”的应用研究仍有待进一步深入。通过机器学习可以充分挖掘房地产市场“大数据”中的有效信息，以实现对房地产市场的有效划分并服务于多种应用，例如在房地产评估中确定同一供需圈的范围、在土地分等定级中划分市场区片等。“大数据”的有效挖掘对房地产市场的分析与管理有十分重要的意义。本文从模型的精度和实践的可行性等角度出发，研究构建房地产市场划分模型。

二、数据获取与处理

（一） 数据选取

一般而言，划分房地产市场区片有两种方法，一是根据每个区域楼盘的特征因素，如房屋年限、容积率、

交通通达度、设施状况等进行划分；二是直接利用每个区域楼盘的市场均价进行划分。事实上，区域楼盘的很多特征因素与区域楼盘的市场均价之间存在较强的线性相关关系，因而本文选择以武汉市住宅房地产楼盘（不含别墅）的市场均价为特征变量，研究房地产市场区片的划分。选取武汉市住宅房地产市场作为研究对象的原因是：（1）武汉市是典型的多中心发展模式，房地产市场分区明显；（2）武汉市住宅房地产市场较为活跃，交易样本数量较多。剔除别墅的主要原因是：别墅与一般住宅在产品特性、市场供需存在明显的差异，不应归入统一市场类型。

（二） 数据来源

本文所选取的数据来源有两个渠道：（1）房地产均价通过设计爬虫软件抓取的房天下网站房源信息。选择房天下的原因是：房天下收集楼盘数量多，数据类型丰富，满足研究需求；（2）房地产位置来源于百度地图 API 获取的房源经纬度坐标，并通过地理编码连接房地产市场价格与坐标数据。共获得 6154 条原始价格信号，交易时间为 2024 年 3 月；（3）房源坐标数据采用 WGS-84 坐标系统。

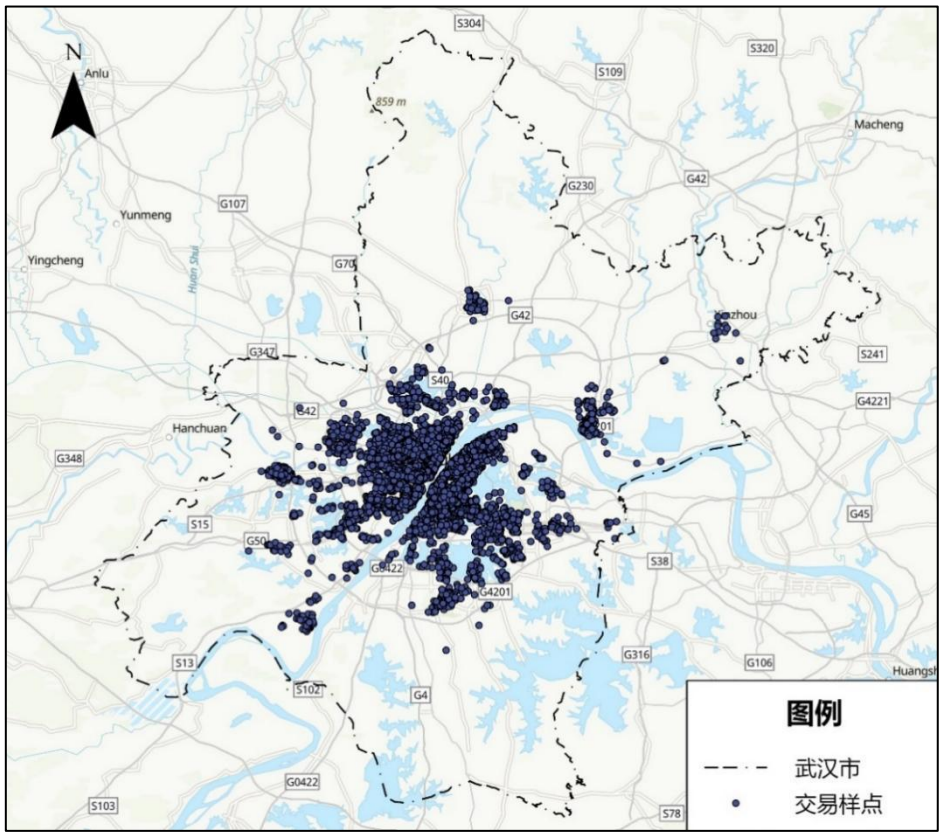

图1 交易样点分布情况

Fig. 1. The Distribution of Trading Sample

三、模型构建

（一） 神经网络模型

神经网络（ANN）是目前机器学习中最常用的非线性分析模型之一，常应用于监督学习、无监督学习

和强化学习等不同场景。在房地产市场中，神经网络模型得到大量的应用，如陈诗沁，王洪伟[4]提出的基于机器学习的房地产批量评估模型，莫丽娟,李燕宁[5]等人提出的基于人工神经网络统计学模型的房地产价值估算方法等。

然而，神经网络模型在房地产市场方面的应用存在着一些无法逾越的问题。从原理上看，神经网络是具有大量相互连接的简单过程的大规模并行系统，可以被视为一种加权有向图，其中人工神经元是节点，带权重的有向边是神经元输出和神经元输入之间的连接[6]。但是这就会带来一个问题，即神经网络的内部是一个无法预测的“黑箱”，尽管神经网络能提供一种确定所有预测变量的总体影响的方法，但变量之间的相互作用却难以解释，阈值的权重也难以确定[7]。这种解释力的缺乏是将神经网络应用于房地产市场的一个主要问题，因为各变量（即影响房地产市场状况的各个因素）之间的作用对房地产市场会产生较大的影响，且在实际的房地产市场应用中，如房地产估价、房地产市场区域划分等，需要综合考虑影响房地产市场状况的各个因素得到一个综合性的结果。而神经网络模型往往只能给出一个概论性的结果，缺乏对影响因素的具体解释。

（二） K 均值（K-Means）模型

1.K 均值算法

K 均值算法是由 Hartigan 和 Wong[9]于 1975 年提出的一种机器学习聚类算法，与神经网络不同，属于一种典型的无监督学习方法。其基本原理是将 N 维的 M 个点划分为 K 组，使得每组内数据的平方和最小，即每个组内的点都距离该组均值（聚类中心）最近。K 均值算法的目标是找到使得下式满足的聚类S_i：

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \qquad (3.1)$$

式中μ_i是S_i中所有点的均值，k是分组的数量，x是S_i中的向量。

因此，以市场均价为特征变量时候，K 均值算法可以很好的将价格接近且位置相接近的样点划为一组。相对神经网络算法，K 均值算法能够很好解释变量之间的关系，即每一组变量与其平均值相接近，对房地产市场具有相近的影响度。

2.模型的实现

（1）依据行政区划划分子集

由于武汉市不同行政区及功能区的房地产市场拥有不同的特性，因而首先按行政区划及功能区划，将数据集划分为 15 个子集。

子集编号	行政区	样本数量	子集编号	行政区	样本数量
1	蔡甸	277	9	江岸	899
2	东湖高新	388	10	江汉	620
3	东西湖	347	11	江夏	238

4	沌口	131	12	硚口	458
5	汉南	61	13	青山	204
6	汉阳	507	14	武昌	1002
7	洪山	624	15	新洲	156
8	黄陂	242	合计	--	6154

表 1 房地产市场样点子集划分表

Fig 1. Division of subset of real estate market sample

（2）对子集进行聚类分析

输入房地产市场样点价格、经度、纬度三个维度的数据，使用 K 均值算法对子集中的交易样点进行聚类分析，得到每个子集中市场价格相近且距离相近的组别。

（3）将点数据扩充为面数据

由于我们需要的是面状市场均质区划分结果，而 K 均值算法仅对交易样点进行了分类，因而可以通过创建泰森多边形的方法，将点要素转换为面要素，融合并裁剪后得到最终的划分结果（图 2）。在此划分结果的基础上，还可以进一步按地物界限进行整饰，本文不做赘述。

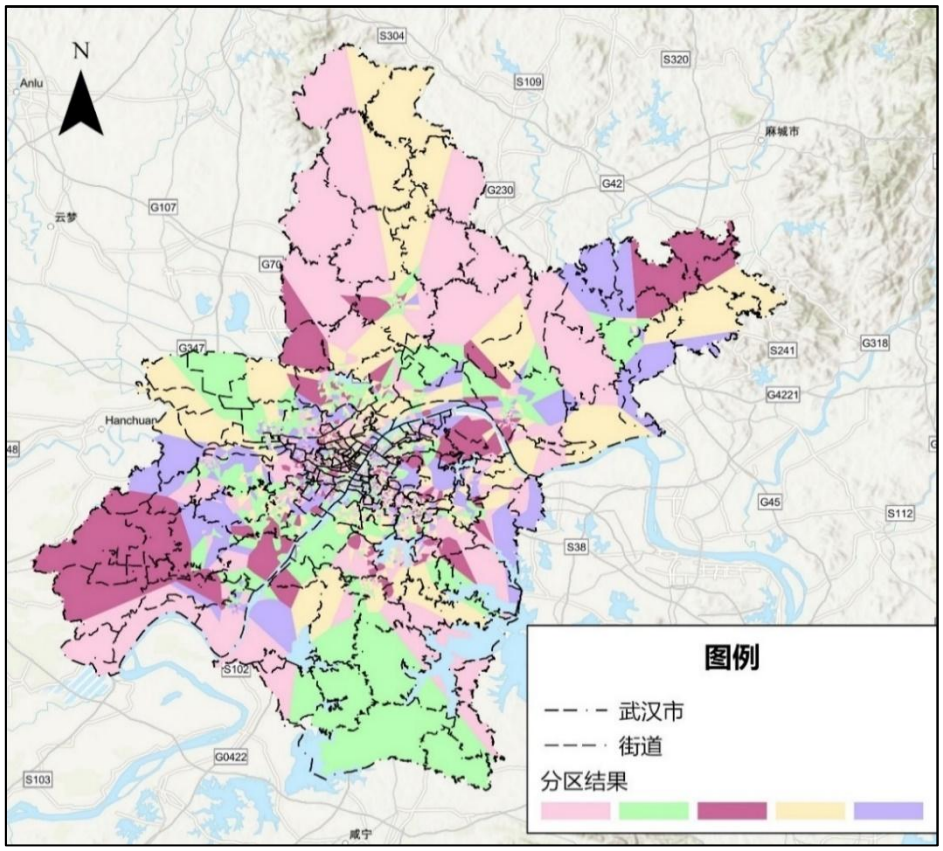

图2 K 均值聚类分析结果

Fig. 2. Result of K-Means Clustering

3.模型存在的问题

从图 2 中可以看出，K 均值模型可以较好的依据房地产市场分布情况划分出市场均质区域。但是该模

型还是存在着一些问题。以武昌区的徐家棚街道为例，将图片放大后（图 3），可以明显看到，交易样点中的异常点对聚类结果影响较大，很多异常点被独自划分为一个市场区域，使得分区结果非常零碎，难以进行细碎图斑处理，进而影响房地产市场分析与管理的实际应用。如果采用人工方式调整异常点，工作量将会十分巨大。

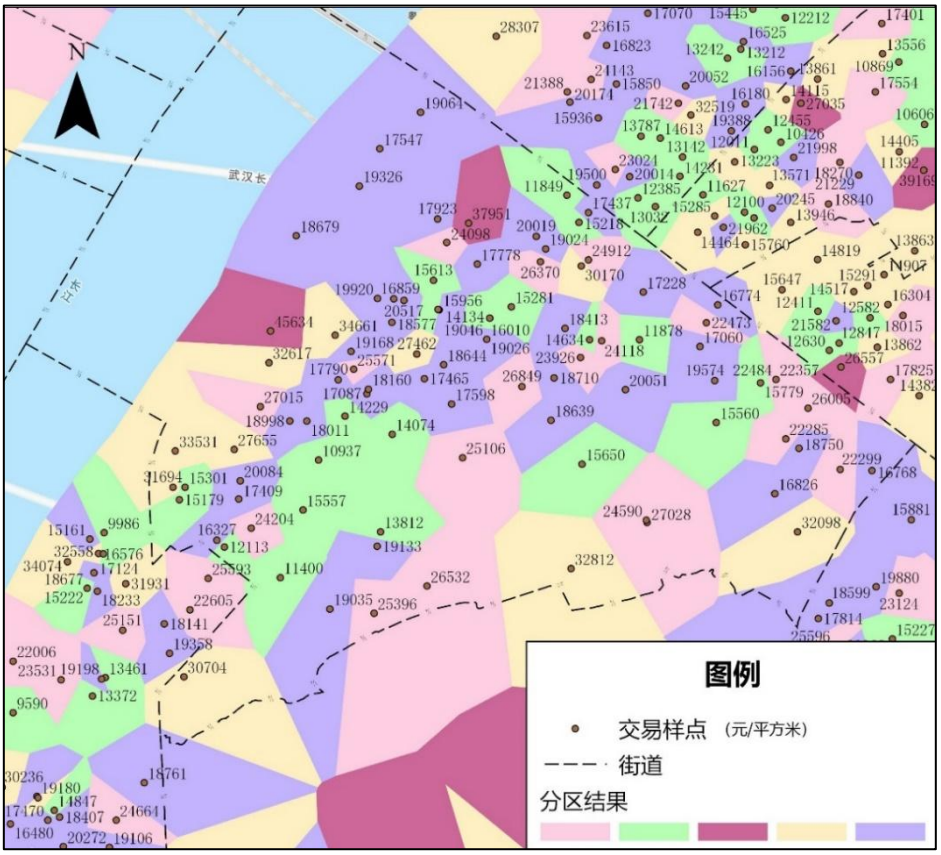

图3 K 均值聚类分析结果局部放大图

Fig. 3. Partial Enlargement of K-means Clustering Analysis Results

从 K 均值算法的基本原理来看，由于该算法的划分依据是组内数据的平方和，组内的每个影响因素具有相同的权重，但数据中价格的极差远大于位置坐标的极差，使得算法结果受到价格因素的影响大于受到坐标因素，即位置因素的影响，进一步强化了异常点对分类结果的影响。

考虑到这些问题，尽管 K 均值模型有着较高的分类精度，但其难以满足复杂情况下房地产市场划分的实践要求。

（三） 基于标准化的 K 均值模型

针对上述问题，可以考虑采用 0-1 标准化以及 Z-Score 标准化来减小价格和位置坐标单位不同的问题，或者使用过自定义距离（使用参数 α 来控制坐标和价格权重）来尝试解决异常点的问题。但通过实验发现，标准化处理后的数据虽然能一定程度解决异常点被单独划分为单独区域的问题，但是仍无法解决不同区域相互交错分布的问题，生成效果与 K 均值模型相比没有明显的提升，分类到不同区域的样本点互相夹杂分

布，使得分类结果存在大量噪声，因此不论使用何种方式对分类结果进行处理，都无法达到一个比较理想的最终结果。

考虑到数据中价格为一维向量，而位置坐标为二维向量，两者的维度不同，因而 K 均值模型和基于标准化的 K 均值模型无法在两个维度上同时对数据进行较好的分类，从而产生夹杂分布的问题。

（四） 基于 K 均值的价格与坐标分离模型

如前文所诉，由于房地产价格与位置坐标难以在同一个维度下直接运用 K 均值模型计算分区状况，因此可以考虑用 K 均值算法分别对算房地产的价格与位置坐标进行聚类。将两个聚类结果结合起来即可得到初步的房地产市场划分情况。其基本步骤如下：

（1）依据行政区划划分子集

由于武汉市不同行政区及功能区的房地产市场拥有不同的特性，因而首先按行政区划及功能区划，将数据集划分为 15 个子集。

（2）对子集房地产价格进行聚类分析

输入房地产市场样点价格一个维度的数据，对子集中的房地产交易样点的价格进行聚类分析，得到每个子集中市场价格相近的组别S_{pi}：

$$S_{pi} = (x_1, x_2, \dots, x_n)^T \tag{3.2}$$

式中S_{pi}为第i个子集中的房地产价格聚类结果，x_n为子集中第n个交易样点价格聚类的簇号。

（3）对子集房地产位置坐标进行聚类分析

输入房地产市场样点经度、纬度两个维度的数据，对子集中的房地产交易样点的位置坐标进行聚类分析，得到每个子集中空间位置坐标相近的组别S_{ci}：

$$S_{ci} = (y_1, y_2, \dots, y_n)^T \tag{3.3}$$

式中S_{ci}为第i个子集中的房地产位置坐标聚类结果，y_n为子集中第n个交易样点坐标的簇号。

（4）计算聚类结果

根据房地产价格的聚类结果及房地产位置坐标的聚类结果，计算最终的聚类结果S_i：

$$S_i = S_{pi} + kS_{ci} \tag{3.3}$$

式中S_i为第i个子集中聚类结果，k为一个常数，其目的是为了使每两两相加的聚类簇号值唯一。

（5）将点数据扩充为面数据

通过创建泰森多边形的方法，将点要素转换为面要素，融合相同分类的图斑并裁剪后得到初步的划分结果，但此时仍存在一些细碎图斑需要处理。

（6）处理细碎图斑

对于小于 1 平方公里的图斑，利用融合（Eliminate）算法，将面与具有最大面积或最长共享边界的相邻

面合并来消除细碎图斑。在此基础上，统计小于 5 平方公里且样点数小于 5 的图斑，再次进行融合，得到最终的划分结果（图 4）。在此划分结果的基础上，还可以进一步按地物界限进行整饰，本文不做赘述。

（7）计算区域价格均值

通过计算区域内房地产市场样点的平均价格，还可以体现每个分区的平均价格及各分区间的价格趋势（图 5）。

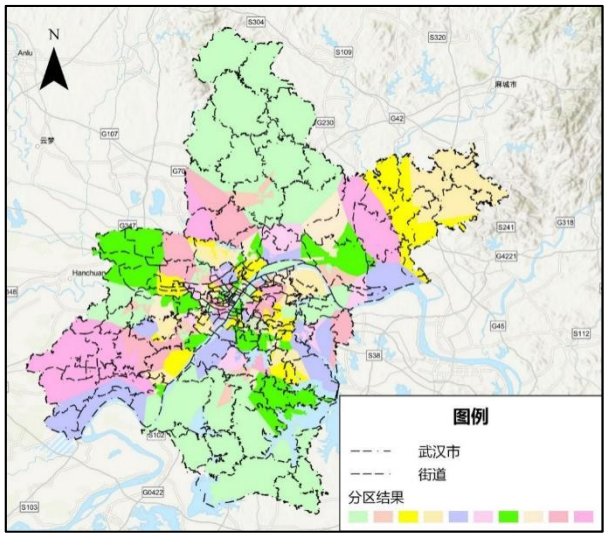

图4 基于 K 均值的价格与坐标分离模型聚类分析结果

Fig. 4. Result of K-Means Clustering with the Separation of Price and Coordinate

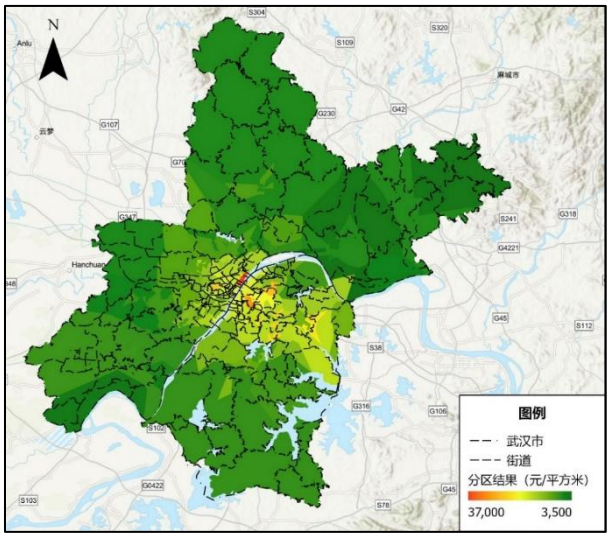

图5 基于 K 均值的价格与坐标分离模型聚类价格区间

Fig. 5. Price Division of K-Means Clustering with the Separation of Price and Coordinate

四、模型构建效果对比

（一） 与武汉市《房地产市场区域板块划分》的对比

《房地产市场区域板块划分》（DB4201T 639-2020）是由武汉房地产经济行业协会等[9]行业专业机构起草编撰的武汉市地方标准文件，对武汉市房地产市场范围进行了划分。文件根据地理相接近，房屋、交通、资源环境、房地产价格等属性相似的原则，以人工或自然边界为界限将武汉市房地产市场划分为了 160 个板块。该文件可以作为验证基于 K 均值的价格与坐标分离模型划分结果准确性的依据。

将基于 K 均值的价格与坐标分离模型结果与《房地产市场区域板块划分》进行叠加（图 6），可以看出该模型的区域划分结果与《房地产市场区域板块划分》有着较好的契合度。由于未对模型结果进行细致的边界整饰，模型结果的边界与《房地产市场区域板块划分》的区域边界不完全重合。

将《房地产市场区域板块划分》结果和基于 K 均值的价格与坐标分离模型结果分别与交易样点叠加，可以得到交易样点在两种划分方式下分别落入了哪个划分区域，进而定量分析基于 K 均值的价格与坐标分离模型结果与《房地产市场区域板块划分》结果的相关性，确定其划分的准确度。由于房地产市场区域划分结果不是简单线性相关关系，因此可以计算归一化互信息（NMI）值来比较两个结果的相关性。相对传统的相关系数，互信息（MI）可以较好的衡量非线性关系[8]，而归一化互信息是对互信息进行归一化处理，去除

了量纲对结果的影响，使其值区间落入[0,1]，更易于定量比较各结果之间的相关性。

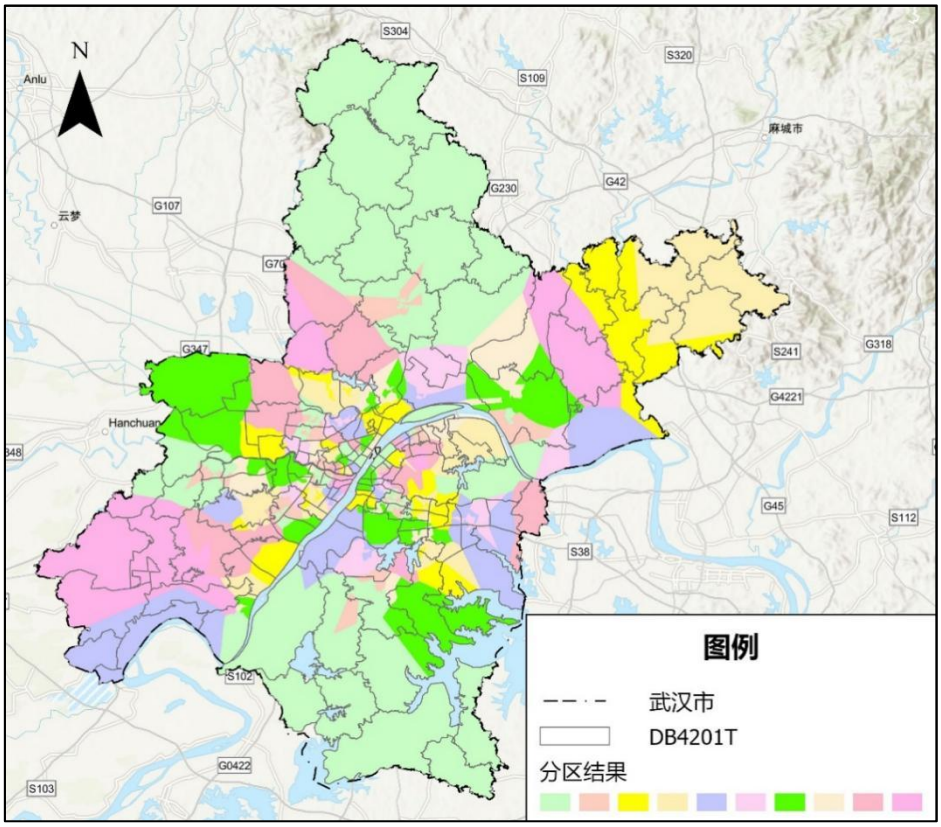

图6 《房地产市场区域板块划分》与基于 K 均值的价格与坐标分离模型聚类分析结果对比

Fig. 6. Comparison of Regional Division of the Real Estate and K-Means Clustering with the Separation of Price and Coordinate

归一化互信息的计算公式如下：

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{4.1}$$

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{4.2}$$

$$H(X,Y) = -\sum_{x \in X, y \in Y} p(xy) \log_2 p(xy) \tag{4.3}$$

$$NMI(X,Y) = \frac{H(X) + H(Y)}{H(X,Y)} \tag{4.4}$$

式中$NMI(X,Y)$表示种分类方法X和Y的归一化互信息值。其中X，Y是离散的随机变量，且$p(x)$和$p(y)$是其分别的概率密度分布，$p(xy)$则为其联合概率分布，$H(X)$和$H(Y)$表示其分别的信息熵，$H(X,Y)$是联合信息熵。

由于归一化互信息的取值位于[1,2]，为了统一，将归一化互信息通过线性拉伸变换到[0,1]：

$$SNMI(X,Y) = \frac{NMI(X,Y) - min}{max - min} \tag{4.5}$$

式中max、min表示所有$NMI(X,Y)$的最大值和最小值，$SNMI(X,Y)$表示最终的归一化互信息。

将《房地产市场区域板块划分》和基于 K 均值的价格与坐标分离模型中划分的板块分别从 1 开始按顺

序递增编号，计算得到两种方法的归一化互信息值为 0.8283，说明两种划分方式之间有较高的相关性。

（二） 与克里金插值法的对比

克里金插值（Kriging Interpolation）可以通过计算数据点之间的空间相关性（即变异函数或协方差函数）来估计任何坐标的值，是空间预测中最常用的方法之一[9]。克里金插值法数学基础和应用与机器学习有一些相似之处，如两者都依赖于已有数据来进行预测，且某些机器学习方法（如高斯过程回归）和克里金插值在数学上有相似之处，特别是在处理空间相关性方面。一般而言，利用克里金插值法划分房地产市场情况的基本思路为：（1）使用克里金插值法对样本进行插值计算；（2）对插值结果进行聚类；（3）对聚类结果按地物界限进行整饰，得到市场分布情况。本文主要对比基于 K 均值的价格与坐标分离模型结果与克里金插值结果的差异，故未按地物界限整饰聚类结果。

克里金插值法包括使用漂移量解析函数的泛克里金（Universal Kriging），直接使用辅助预测因子来产生克里金权重“外部漂移克里金法”（Kriging with External Drift），以及主要用辅助数据（Auxiliary Data）解释的不同局部均值来整合回归的“回归克里金法”（RK）等 [11][12]。从数学上来看，他们都属于插值法，由于本研究中没有其他辅助变量的信息（只有空间坐标），因此泛克里金是最合适进行比较的地统计学方法。

以武汉市行政区边界为插值边界及掩膜，使用范克里金插值法对所收集的武汉市房地产成交数据进行插值，得到武汉市房地产市场分布情况（图 7），从图中可以看出，在处理大数据模型的时候，克里金插值法在边界区域存在显著的过拟合现象，如西南角的远城区因缺少足够的样点而被推导成为市场价格较高的区域（如图 8 所示），但通过查询该区域的房价与地价，该区域实际却应该属于市场价格较低的区域。

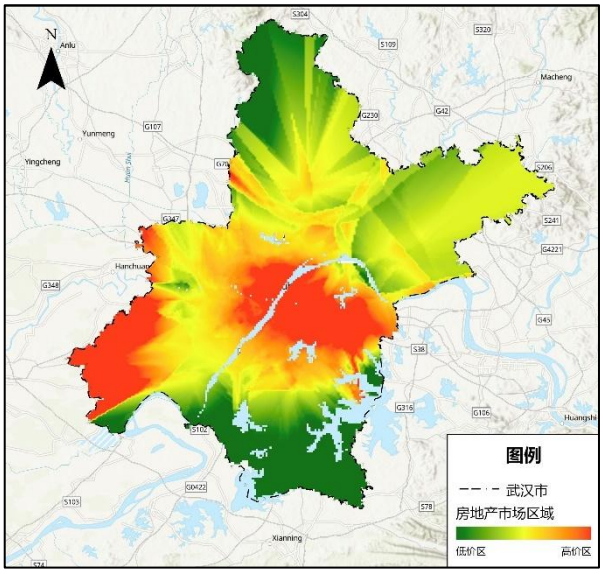

图7 泛克里金插值结果

Fig. 7. Result of Universal Kriging

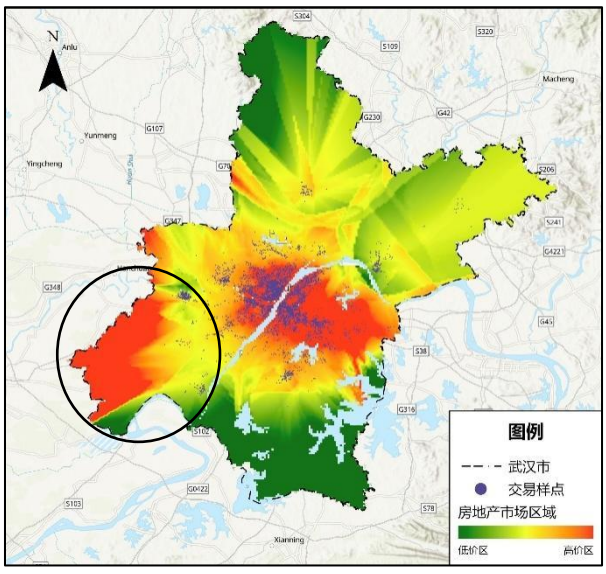

图8 交易样点分布与克里金结果对比

Fig. 8. Comparison of Sample Distribution and Kriging Result

将泛克里金插值结果与《房地产市场区域板块划分》叠加（图 9），虽然插值结果未按地物界限进行整

饰，仍然可以看出中心城区插值结果的边界与《房地产市场区域板块划分》的契合度一般，且远城区仍然会因为过拟合等原因而产生一定偏差。两种分区结果的归一化互信息值为 0.4063，也说明单纯的克里金插值法并不是特别适合直接用于房地产市场的划分。

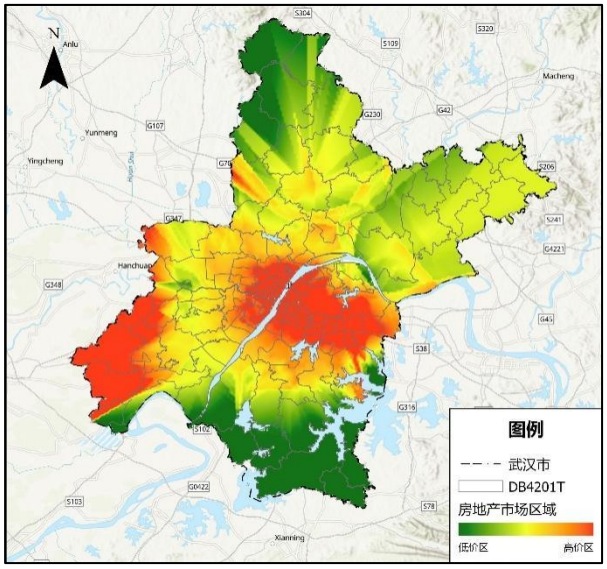

图9 《房地产市场区域板块划分》与克里金对比

Fig. 9. Comparison of Regional Division of the Real Estate Market and Kriging Result

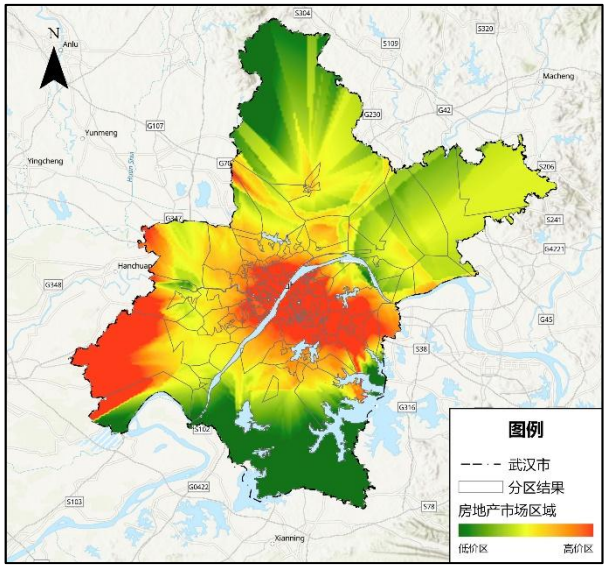

图10 基于 K 均值的价格与坐标分离模型与克里金对比

Fig. 10. Comparison of K-Means Clustering with the Separation of Price and Coordinate and Kriging Result

将基于 K 均值的价格与坐标分离模型结果与克里金插值结果进行叠加（图 10），可以发现基于 K 均值的价格与坐标分离模型具有更好的分区效果，且每个分区也能较好的覆盖插值结果。两种方法的分区结果的归一化互信息值为 0.3968，说明其之间的相关性较低。究其原因，还是克里金插值原理是依据现有结果预测未知区域的数值，难以过滤异常值，且在数据稀疏的区域容易产生过拟合现象。克里金插值法由于其预测性质和对数据点间空间关系的依赖，在处理市场划分时不如直接的聚类分析方法精确，尤其是在需要考虑市场板块的明确界限和异常值处理的场景下。

五、总结

神经网络作为目前大数据分析中最常用的非线性分析模型之一，在房地产市场分析中扮演着日益重要的角色。但由于其内部复杂且不透明的运算过程，导致难以解释各因素如何具体影响最终结果。在房地产市场分析中，这种缺乏透明度和解释性的特点成为一大局限，而 K 均值聚类模型是根据数据的最小平方和进行聚类，可以解释为每个分组属性上相接近，拥有相似的市场条件，能够较好的解决这个问题。

然而普通 K 均值模型难以解决房地产交易样点的价格与位置坐标单位不同的问题，虽然该问题可以通过标准化来解决，但维度不同的问题依然存在。因而提出基于 K 均值的价格与坐标分离模型。

在模型的准确性上，将其结果与《房地产市场区域板块划分》的比较，可以发现两者拥有较高的归一化互信息值（0.8283），这表明这两种方法在板块划分上有着高度的一致性或信息重叠，说明改进后的 K 均值

模型能够较好地捕捉到实际市场板块的结构。当将克里金插值法的结果与《房地产市场区域板块划分》进行比较时，归一化互信息值降低至 0.4063，表明其结果与标准划分的契合度较低，尤其是在处理数据稀疏的远城区时，过拟合问题导致了较大的偏差。同时，将基于 K 均值的价格与坐标分离模型与克里金插值结果进行对比，归一化互信息值为 0.3968，也显示出较低的相关性。这说明 K 均值模型在一定程度上能够提供与克里金插值不同的视角，可以避免克里金插值法由于其预测性质和对数据点间空间关系的依赖所产生的问题。

综上所述，基于 K 均值的价格与坐标分离模型在处理大数据房地产市场信息上表现出了更好的适应性和准确性。在大数据环境下该模型能较为有效的对数据进行处理，并形成标准化的流程。目前模型仍然有较大的改进空间，如机器学习中有专门用于空间位置聚类的密度聚类算法（DBSCAN）、OPTICS 算法等，可以考虑将此类算法融入基于 K 均值的价格与坐标分离模型计算位置坐标的聚类。此外，由于 K 均值分类的初步结果为离散的点结果，可以考虑融合计算机视觉大数据算法中的边缘检测算法（Edge Detection）直接划分出边界结果。另一方面，还可以考虑适当引入更多的非线性相关的分类变量，如市场预期、基础设施状况等，以进一步提高模型的准确度与泛用性。

参考文献

[1] 曲卫东.土地估价信息系统(LAIS)[J].中国土地科学，2003，17(2):6.DOI:10.3969/j.issn.1001-8158.2003.02.005.

[2] 杜丹阳,李爱华.大数据在我国房地产企业中的应用研究[J].中国房地产:学术版, 2014.

[3] Yfantis E A, Flatman G T, Behar J V. Efficiency of Kriging Estimation for Square, Triangular, and Hexagonal Grids[J]. Mathematical Geology, 1987, 19(3):183-205.DOI:10.1007/BF00897746.

[4] 陈诗沁,王洪伟.基于机器学习的房地产批量评估模型[J].统计与决策, 2020(9):181-185.

[5] 莫丽娟,李燕宁,吴骞.基于人工神经网络统计学模型的房地产价值估算方法:CN201210283427.8[P].CN103578057A[2024-06-07].

[6] Jain A K, Mao J, Mohiuddin K M. Artificial Neural Networks: A Tutorial[J]. Computer, 2015, 29(3):31-44.DOI:10.1109/2.485891.

[7] Jackson O D A. Illuminating the “Black Box”: A Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks[J]. Ecological Modelling, 2002.DOI:10.1016/S0304-3800(02)00064-9.

[8] 李亮,舒宁,王琰.利用归一化互信息进行基于像斑的遥感影像变化检测[J].遥感信息，2011(6):5.DOI:10.3969/j.issn.1000-3177.2011.06.004.

[9] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.DOI:10.2307/2346830.

[10] DB4201/T 639-2020,房地产市场区域板块划分[S].

[11] Hengl T, Heuvelink G B M, Stein A. A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging[J]. Geoderma, 2004, 120(1-2):75-93.DOI:10.1016/j.geoderma.2003.08.018.

[12] Hengl T, Heuvelink G B M, Rossiter D G. About Regression-kriging: From Equations to Case Studies[J]. Computers & Geosciences, 2007, 33(10):1301-1315.DOI:10.1016/j.cageo.2007.05.001.