1 对平行进口环境制约因素及企业绩效的实证分析

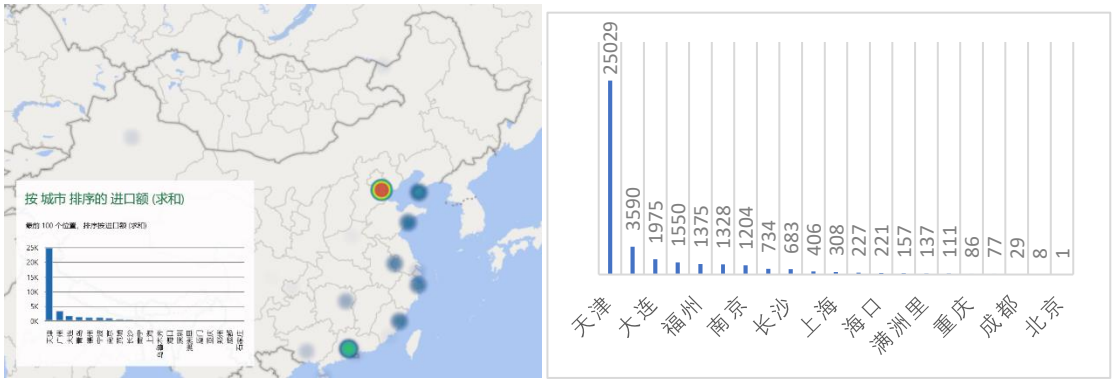
1.1 对平行进口环境制约因素的影响分析

1.1.1 聚类分析

聚类指标选取

自我国 2014 年 8 月份陆续开设平行进口试点区域以来，已经设立有 21 个进口汽车整车口岸，其中 18 个具备平行进口试点资质。这些试点区域主要分布在东部沿海地区，少数诸如重庆、成都等地区也靠内陆港获得一席之地。

图 1.1 汽车平行进口口岸进口额热力图及数据



由图 1.1 知，试点区域由华中地区一路延展至海南，地域差异大，经济发展不平衡，且运输条件也因地而异，对于平行进口地区制约关系的聚类分析不能简单以单一指标作为分类标准，而是应该建立科学的指标体系来进行聚类分析。

在对不同区域的平行进口额进行指标分析时，需要对影响平行进口各个环节的主要方面进行选取，即：对于环节末，要依靠地区经济发展水平来判断进口需求；对于环节中，需通过港口吞吐量来判断进口能力；对于环节始，进口源不同，导致进口质量也有较大差异。故此，本文选取如下表所示的聚类指标对不同区域不同口岸按进口需求和实力进行聚类分析。

表 1.1 聚类指标

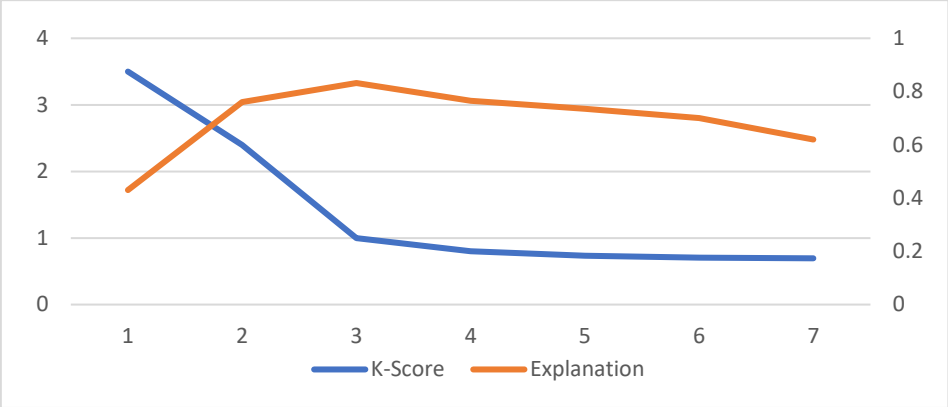
指标类型	聚类指标量化
进口需求	人均可支配收入
	恩格尔系数
进口实力	国民生产总值（GDP）
	汽车保有量
	海港吞吐量

K-Means 聚类结果分析

常见聚类方法有 K-Means 聚类、均值漂移聚类、DBSCAN、GMM 聚类、凝聚层次聚类以及图团体检测（GCD）等。考虑到所探讨的平行进口区域聚类特征明显，且其聚类目标明确，同时存在标签过少和一次性等问题，而 K-Means 聚类方法对目标点维度要求较低，速度较快，在给定聚类数时聚类效果通常较好，故本文选取 K-Means 方法，运用 Python 中的机器学习库 Sklearn 进行聚类。

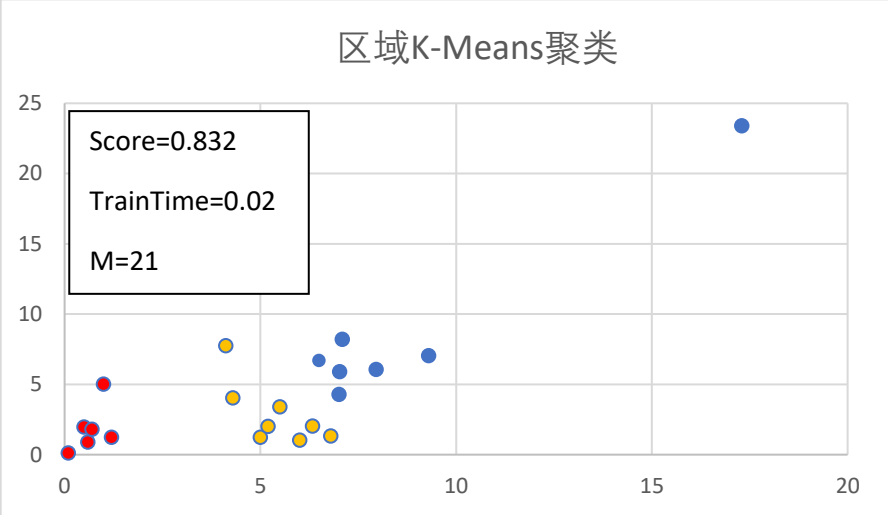
首先，对天津等 21 个口岸进行 LabelEncoder 编码，其值为 1~21。在聚类开始前，要先对数据进行均值归一化（Batch Norm），以避免因单项数值过大或过小带来的失衡现象。聚类 k 值评估：

图 1.2 K-Score 与解释率



表明，当 K=3 即分类数量为 3 组时，能更好的解释样本差异性。取 K=3，并取 max_iter=20,利用 TSNE 将数据降至 2 维，可视化如下：

图 1.3 区域 K-Means 聚类结果



注意到，由于天津港进口额占比过高，吞吐量也较大，虽然分为三类之一，但考虑到其特殊性，在此将其归为特大特殊区域；且由于黄埔隶属于广州市，所以将其与广州港归为一类代表广州。根据其输出矩阵，具体划分的结果如下表所示：

表 1.2 聚类分析结果表

类别	一类区域	二类区域	三类区域	特大特殊区域
地区	广州，大连，青岛，福州，宁波，南京	长沙，南宁，上海，海口，深圳，厦门	重庆，郑州，北京，石家庄，成都，满洲里，乌鲁木齐	天津
地区数	6	6	7	1

利用轮廓系数（Silhouette Coefficient）S 判断聚类效果，其中：

$$S = \frac{ab}{max(ab)}$$

a 是每一个类中样本彼此距离的均值， b 是一个类中样本与其最近的那个类的所有样本的距离的均值。计算得 S=0.832，表明类之间相关性较弱，聚类效果较好。

由聚类分析得，第一类区域主要包含的城市（地区）为广州、大连、青岛、福州、宁波、南京，均为我国一线或较发达城市，其港口吞吐量都较大；地二类区域包含长沙，南宁，上海，海口，深圳，厦门，表明除港口和 GDP 外，平行进口额的一个重要制约因素为汽车保有量，在汽车保有量较饱和时，平行进口热度骤然下降；地三类区域包含重庆、郑州、北京、石家庄、成都、满洲里、乌鲁木齐，其均为内陆城市，进口能力较弱，需求较小，且处于平行进口发展初级阶段，进口数量自然较为保守，而北京机动车数量过饱和也导致其有其心而力不足。

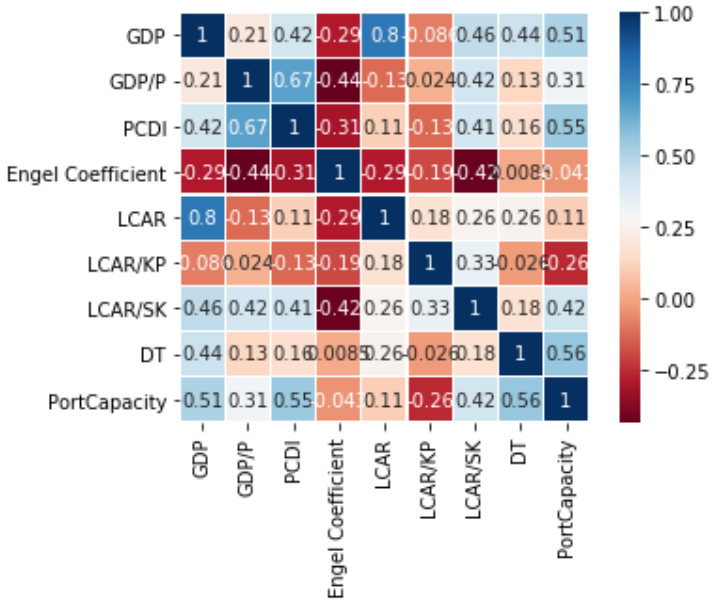
1.1.2 关联性分析

在对区域进行聚类之后，根据影响因素建模之前，首先应对变量间的关联性进行分析。

实际上，在多参数附加超参数拟合问题中（参数量级通常在 10^4 之上），为了减少变量间关联性，加快拟合速度，通常会对参数进行主成分分析（PCA），之后再进行变量间的关联性分析。考虑到本例中样本与参数数量均较少，进行 PCA 无实际意义且容易使模型过拟合，故在此只进行关联性分析。

利用 Seaborn 生成关联性热力图如下：

图 1.3 关联性分析 HeatMap



其中，GDP 为国民生产总值，GDP/P 为人均国民生产总值，PCDI 为人均可支配收入，Engle Coefficient 为地区恩格尔系数，LCAR 为汽车保有量，LCAR/KP 为千人均汽车保有量，LCAR/SK 为每平方公里汽车保有量，DT 为平行进口试点发展时间，PortCapacity 为海港吞吐量。

由图 1.3 可知，人均国民生产总值与人均可支配收入存在 67%的正相关度，可根据拟合结果酌情保留或删除；城市生产总值与汽车保有量存在高达 80%的关联度，但考虑到其他方面关联度不高，能够组合以保持较高解释率，故在此可保留全部指标作为模型拟合特征。

1.1.3 实证分析与相关检验

样本以及变量选取

本文将选取 2014-2018 年我国平行进口口岸的进口额及影响因素的数据作为样本。进口额相关数据出自《中国城市统计年鉴》、《中商产业研究院研究报告》、各省份统计局、国家商务部、国家发改委、国家经济信息中心、中国行业研究网等政府和企事业单位。

在平行进口环境制约因素分析部分，我们已经对诸如经济因素（GDP）、收入因素(PCDI)、地区进口实力（PortCapacity）、城市规模因素（LCAR/SK 和 LCAR/KP）等影响地区进口额的制约关系做了理论分析和实际验证，并且检验了相关性，达到了 83.2%的解释率。这里暂且选择国民生产总值、人均国民生产总值、人均可支配收入、地区恩格尔系数、汽车保有量、千人均汽车保有量、每平方公里汽车保有量、平行进口试点发展时间、海港吞吐量这九个指标作为自变量，时间、空间一体化研究它们对汽车平行进口额的影响程度及拟合关系。

首先，选取 2014-2018 年我国 21 个汽车整车平行进口口岸所在城市（地区）的进口额和九大影响因素的年度或半年度数据，**项目后期将增加样本数量，获得完备数据和影响因子，**通过比较不同回归模型检验模型准确率。由于平行进口本身发展时间较短，时间序列样本较少，模型极易过拟合，即时间序列应作为相关性的辅助分析工具，而主要拟合地域模型。

同时，由于东北地区汽车平行进口正处于初级阶段，在建模分析中，时间序列过短，东北地区和天津的地缘差别大，发展规模不同，且天津进口额过大使得其他城市在回归中影响较少，实际参考价值有限，所以应着重调整天津权重，或在回归中剔除天津进行建模，配合正则化以改善单变量差异过大所带来的过拟合或欠拟合现象。

当样本数较多时，过拟合问题较少出现，而对于汽车平行进口问题，相对于时间的样本数为 T=4，而相对于空间的样本数 M=21，数量级均较小，此时按类拟合更容易捕捉特征的影响，故按 1.1.1 中的聚类结果分为三类进行拟合。

模型选择和比较

在计量模型的回归问题中，除了常见的 OLS 以外，在机器学习领域还有 svm（支持向量机）回归、KNN 回归、随机森林回归、Adaboost 回归、GBRT 回归和 Bagging 回归等，这些回归方法的差异主要在于是否有核函数，即 kernel，以及是否有决策，即决策树。实际上，核函数的本质作用在于隐含了低维到高维的映射，从而可以避免直接计算高维内积；决策树则更注重单个结果对后续结果的影响。纵然 OLS 最为常见，考虑到本课题所选数据类别性强，回归性弱，应当选取弱分类器和回归模型相结合的建模方法。

同时，对一些影响因子较多的问题中，权重计算是重中之重，此时可选取层次分析模型建立线性拟合方程，通过对权重的一致性检验和建立模糊矩阵来作为拟合模

型的评价结果。然而，层次分析模型毕竟作为线性模型（实际上是多项式模型），并不能很好解释高维非线性问题。故本文拟选用 OSL、SVM 和随机森林（RF）进行回归拟合，并通过 Grid Search 来调试超参数，进而选出最佳模型。

模型拟合

通过搜集全国 21 个试点城市（地区）以及相关影响因素数据，对除天津之外这三类区域分别采取 OLS、SVM、RF 回归方法，采取 lasso 正则化，即

$$\min_w \frac{1}{n}\| y - Xw \|^2, s.t. \| w \|_1 \le C$$

其中：w 为参数矩阵，y 为预测结果，X 为变量，C 为常数。上式可与回归方程化简为：

$$\theta_j = \theta_j(1 - a\frac{\lambda}{m}) - a\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i,\ j = 1,2...m$$

其中：θ_j 为参数矩阵，λ 为正则化系数。

借助 stata14 进行 OLS 回归分析得出以下结果：

图 1.4 stata OLS 回归结果 1

Source	SS	df	MS	Number of obs	=	21
				F(10, 10)	=	17.33
Model	547294259	10	54729425.9	Prob > F	=	0.0001
Residual	31579121.9	10	3157912.19	R-squared	=	0.9454
				Adj R-squared	=	0.8909
Total	578873381	20	28943669	Root MSE	=	1777.1

ImportV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDP	-.2967542	.1292732	-2.30	0.045	-.5847928	-.0087157
GDPP	.0059553	.0125229	0.48	0.645	-.0219474	.033858
PCDI	-.1366536	.0704278	-1.94	0.081	-.2935766	.0202694
EngelCoefficient	-30704.12	12536.43	-2.45	0.034	-58637.02	-2771.224
LCAR	17.87095	8.475151	2.11	0.061	-1.012863	36.75476
LCARKP	-7.616322	10.83522	-0.70	0.498	-31.7587	16.52606
LCARSK	1.578113	1.93474	0.82	0.434	-2.732756	5.888982
DT	9.769649	418.41	0.02	0.982	-922.5059	942.0452
PortCapacity	1.472498	1.198288	1.23	0.247	-1.197454	4.142449
Emphasis	888.7086	80.34077	11.06	0.000	709.6983	1067.719
_cons	14552.15	5721.077	2.54	0.029	1804.799	27299.5

由图 1.4 知 Prob>F = 0.0001<0.05，说明所选模型具有统计意义。在显著性水平为 0.05 的条件下，有 GDP、EngelCoefficient、Emphasis 通过了显著性检验，