参赛编号：20240690074

2024年第五届“大湾区杯”粤港澳

金融数学建模竞赛

题 目　基于$ARIMA$和GAM模型的大湾区经济预测探究

摘　要

粤港澳大湾区，作为全球四大湾区之一，由九个城市及两个特别行政区组成，是中国经济发展的重要引擎，具有重要的战略意义。本文旨在通过建立数学模型，分析和预测粤港澳大湾区未来5至10年的经济发展趋势，并为政策的制定建言献策。

针对任务一，本文收集了2000年至2023年间涵盖人口、科技、教育、商贸和交通等领域的经济指标数据。采用**$Lagrange$线性插值法**填补数据中的缺失值，并进行标准化处理。在定性分析阶段，我们计算**$Spearman$相关系数**，发现GDP与工业企业研发经费、城市人口等因素之间存在显著的正相关关系。在定量分析阶段，由于**方差膨胀因子**（$VIF>10$），本文采用基于**交叉验证**的**岭回归**模型对多个影响因素进行深入分析，以有效处理多重共线性问题。分析结果表明，**城市人口**、**货物进出口货值**、**工业企业研发经费**、**高校在校学生数**和**教育经费**这5个重要因素对粤港澳大湾区的经济发展影响显著。

针对任务二，本文首先基于5个重要因素建立了**$ARIMA$模型**，预测未来5至10年的变化趋势。结果显示，受全球经济和疫情影响，大湾区人口增长将放缓，后疫情时期的生活成本上升可能导致人口流出。随后，通过**广义相加模型（GAM）**分析GDP与各变量的非线性关系，结果表明城市货物进出口和高校在校人数增加会引发GDP波动，突显经济增长的多重动态影响。模型拟合度较高，验证了变量选择的有效性。基于此，本文预测大湾区未来5至10年的GDP将显著增长，并提出多项政策建议以促进其可持续发展。

针对任务三，首先获取**东京湾区**主要数据指标，我们预测了东京湾区未来5-10年的GDP趋势，结果显示，其未来几年增速可能放缓，甚至在2029年后可能下降。之后我们**量化分析**了两湾区5个重要因素上的**异同**。结果表明，粤港澳大湾区更依赖人口和贸易增长，而东京湾区则更重视科研和高等教育投入。此外，预测显示未来粤港澳大湾区的教育经费投入增速将超过东京湾区，这将为区域经济长期发展奠定基础。

针对任务四，我们根据模型结果析了粤港澳大湾区未来5-10年的经济走势，指出经济增长主要依赖稳定的进出口贸易和研发投入，同时强调城市人口和教育对长远发展的潜在影响。为此，提出了**五项发展建议**：加大科技创新支持力度，强化全球资源整合、促进高等教育与产业融合、优化人口结构以吸引高端人才，以及增加教育和人力资本投资。这些策略旨在提升区域经济竞争力，促进可持续发展。

关键词：$Spearman$相关系数、岭回归、$ARIMA$模型、广义相加模型(GAM)、量化分析

一、 问题重述

1.1 问题背景

粤港澳大湾区是由广州、深圳、珠海、佛山、惠州、东莞、中山、江门、肇庆九市和香港、澳门两个特别行政区组成的城市群。作为全球四大湾区之一，粤港澳大湾区在金融、产业、科技等方面具有重要地位，其生产总值（GDP）约占全国经济总量的 1/9。大湾区的经济发展非常多元化，涵盖了制造业、金融业、贸易、物流、旅游等多个领域。广州和深圳是大湾区的两个核心城市，广州以其悠久的历史和文化底蕴著称，而深圳则以其快速发展的现代化城市形象闻名[1]。香港是大湾区的重要金融中心，拥有世界一流的金融服务和自由贸易港地位。澳门则以其博彩业和旅游业闻名，是一个国际知名的旅游胜地[2]。大湾区的经济发展潜力巨大，具有重要的战略意义。

1.2 问题的提出

根据问题背景以及所收集到的数据，本文将建立模型解决下述问题：

问题一：分析和评价影响粤港澳大湾区经济发展的因素。利用历史数据，建立数学模型，量化各因素对经济发展的影响，并选择影响未来 5-10 年经济走势的若干重要因素。

问题二：基于提取的主要因素，建立经济预测模型，预测未来 5-10 年粤港澳大湾区的经济走势，并设计策略和方案，促进区域的快速发展。

问题三：选取其他湾区，运用相同的分析方法或模型，预测其未来 5-10 年的经济走势，并量化分析不同湾区之间发展的异同。

问题四：根据建模分析，向决策部门提供建议。简报内容包括建模依据、未来 5-10 年湾区经济的预测及采取措施对经济变化的影响。

二、 问题分析

2.1 任务一的分析

针对任务一，我们将对收集的数据进行预处理，以确保其完整性和准确性。使用$Lagrange$线性插值法填补缺失值，并对不同单位和量纲的数据进行标准化，以便于比较。我们将应用$Spearman$相关系数进行定性分析，检查各指标与GDP的关系，尤其是对经济发展的影响。定量分析中，将运用岭回归模型探讨指标与GDP的关系，通过计算方差膨胀因子（VIF）检验多重共线性，确保模型的可靠性。岭回归的惩罚项有助于处理多重共线性问题。最终分析结果并提出相应建议，以助力大湾区的经济发展。

2.2 任务二的分析

基于提取的主要因素，我们将建立一个经济预测模型，预测未来 5-10 年粤港澳大湾区的经济走势。首先，选择合适的预测模型，如时间序列分析、回归分析或机器学习模型。然后，使用历史数据对模型进行训练和验证，确保模型的准确性。接着，利用模型对未来的经济走势进行预测，分析各主要因素对经济发展的影响。最后，根据预测结果，设计相应的策略和方案，促进粤港澳大湾区的快速发展。

2.3 任务三的分析

在任务三的模型建立与求解中，我们将分析世界四大湾区的发展差异与共同点。首先，通过查找和预处理东京湾区的城市人口、货物进出口和研发经费等数据，为经济预测奠定基础。接着，运用数学模型预测东京湾区未来 5-10 年的GDP走势，识别其经济增长表现和挑战。随后，将比较粤港澳大湾区、东京湾区在经济增长、人口流动和产业结构等方面的异同。这一量化分析旨在揭示各个湾区的竞争优势和发展瓶颈，为政策制定者提供参考，促进区域协同发展。

2.4 任务四的分析

在本次分析中，我们将撰写一份简报，重点关注粤港澳大湾区未来 5-10 年的经济预测及政策建议。简报将详细描述模型构建的依据，包括城市人口、货物进出口和研发经费等数据指标，以确保分析的全面性和准确性。同时，我们将运用定量分析方法，对各项指标进行回归分析和时间序列预测，揭示经济走势的潜在影响因素。最后，简报将提出相应的政策建议，包括增加科技投入、优化人口结构和提升物流效率等，以帮助决策部门制定应对未来经济挑战的有效策略。

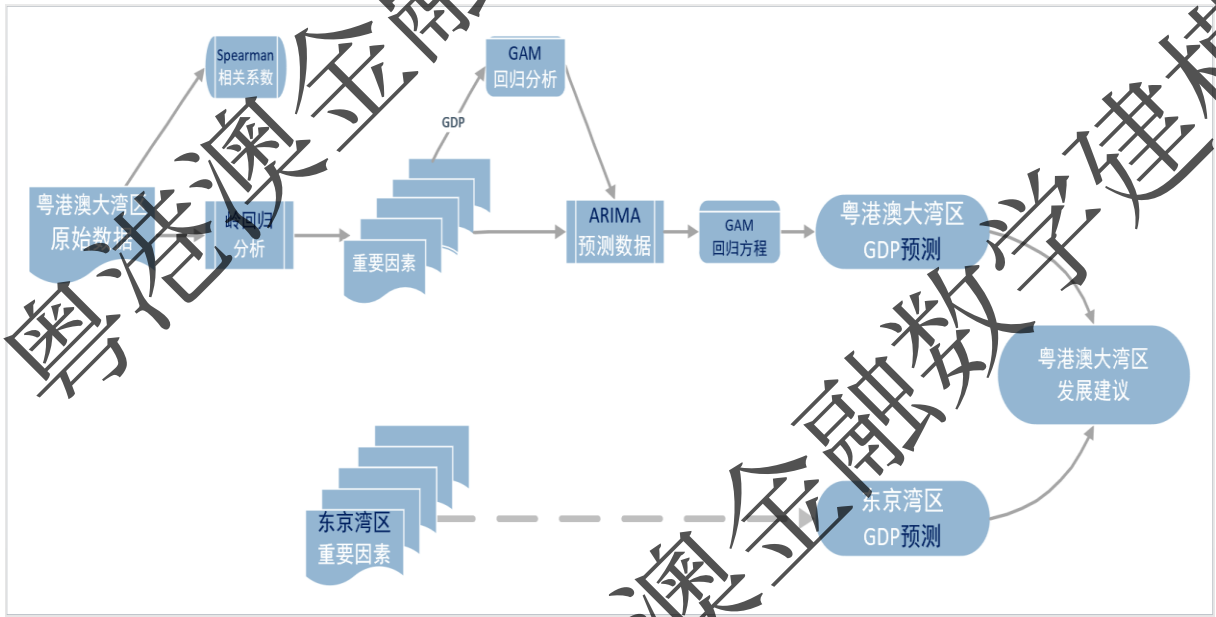


图 1 任务 1-4 流程图

三、 模型假设

1、假设在模型分析的时间范围内，粤港澳大湾区的政策环境保持稳定。

2、假设粤港澳大湾区经济增长受经济周期的影响，经济波动与外部经济环境相关。

3、假设模型中考虑的各因素在时间上是同步变化的，即自变量与因变量之间的关系在分析期间保持一致。

四、 符号说明

符号	符号意义
X_i	影响因素（粤港澳大湾区）
T_i	重要影响因素（东京湾区）
Y	生产总值（粤港澳大湾区）
S	生产总值（东京湾区）
Z	标准化影响因素（粤港澳大湾区）
λ	正则化参数
$s(x)$	特征函数

五、 模型的建立与求解

5.1 任务一的模型建立与求解

5.1.1 数据查找与预处理

在进行数据分析之前，我们需要对收集到的粤港澳大湾区的各项数据进行预处理。根据国家统计局等网站，本文从人口、科技、教育、商贸和交通运输五个方面获取2000年-2023年粤港澳大湾区的主要数据指标，包括城市人口X_1（万人）、城市就业人口X_2（万人）、0-14岁比例X_3、15-64岁比例X_4、65岁以上比例X_5、城市货物进出口货值X_6（亿美元）、外商投资企业投资总额X_7（百万美元）、技术市场成交额X_8（亿元）、工业企业研究发展$R\&D$经费X_9（万元）、普通高等学校在校学生数X_{10}（万人）、教育经费X_{11}（万元）、公路里程X_{12}（万公里）、货运量X_{13}（万吨）、生产总值Y（亿人民币），并将数据汇总于附件1和附件2，同时绘制影响粤港澳大湾区经济发展的指标体系图，如下图所示。

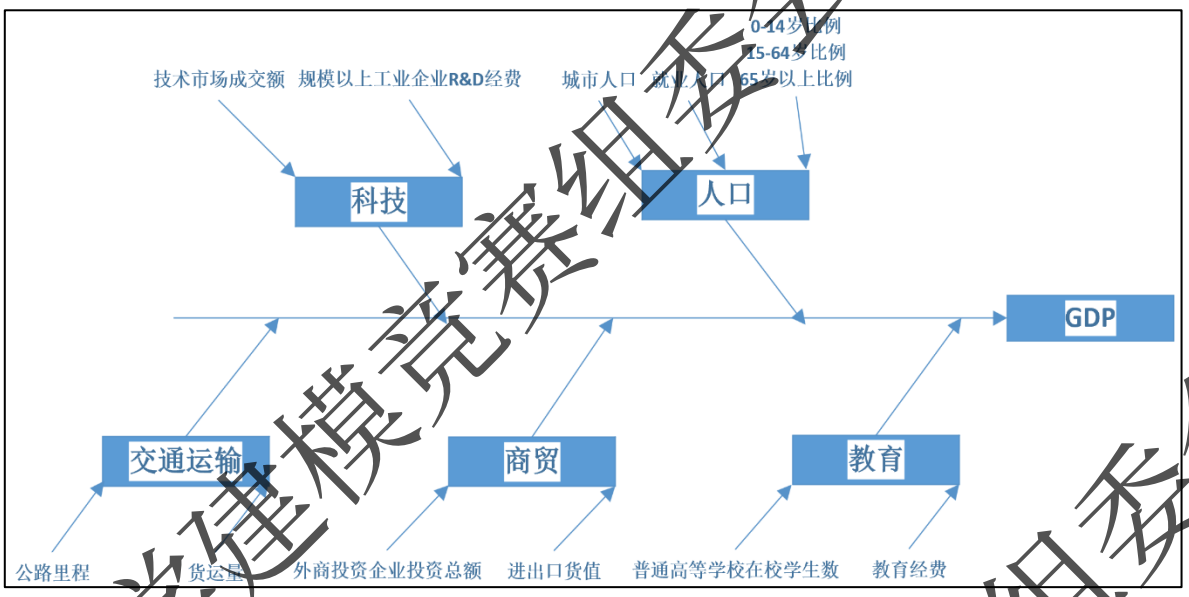


图 2 所选取影响 GDP 因素

对于附件 1 和附件 2 中的缺失值，本文结合其相邻两点的数据，采用 Lagrange 线性插值公式：

$$l_1(t) = \frac{t - t_1}{t_0 - t_1} f(t_0) + \frac{t - t_0}{t_1 - t_0} f(t_1)$$

计算出对应的插值$l_1(t)$，补足数据中所有的缺失值。

根据附件 1 的数据，本文先计算 2000 年至 2023 年大湾区 GDP 的增速，并将 GDP 及其增速绘制成图 2。

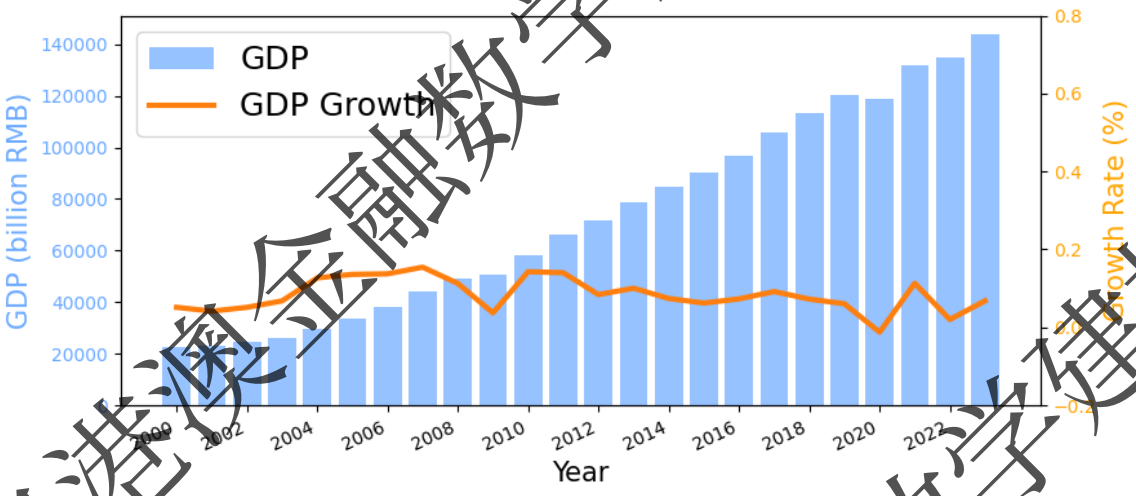


图 3 粤港澳大湾区 GDP 及增速可视化

上图显示了粤港澳大湾区 2000 年至 2023 年间的 GDP 总量及其增长率的变化趋势。在此期间，GDP 总量呈现出稳定的增长趋势，尤其在 2008 年全球金融危机后，GDP 依旧保持较快的增长，显示出大湾区经济的韧性和强劲的增长动能。而 GDP 增长率的波动幅度较大，尤其在 2008 年和 2020 年受到了外部经济环境的冲击，增长率出现显著下降，反映出全球经济形势对大湾区的影响。整体而言，尽管增长率有所波动，粤港澳大湾区

的经济规模持续扩大，显示出作为中国重要经济区域的强劲增长趋势和区域发展潜力。

为了探究大湾区经济发展与上述 13 个指标之间的关系，本文将从定性和定量两个方面进行分析。

5.1.2 经济走势影响因素的定性分析

针对附件 2 的数据，根据标准化公式：

$$z_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{n} {x_{ij}}^2}$$

对 13 个不同单位和量纲的数据指标进行标准化处理，使其在同一尺度上进行比较，形成统一的数据集，并得到标准化处理后的矩阵$Z = (Z_1, Z_2, \ldots, Z_{13})$，为后续的分析和建模提供可靠的基础。

首先本文将进行定性分析。Spearman相关系数可以用于检验非正态分布数据间的相关系数，根据Spearman相关系数计算公式：

$$r_{ij} = 1 - \frac{6\sum_{i=1}^{n} d_m^2}{n(n^2-1)}$$

其中d_m为x_{im}和x_{jm}之间的等级差。使用 python 计算附件 2 中各数据之间的相关系数，并绘制出如下的相关性分析热力图。

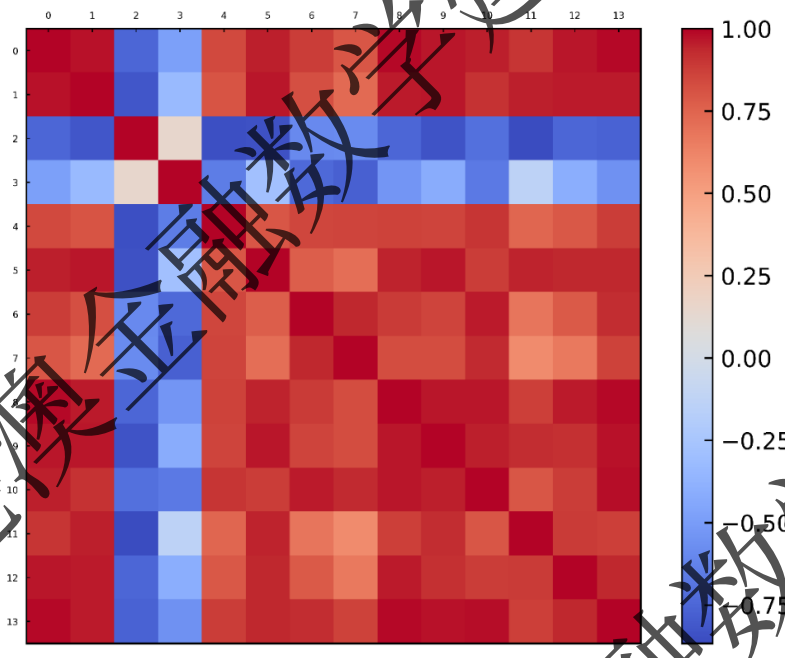


图 4 各指标相关性分析热力图

由上图结果可知，GDP 与 0-14 岁比例、15-64 岁比例呈负相关且相对较弱；与其余 11 个指标均呈现正相关，其中 GDP 与工业企业 R&D 经费相关性最强，相关系数为 0.99，与城市人口相关性次强，相关系数达到了 0.98，而与技术市场成交额相关性相对较弱，相关系数为 0.86。

5.1.3 经济走势影响因素的定量分析

随后，本文将运用岭回归来定量分析对 13 个指标与粤港澳大湾区 GDP 之间的关系：

① 首先通过 VIF 检验是否存在多元共线性：

由于本文将根据这 13 个指标来对理赔额进行回归分析，所以需先借助方差膨胀因子$VIF(Variance\ Inflation\ Factor)$来对上述 13 个自变量进行多重共线性检验。如若存在多重共线性，即某一解释变量可以由其他的解释变量线性表示，则会导致回归分析中的系数$\widehat{\beta_0},\dots,\widehat{\beta_{13}}$不可识别，从而无法定义最小二乘法。

根据第 m 个自变量的VIF计算公式：

$$VIF_m = \frac{1}{1 - R^2_{1\sim a\backslash m}}$$

其中，

$$R^2_{1\sim a\backslash m} = \frac{\sum_{i=1}^{12}(z_m - \widehat{z_m})^2}{\sum_{i=1}^{12}(z_m - \overline{z_m})^2}$$

是将第 m 个自变量作为因变量，对剩下的$(a-1)$个自变量进行线性回归后得到的拟合优度。由定义可知，回归模型的VIF值为：

$$VIF = max\{VIF_1, VIF_2, \cdots VIF_k\}$$

数据和上述公式可知，该回归模型的 VIF 值为：

$$VIF = max\{VIF_1, VIF_2 \cdots VIF_5\} = 442.1657$$

即该回归模型的$VIF > 10$，由此可得该回归模型中，存在严重的多重共线性。

② 建立岭回归模型：

岭回归$(Ridge\ Regression)$是一种用于应对多重共线性（多个自变量高度相关）的问题的回归方法[4]。在多重共线性存在的情况下，普通最小二乘法（OLS）会导致回归系数不稳定，并可能产生很大的方差，从而导致预测不准确。岭回归通过引入一个惩罚项来缩小回归系数，从而减小模型的方差，使得回归模型更加稳定。

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_{13} z_{i(13)} + \xi_i$$

该过程通过最小化以下公式来估计$(\beta_0, \beta_1, \dots, \beta_p)$的值：

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j z_{ij}\right)^2$$

岭回归与最小二乘法非常相似，不同之处在于系数是通过最小化稍微不同的量来估计的。具体来说，岭回归系数估计$\widehat{\beta^R}$，是通过最小化以下公式的值来获得的：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS\ \lambda\sum_{j=1}^{p}\beta_j^2\ ,$$

其中$(\lambda \geq 0)$是一个调节参数，需要单独确定。上述方程平衡了两个不同的标准。与最小二乘法一样，岭回归通过使RSS尽可能小来寻求较好的数据拟合系数估计。然而，第二项$\lambda\sum_{j=1}^{p}\beta_j^2$，称为收缩惩罚$(shrinkage\ penalty)$，在$(\beta_1, \beta_2 \ldots, \beta_p)$接近零时很小，因此具有将$\beta_j$的估计值向零收缩的效果。

岭回归会根据每个λ的值生成不同的系数估计 $\widehat{\beta_\lambda^R}$。选择合适的 λ 值是至关重要的，我们使用交叉验证来评估，如下图所示。

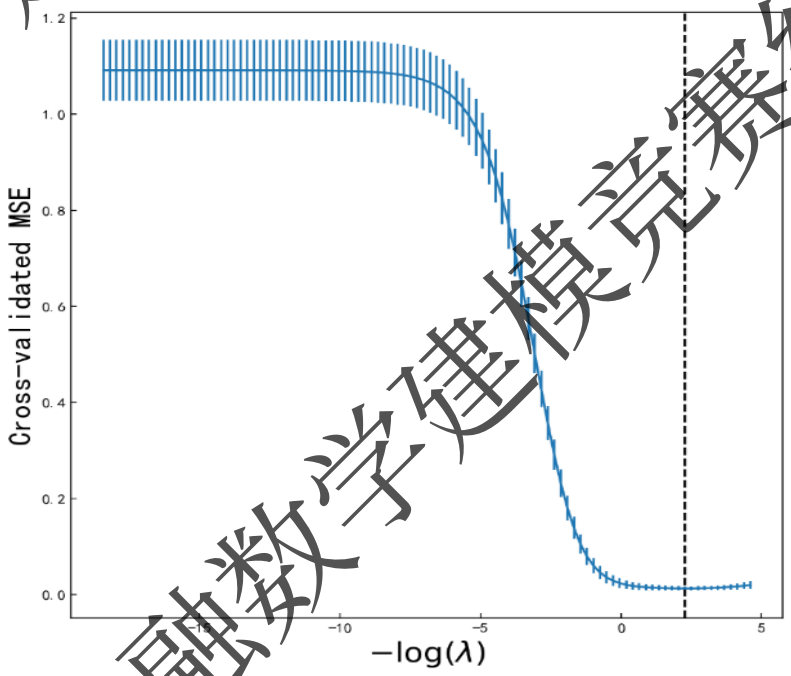


图 5 岭回归交叉验证图

上图展示了在交叉验证中，均方误差（MSE）随正则化参数λ（通过$-\log(\lambda)$转换后的表现）变化的趋势，用于选择最佳的正则化强度。横轴显示$-log(\lambda)$，其中λ是正则化参数。通常，较小的λ值对应较少的正则化，而较大的λ值对应较强的正则化。当$-log(\lambda)$增较大（向右）时，表示λ值较小，正则化强度较低。纵轴显示了交叉验证的均方误差（MSE），越小表示模型在验证集上的预测表现越好。每个点上的垂直误差棒较短的区域表示模型预测误差较稳定，而误差棒较长的区域则表示预测误差波动较大。在图的左侧$-log(\lambda)$较小，即λ较大，MSE保持较高且稳定，模型正则化过强，限制了模型的复杂度，导致欠拟合。随着λ的减小，$-log(\lambda)$增加，MSE逐渐下降，模型的拟合效果得到改善。在$-log(\lambda)\approx 0$ 附近，MSE达到最小值，此时模型在交叉验证中的表现最佳。虚线通常表示最佳λ的位置，即交叉验证过程中误差最低时对应的λ值。在这个λ下，模型在偏差与方差之间达到了较好的平衡，即既不会欠拟合，也不会过拟合。

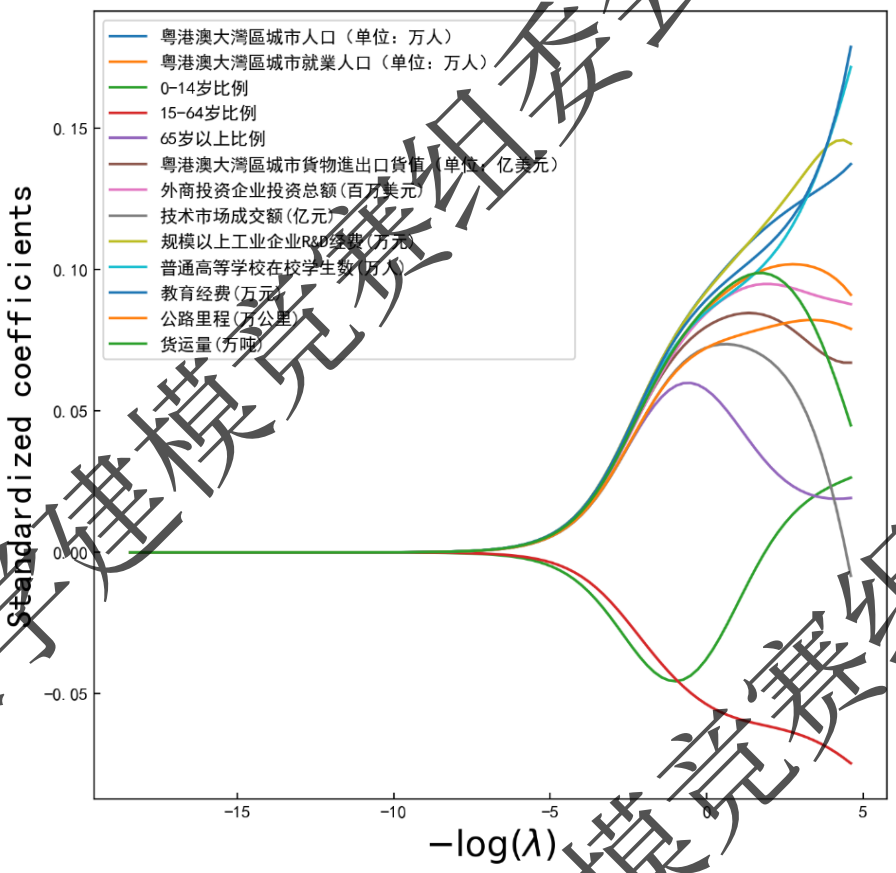


图 6 岭回归结果图

根据上图岭回归的结果，我们对影响粤港澳大湾区经济发展的各项指标及其系数进行了总结。城市人口与城市就业人口的正系数表明，人口增长和就业机会的增加对经济发展有积极影响。0-14 岁比例和 15-64 岁比例的负系数暗示，青少年和劳动年龄人口比例过高可能对经济发展产生一定压力，65 岁以上比例的正系数说明，老龄化社会的影响可能会带来一定的经济负担。工业企业 R&D 经费和技术市场成交额的正系数表明，技术创新与研发投入是推动经济增长的重要因素。这些结论与 5.1.2 中定性分析的结果基本一致，因此认为本文的模型效果良好。

由上述模型，可得标准化后 13 个指标的岭回归系数，如下表所示。

表 1 标准化后的岭回归系数

指标	岭回归系数	指标	岭回归系数
城市人口	0.000086	技术市场成交额	0.000075
城市就业人口	0.000084	工业企业 R&D 经费	0.000086
0-14 岁比例	-0.000066	高校在校学生数	0.000085
15-64 岁比例	-0.000048	教育经费	0.000085
65 岁以上比例	0.000077	公路里程	0.000076
货物进出口货值	0.000082	货运量	0.000082
外企投资总额	0.000080		

由上表可知城市人口与工业企业 R&D 经费对粤港澳大湾区 GDP 影响最大，标准化后岭回归系数均达到了 0.000086，即城市人口和工业企业 R&D 经费每增加 1，GDP 将增加 0.000086。选取其中标准化后岭回归系数最大的 5 个指标，由于排名第 5 的城市就业人口与城市人口同属于人口方面的因素，因此为了考虑得更全方面，本文选取排名第 6 的指标货物进出口货值代替城市就业人口。最终选取城市人口、工业企业 R&D 经费、高校在校学生数、教育经费以及货物进出口货值作为影响未来 5-10 年的大湾区的经济走势的重要因素，并将这 5 个重要因素的数据汇总于附件 3。

5.2 任务二的模型建立与求解

5.2.1 五个重要因素预测模型的分析与求解

以城市人口因素为例，本文优先考虑通过建立$ARIMA(p,k,q)$时间序列模型[3]：

$$\Delta^k x_{1n} = \xi_0 + \sum_{j=1}^{p} \xi_j \Delta^k x_{n-j} + \varepsilon_n + \sum_{j=1}^{q} \eta_j \Delta^k \varepsilon_{n-j}$$

来预测粤港澳大湾区的城市人口未来 5-10 年的数量。$ARIMA$模型，是一种结合了自回归（AR）和滑动平均（MR）的时间序列预测方法。它通过差分处理使非平稳时间序列数据变得平稳，然后利用AR和MR部分来建模数据的自相关性和随机干扰。

① 自回归$AR(p)$部分：模型中的$\xi_0 + \sum_{j=1}^{p} \xi_j \Delta^k g_{t-j}$代表了时间序列中前$p$期的观测值对当前期的影响。

② 差分阶数（k）：若时间序列$\{g_t\}$满足以下条件：

$$\begin{cases} E(x_{1n}) = E(x_{1(n-s)}) = u \\ Var(x_{1n}) = Var(x_{1(n-s)}) = \sigma^2 \\ Cov(x_{1n}, x_{1(n-s)}) = \gamma_s \end{cases}$$

均值u为固定常数，方差存在且为常数，协方差只与间隔s有关，与t无关，则称$\{g_t\}$为协方差平稳，又称弱平稳，此时取$k=0$。如果时间序列数据不平稳，需要进行平稳化处理。不平稳的时间序列根据k阶差分公式：

$$\Delta^k x_{1t} = (1-L)^k x_{1t}，\text{其中} L^i x_{1t} = x_{1(t-i)}$$

每进行一阶的差分处理后，计算数据的自相关系数（ACF）和偏自相关系数（$PACF$），直至进行k阶差分处理后的ACF和$PACF$的平均值趋近于 0，这样的时间序列数据是平稳的。其中样本的ACF计算公式为：

$$r_s = \frac{\sum_{j=16+1}^{M} (x_{1j} - \bar{x})(x_{1(j-s)} - \bar{x})}{\sum_{j=16+1}^{M} (x_{1j} - \bar{x})^2}$$

③ 移动平均（MR）部分：模型中的$\varepsilon_n + \sum_{j=1}^{q} \eta_j \Delta^k \varepsilon_{n-j}$代表了前$q$期的随机误差项对当前期的影响。

④ 常数项（ξ_0）：c是模型的截距项，代表当所有其他变量为零时的预期值。

⑤ 随机误差项（ε_t）：ε_t代表了不可预测的随机扰动，它应该是一个白噪声序列，其自相关系数为：

$$r_s = \begin{cases} 0, & s \neq 0 \\ 1, & s = 0 \end{cases}$$

估算完成$ARIMA$时间序列模型后，需要对残差进行白噪声检验（Q检验），分别设定原假设和备择假设：

$$H_0: r_1 = r_2 = \cdots = r_s = 0, \quad H_1: r_i (i = 1,2,\ldots,s) \text{至少有一个不为} 0$$

在H_0成立的条件下，构造统计量Q：

$$Q = M(M+2) \sum_{j=1}^{16} \frac{r_j^2}{M-j} \sim \chi^2_{16-n}$$

其中$n = p + q + 1$表示模型中的未知参数的个数。如果检测得出残差为白噪声，则说明该模型能够完全识别出时间序列数据的规律，即该模型可被接受；如果残差不为白噪声，则说明还有部分信息没有被模型识别，需要进行模型的修正来识别这一部分信息

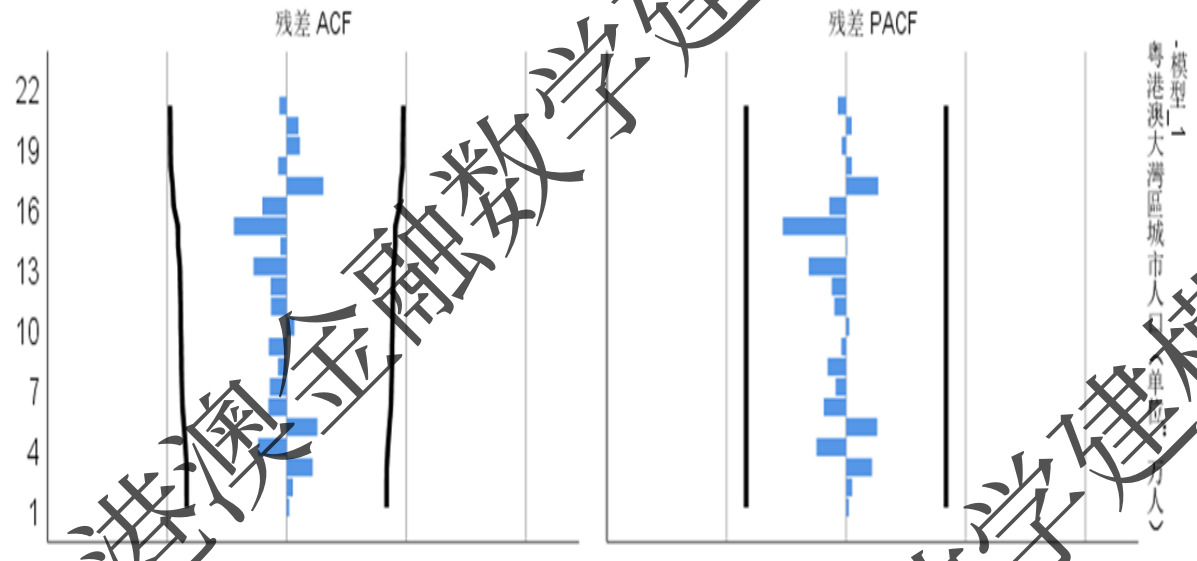


图 7 大湾区过去 24 年每年城市人口时序的残差 ACF 和 PACF

根据附件 3，对粤港澳大湾区城市人口进行残差的 ACF 和 PACF 分析，得到对应的残差图，如图 7 所示。分析图 5，过去 24 年每年城市人口时序的残差值均没有超出阈限范围的，且其白噪声检验的显著性分析值$p = 0.318$，本文认为在显著性水平为 95%的检验中设定的阈限之内可以接受原假设，因此认为其预测结果良好。

通过 SPSS 软件的运行计算，得到的最佳预测模型为$ARIMA(1,2,0)$，将预测结果汇总到附件 4 中，绘制其时间序列预测曲线图，并进行分析。

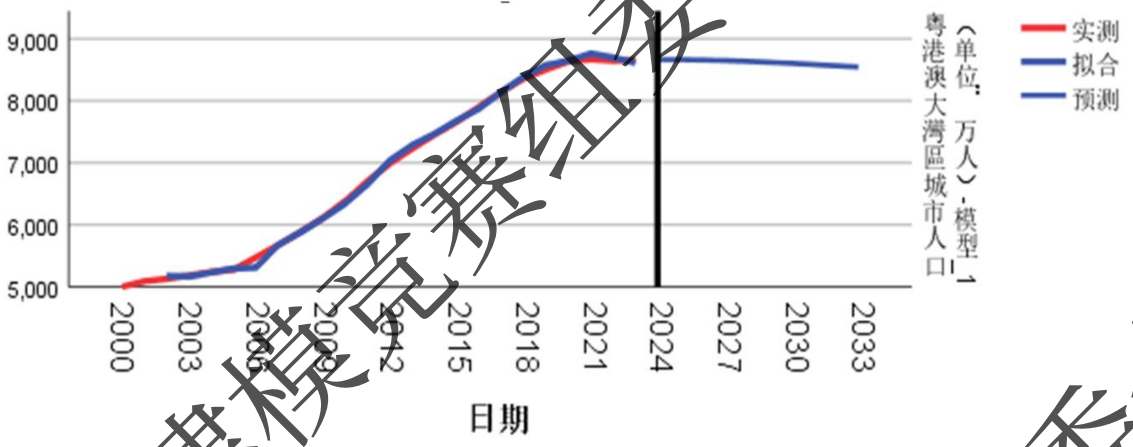


图 8 基于$ARIMA$模型的未来 5-10 年每年城市人口的数量预测曲线图

分析图 8，在经历了 2006 年到 2020 年大湾区城市人口的数量飞速增长后，由于疫情的影响，城市人口的数量在近三年略有减少，并且在未来 5-10 年，$ARIMA(1,2,0)$模型的预测曲线也呈现出缓慢下降的趋势，这一现象与当前全球“后疫情时期”经济环境下行以及社会生活成本逐渐提高的大背景下，城市人口数量减少的必然结果相对应。

其余 4 个重要因素，基于$ARIMA$模型的未来 5-10 年每年的预测值均存放于附件 4 中。

5.2.2 GDP 回归分析模型的分析与求解

在根据 5.1.3 中所选出来的五个重要因素对粤港澳大湾区 2000 年-2023 年每年的 GDP 总量进行回归分析时，城市人口、货物进出口货值等自变量与 GDP 因变量的关系尚不明确。而当自变量与应变量间关系不明确时，通常可以使用广义相加模型来检测比变量间是否具有非线性关系。因此，本文考虑采用广义相加模型对 GDP 进行回归分析。

广义相加模型$(Generalized\ additive\ models,\ GAM)$，通过光滑样条函数、核函数或者局部回归光滑函数，对变量进行拟合。GAM采用模型中的每个预测变量并将其分成多个部分，然后将多项式函数分别拟合到每个部分。GAM的原理是最小化残差（拟合优度）同时最大化简约性（最低可能自由度）。回归模型中部分或全部的自变量采用平滑函数，降低线性设定带来的模型风险。

对于上述五个重要因素，相比于普通的多元线性回归模型：

$$y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i6} + \gamma_3 x_{i9} + \gamma_4 x_{i(10)} + \gamma_5 x_{i(11)} + \epsilon_i$$

为了实现每个因素与 GDP 之间的非线性关系[4]，本文需要将每个线性分量$\gamma_j x_{ij}$替换为非线性平滑函数$f_j(x_{ij})$。然后，建立GAM多元回归非线性方程：

$$y_i = \gamma_0 + f_1(x_{i1}) + f_2(x_{i6}) + f_3(x_{i9}) + f_4(x_{i(10)}) + f_5(x_{i(11)}) + \epsilon_i$$

其中$f_1(x_{i1}),\ldots,f_5(x_{i(11)})$分别为城市人口、货物进出口货值等五个因素与因变量 GDP 之间的非线性平滑关系函数。本文选取平滑样条函数[4]：

$$f_j(x_{ij}) = argmin_{g_j(x)}\left(\sum_{i=1}^{n}(y_i - g_j(x_{ij})^2 + \lambda\int[g_j''(t)]^2 dt\right)$$

来描述各个重要因素与 GDP 之间的关系。

根据 GAM 模型得出各自变量与 GDP 的关系曲线，并绘制成曲线图，如下图所示。

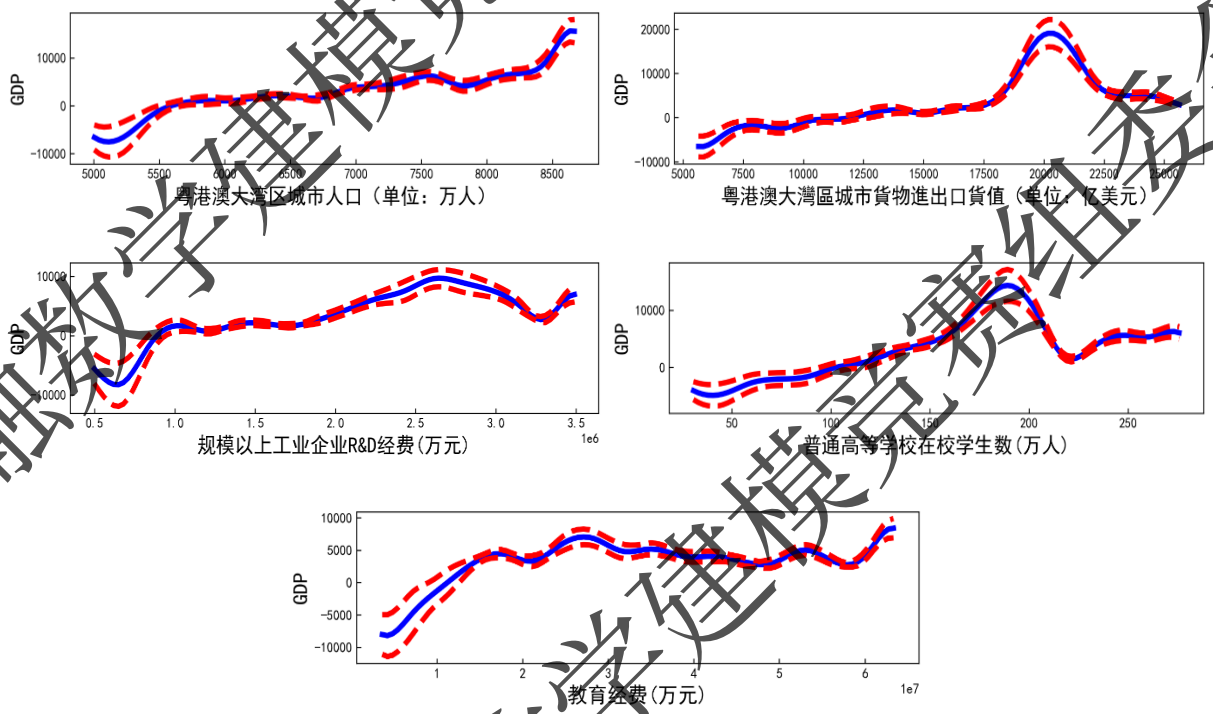


图 9 GAM 模型下各自变量与 GDP 的关系曲线图

根据图 9，可以清晰地得到以下结论：

① 红色虚线代表 95%的置信区间。

② 随着城市货物进出口货值和高校在校人数的增加，大湾区的 GDP 呈现出较为明显的波动性。

③ 城市人口和教育经费与大湾区 GDP 呈现出正相关关系。具体来说，城市人口和教育经费越多，GDP 往往越高。

④ 城市货物进出口货值和高校在校人数对 GDP 的影响则呈现一种先增后减的模式。随着城市货物进出口货值和高校在校人数的增加，大湾区 GDP 起初上升，但当城市货物进出口货值和高校在校人数分别超过 21000 亿美元和 190 万人时，GDP 开始急剧下降。

运行 python 程序求解拟合回归的非线性方程，得到特征函数s，相关结果如下图所示：

表 2 特征函数信息表

Feature Function	*Lambda*	*Rank*	$P > x$	*Sig.Code*
$s(0)$	[0,6]	20	1.11e-16	***
$s(1)$	[0,6]	20	1.11e-16	***
$s(2)$	[0,6]	20	1.11e-16	***
$s(3)$	[0,6]	20	1.11e-16	***
$s(4)$	[0,6]	20	1.11e-16	***
intercept		1	7.33e-15	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

由上表可知：

- 特征函数 s(0)的 P 值为 1.11e-16，非常显著（***）。
- 特征函数 s(1)的 P 值为 1.11e-16，非常显著（***）。
- 特征函数 s(2)的 P 值为 1.11e-16，非常显著（***）。
- 特征函数 s(3)的 P 值为 1.11e-16，非常显著（***）。
- 特征函数 s(4)的 P 值为 1.11e-16，非常显著（***）。
- 截距项的 P 值为 7.33e-15，非常显著（***）。

因此，上述五个自变量对模型的预测有重要贡献，这也证明了 5.1.3 中我们所选的影响因素是相对来说最重要的。另外，截距项也非常显著，这表明在没有其他特征的情况下，模型对目标变量的平均预测值是非常重要的。

另外，由拟合优度的计算公式：

$$R^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} \text{(k 为自变量的个数)}$$

其中

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

为误差平方和，

$$SST = \sum_{i=1}^{n} (y_i - \overline{y_i})^2$$

为总体平方和。

将相关数据代入上述公式得到调整后的拟合优度为：

$$R^2 = 0.9656$$

由此可知，调整后的拟合优度接近于 1，说明该非线性回归模型拟合程度高，效果较好。

5.2.3 粤港澳大湾区的未来经济走势的预测模型

根据 5.2.1 和 5.2.2 的两个子模型的分析，本文将基于这两个子模型，建立对粤港澳大湾区的经济预测模型，如下图所示。

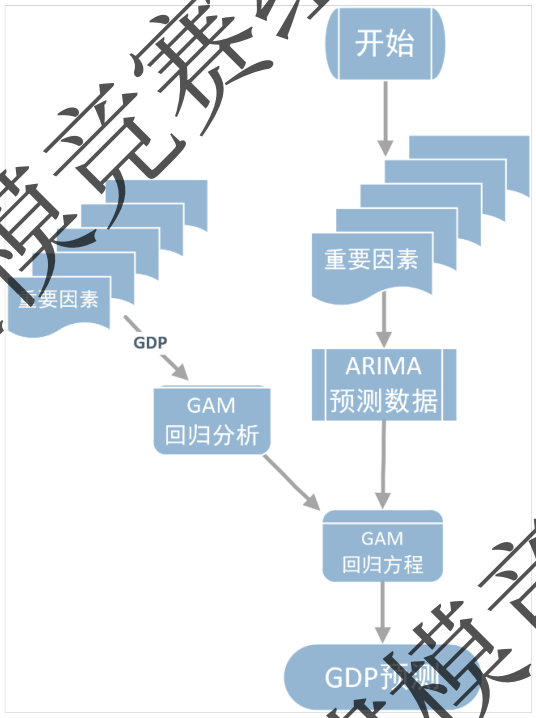


图 10 粤港澳大湾区 GDP 预测流程图

结合 5.2.1 所得到的粤港澳大湾区未来 5-10 年的 5 个重要因素的预测值，以及 5.2.2 所得到的 GAM 回归方程，可以得到大湾区未来 5-10 年的 GDP 预测值，将预测结果汇总于附件 4 中，并绘制预测曲线，如下图所示。

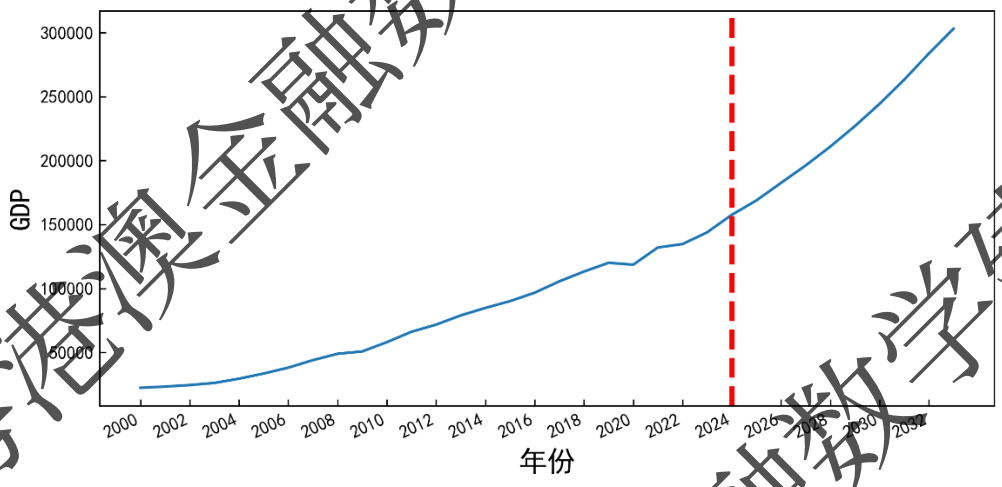


图 11 粤港澳大湾区 GDP 预测图

在上图中，从 2000 年到 2023 年，粤港澳大湾区的 GDP 呈现稳步上升的态势，表现出该区域经济基础的稳固和增长潜力。根据模型的预测，2024 年后的 GDP 增长呈现出显著的加速趋势，GDP 增长曲线趋向陡峭，这表明该区域未来十年内可能经历较快的经济扩张。这种强劲的增长预期可能受到多种因素的推动，包括大湾区经济一体化、科技创

新的加速、以及国际和国内投资的不断注入。尤其是随着政策支持和基建升级，大湾区作为全球领先的经济和金融中心的潜力将进一步释放。投资者可依据这种经济增长的预期，前瞻性地评估在该区域的资产配置机会，尤其是在房地产、金融服务、高科技产业等领域可能带来的投资机会。

5.2.4 粤港澳大湾区未来的策略制定方案

根据上述模型分析，本文将提出以下建议方案，以促进未来大湾区的经济快速发展：

① 优先加大科技创新支持投入力度。

② 优先加大高等教育和人力资本投资。

③ 强化高等教育与社会产业的联动，以促进大湾区经济的良性循环发展。

④ 推动国际贸易和物流枢纽建设。

⑤ 优化人口结构，吸引全球人才，逐步提升劳动力素质。

5.3 任务三的模型建立与求解

5.3.1 数据查找与预处理

世界四大湾区包括粤港澳大湾区、东京湾区、纽约湾区以及旧金山湾区，各自的地理区位如下图所示。

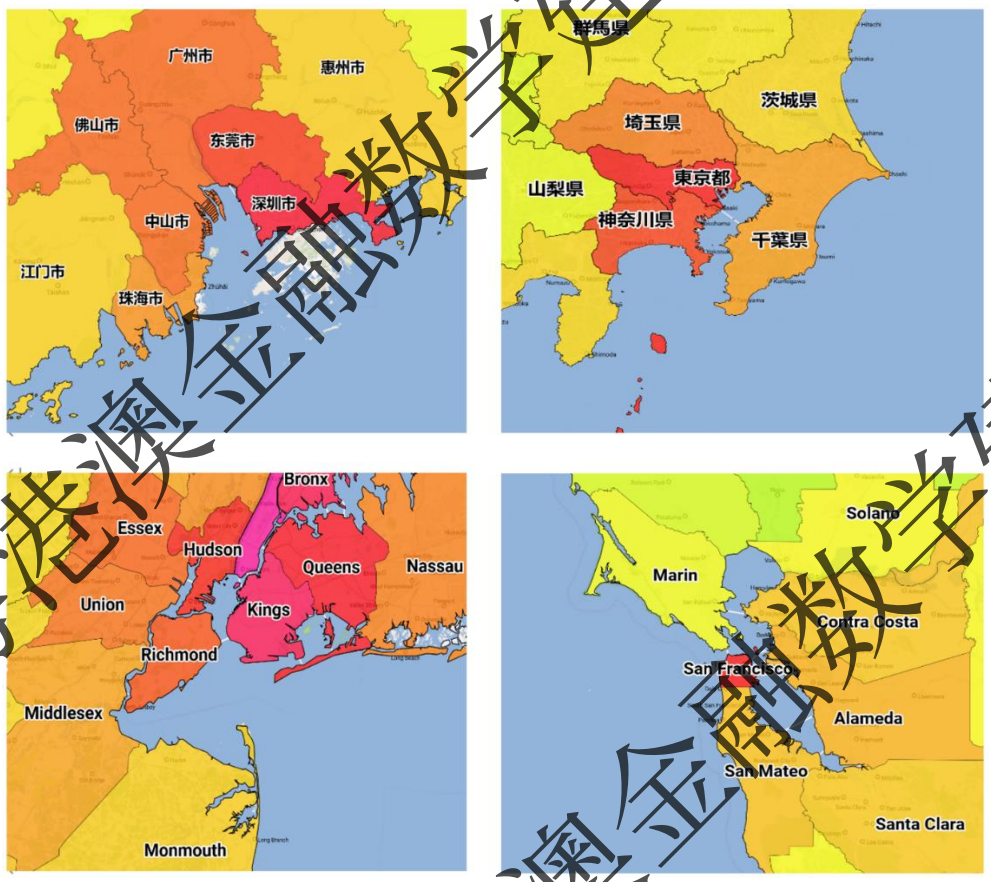


图 12 世界四大湾区示意图

其中东京湾经济总量在 2018 年就达到了 1.8 万亿美元，GDP 比重占全国 41.4%，并且在 2020 年中国社科院对四大湾区经济影响力排名中，得分 0.634，排名第一[5]。本文通过东京统计年鉴等网站，获取 2000 年-2023 年东京湾区的主要数据指标，包括城市人口T_1（万人）、城市货物进出口货值T_2（亿美元）、外工业企业研究发展$R\&D$经费T_3（万元）、普通高等学校在校学生数T_4（万人）、教育经费T_5（万元）、生产总值S（亿人民币），并将数据汇总于附件 5 中。同时根据 5.1.1 的步骤，对上述数据进行预处理。

5.3.2 东京湾区的未来经济走势的预测模型

根据 5.2.3 所运用的数学模型，本文将利用附件 5 的数据，对东京湾区未来 5-10 年的 GDP 进行预测，所得的预测数据均汇总于附件 6 中，并绘制 GDP 的预测曲线，如下图所示。

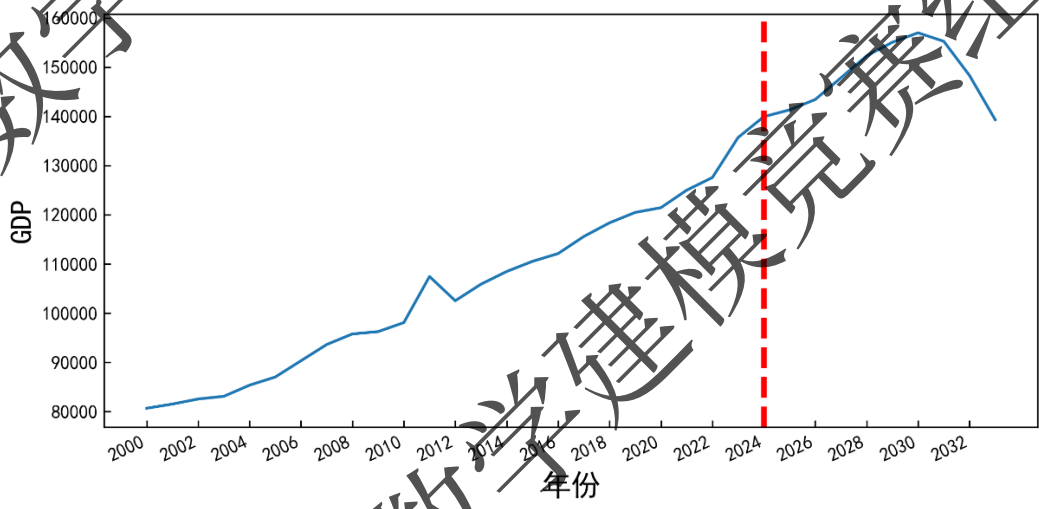


图 13 东京湾区的未来经济走势的预测图

上图展示了东京湾区 GDP 的历史数据与未来十年内的预测趋势，其中在 2000 年至 2023 年期间，东京湾区 GDP 表现出整体上行的增长轨迹，显示了东京湾区作为一个成熟经济体的持续增长和抗风险能力。根据模型的预测，东京湾区的 GDP 增速在未来几年中有放缓的趋势，尤其是 2029 年之后，增长逐渐趋于平缓甚至出现下降，这一变化预示着东京湾区可能在未来面临经济增速的结构性挑战。影响因素可能包括老龄化人口等；同时，随着亚太地区其他经济体快速崛起，东京湾区在全球市场中的相对地位也可能受到影响。

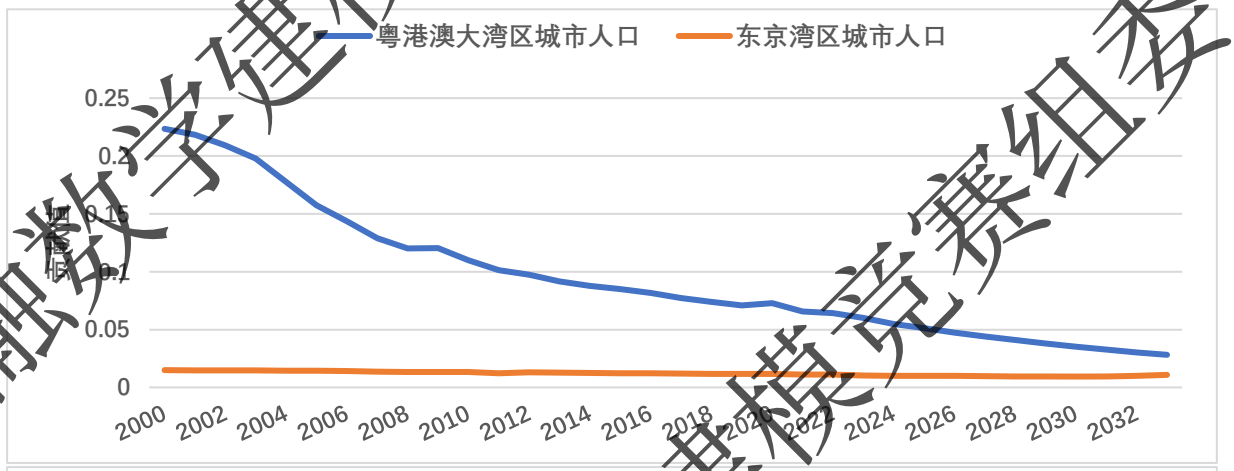
5.3.3 不同湾区发展的异同的量化分析

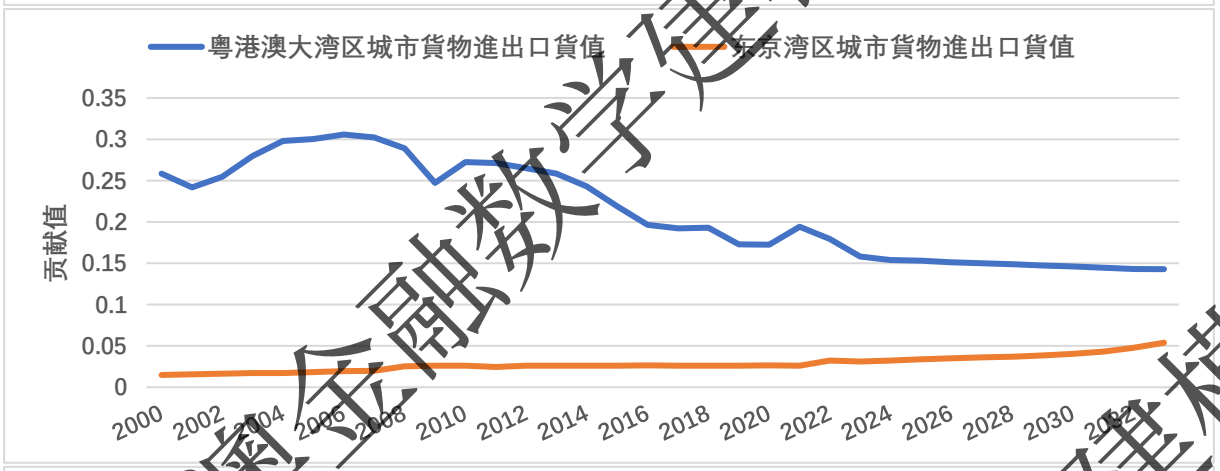
根据附件 3-6 的数据，由于粤港澳大湾区和东京湾区在大多数年份的 GDP 规模体量区别较大，因此为了合理地量化分析这两个湾区发展的异同，本文先根据公式：

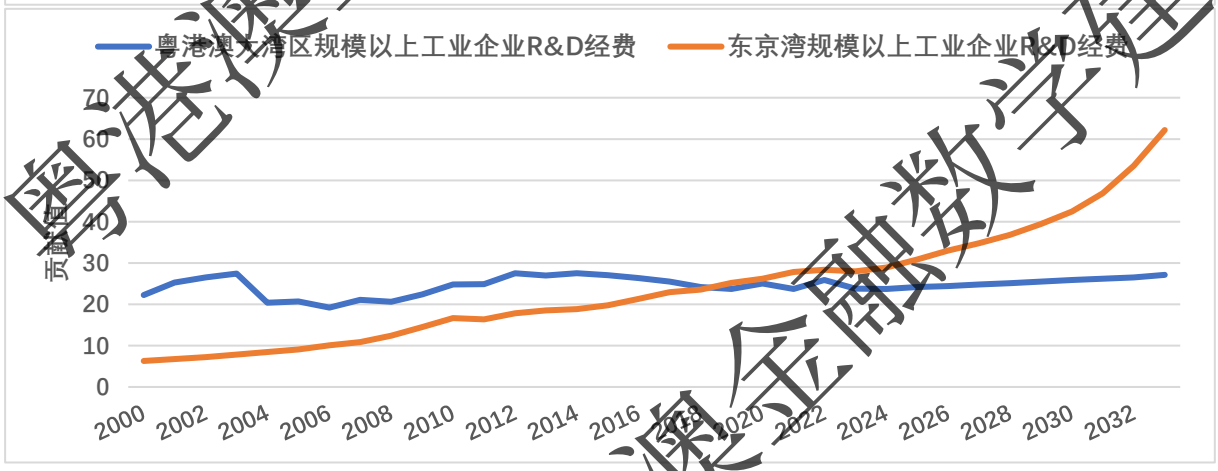
$$\sigma_{ij}^{(X)} = \frac{x_{ij}}{y_j}$$

$$\sigma_{ij}^{(T)} = \frac{t_{ij}}{s_j}$$

分别计算粤港澳大湾区 5 个重要影响因素对 GDP 每一年的贡献值$\sigma_{ij}^{(X)}$，以及东京湾区 5 个重要影响因素对 GDP 每一年的贡献值$\sigma_{ij}^{(T)}$，并将这些计算结果汇总于附件 7 中。对于每一个重要因素，将两个湾区从 2000 年到 2033 年对于各自地区 GDP 的贡献值进行两两对比分析，并绘制如下 5 个重要因素的贡献值曲线图。







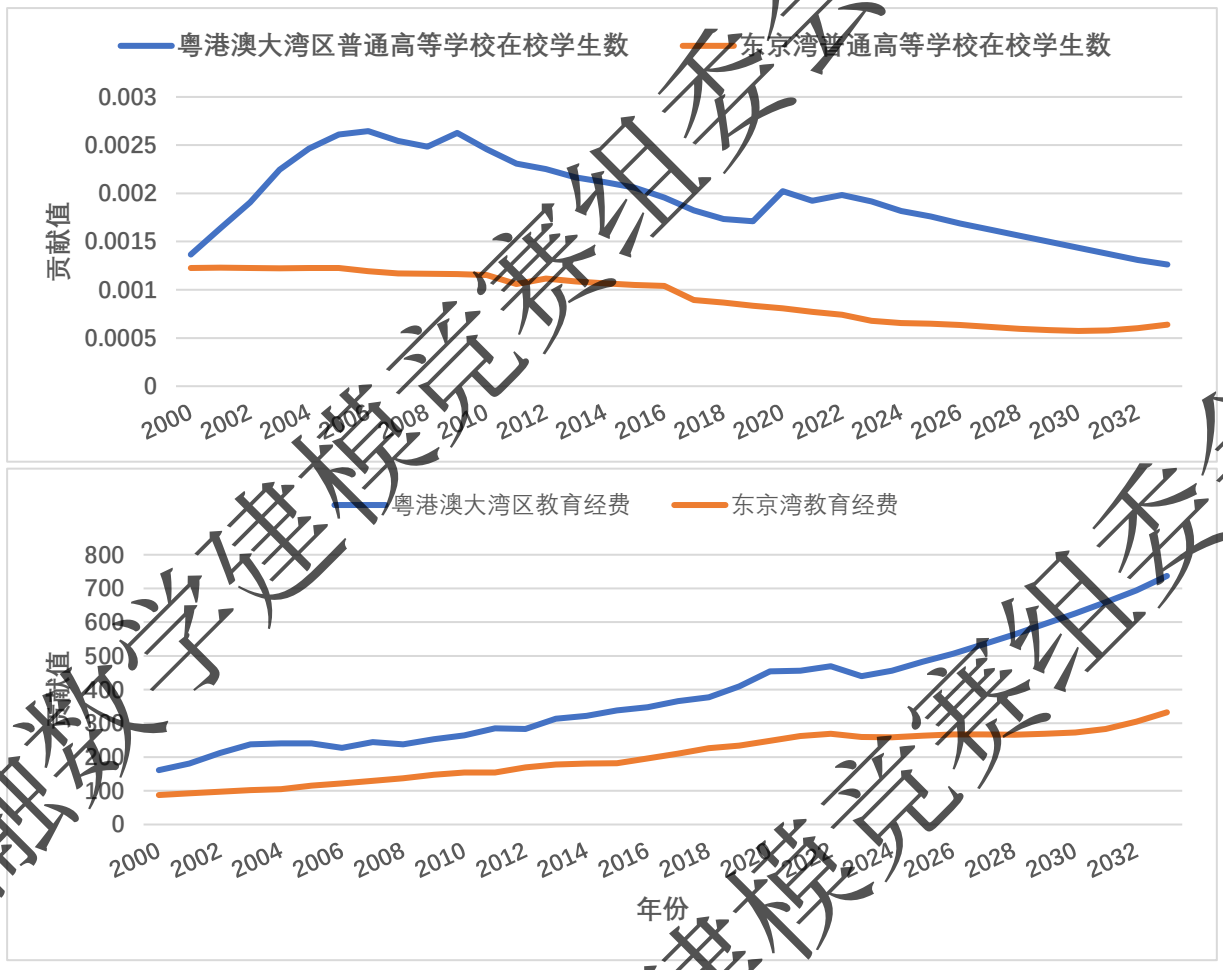


图 14 GDP 增长重要因素的贡献值曲线图

由上图可知：

① 在城市人口方面，长期以来粤港澳大湾区该因素对GDP的贡献值都高于东京湾区的，尤其是在 2016 年以前，远高于东京湾区。这表明粤港澳大湾区更加依赖于人口红利所带来的经济发展，而东京湾区则更注重劳动人口的素质而非数量。

② 在城市进出口货物值方面，长期以来粤港澳大湾区该因素对GDP的贡献值也均高于东京湾区的。这表明坐拥“千年商都”广州和国际金融交易中心香港的粤港澳大湾区在商业贸易方面具备先天的优势，相比东京湾区对GDP能产生更大的贡献。

③ 在工业企业$R\&D$经费方面，两个湾区该因素对GDP的贡献值相近，然而在 2020 年之后，东京湾区在该方面的经费投入开始了显著地增长，而粤港澳大湾区在这方面的投入则可能陷入长期不变的状态。这表明，东京湾区在当下以及将来将更加注重科研方面投入，以提高自身的GDP。

④ 在普通高校在校人数方面，与城市人口因素类似，粤港澳大湾区该因素对GDP的贡献值都高于东京湾区的。这体现了粤港澳大湾区在高等教育方面还是有着更高的投入。

⑤ 在教育经费方面，两个湾区均呈现了逐年上涨的趋势，且在 2024 年后，粤港澳大湾区在教育方面的投入增速相比东京湾区明显提升。这表明了，在未来 10 年里，粤港澳大湾区在教育方面的注重程度要逐渐高于东京湾区在这方面的重视程度。

5.4 粤港澳大湾区发展建言

粤港澳大湾区作为国家级战略经济体，已成为全球瞩目的经济增长极。本简报对未来 5-10 年大湾区的经济走势进行展望，结合对比东京湾区的发展特性，力图为大湾区的高效发展提供数据支持与战略建议。

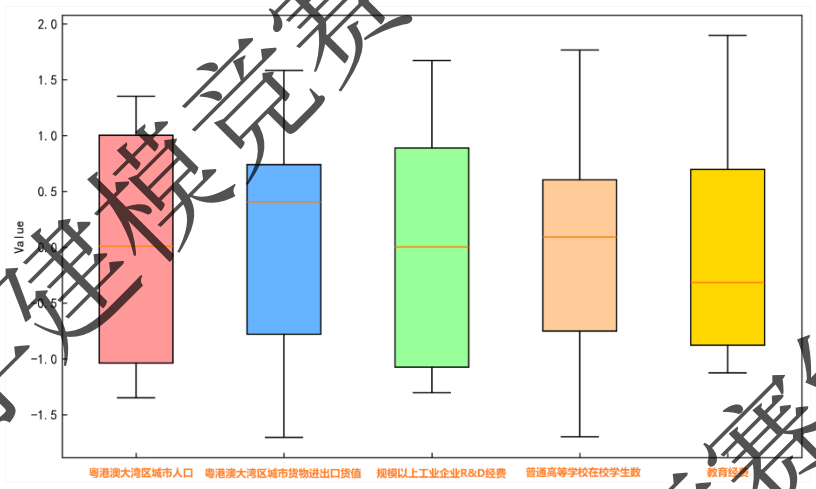


图 15 粤港澳大湾区各重要因素箱型图

上图显示，货物进出口货值对GDP的正向影响较为稳定，是大湾区经济的重要驱动力；$R\&D$经费分布较为适中，反映出研发投入在推动经济创新中作用均衡。城市人口对GDP的影响波动较大；而高等教育在校学生数和教育经费短期内对GDP的直接影响有限，但在长远的人才储备和经济质量提升上具有潜在的正向作用。总体而言，各因素在大湾区的经济发展中各具特色，短期内主要依赖进出口贸易和研发投入，而教育相关因素的长期效应有待进一步释放。

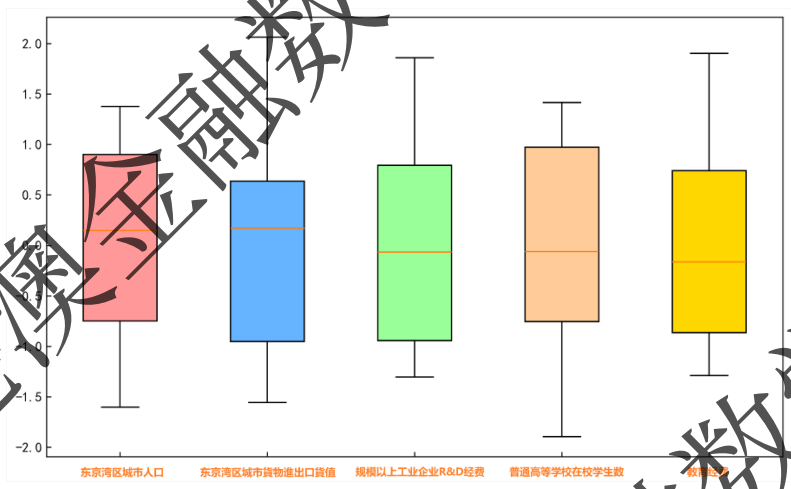


图 16 东京湾区各重要因素箱型图

由上图可知，城市人口对GDP的影响波动较大；货物进出口货值的分布较为集中表明其对GDP的影响较为稳定，与东京湾区成熟的贸易体系相符；$R\&D$经费显示对经济的正向促进作用，是区域创新的关键驱动力之一；在校学生数显示出高等教育对短期经济拉动有限，但具备长期人力资本储备的潜力；教育投入对短期GDP的直接影响不确定性

较大，但有望在未来产生更深远的经济效益。总体而言，东京湾区短期经济增长依赖于稳定的贸易和研发投入，而人口和教育因素则为其长远发展奠定了基础。

因此本文将提出以下的发展建议：

① **加大科技创新支持力度，构建国际创新高地**

粤港澳大湾区应继续加大对高新技术企业的研发支持力度。通过完善的政策引导和财政支持，尤其是在人工智能、新能源等领域，助力大湾区打造全球领先的创新中心。

② **加强全球资源整合，推动对外贸易和供应链效率提升**

粤港澳大湾区应继续强化作为国际贸易和物流枢纽的功能，增强供应链的弹性和抗风险能力。通过打造智能化物流系统和提升进出口货物的附加值，大湾区可以进一步巩固其作为全球贸易重要节点的经济优势。

③ **促进高等教育与产业链深度融合，完善人才培养体系**

为满足高新技术产业的用人需求，粤港澳大湾区可进一步推动教育体系与产业链的深度融合。通过提供跨区域的人才引进政策，大湾区能够有效提升区域内人力资本质量，为创新经济提供有力支撑。

④ **优化人口结构，吸引全球高端人才**

粤港澳大湾区作为中国和亚太地区的经济增长极，应通过一系列优厚的政策，吸引更多高端人才和国际企业入驻。

⑤ **加大教育和人力资本投资，奠定长期发展基础**

教育经费投入是大湾区长期经济增长的重要保障。大湾区应继续加大在高等教育和职业教育领域的投入，为区域内的创新驱动和可持续发展提供扎实的基础。

六、 模型评价

6.1 模型优点

数据驱动：模型基于粤港澳大湾区的历史数据，采用了定量分析和定性分析相结合的方法，确保了模型的科学性和可靠性。

多元因素分析：通过岭回归和广义相加模型(GAM)，模型能够有效处理多重共线性问题，并捕捉变量之间的非线性关系，使得对经济发展的影响因素分析更加全面。

可预测性：利用$ARIMA$模型进行未来 5-10 年经济走势的预测，结合重要经济因素，提供了对区域经济发展的清晰前景。

6.2 模型缺点

数据依赖性：模型的准确性高度依赖于历史数据的质量和全面性，任何缺失或不准

确的数据都可能影响最终预测结果。

假设限制：岭回归和GAM等模型依赖于特定的假设条件（如线性关系和正态分布），在实际应用中可能存在偏差。

复杂性：多元回归和非线性模型的复杂性可能导致模型的解释性降低，决策者在实际应用时可能难以理解模型输出。

6.3 模型改进

数据集扩展：未来可以考虑引入更多维度的数据，如环境因素、社会经济变化等，以提高模型的预测能力和准确性。

模型优化：探索其他机器学习模型（如随机森林、神经网络等），对比不同模型的预测效果，选择最佳模型。

实时更新：建立动态更新机制，定期更新数据和模型，以反映最新的经济动态和政策变化。

6.4 模型推广

区域推广：可以将该模型推广至其他城市群或经济区域，进行类似的经济发展分析和预测，帮助地方政府和企业制定经济策略。

跨领域应用：模型的思路和方法可以应用于其他领域，如环境科学、公共卫生等，进行相关因素的影响分析。

政策支持：将模型的结果和建议纳入政策制定过程，为政府提供基于数据的决策支持，提高政策的针对性和有效性。

七、 参考文献

[1] 赵娟, 左光宇. 数字经济推动粤港澳大湾区社会经济融合发展探究[J]. 老字号品牌营销, 2024, (20): 38-40.

[2] 粤港澳大湾区产业布局及发展研究报告, 马咪, 吾永超

[3] 林蓝玉，陈秀芳，张德飞，$ARIMA$模型在股票中的应用，经济研究导刊，2018，376(26):151-153.

[4] James G, Witten D, Hastie T, et al. An introduction to statistical learning[M]. New York: springer, 2013.

[5] 谭晓丽. 基于网络分析法的世界三大湾区城市群经济发展对比分析[J]. 经营管理者, 2024, (10): 87-89.