

城乡收入差距调控研究——基于乡村振兴与数字经济协同视角

摘要

随着脱贫攻坚战的全面胜利，我国的“三农”工作重点已向推进乡村振兴转移并取得了积极的成效，当前城乡发展的协调性逐渐提高，城乡差距逐渐缩小。但与西方国家和其他发展中国家相比，我国的城乡收入差距仍处于较高的水平，依旧是阻碍我国经济社会发展的瓶颈。与此同时，数字经济已成为推动经济社会发展的新引擎，借力数字经济赋能乡村振兴亦成为关键举措。本文建立耦合模型，厘清数字经济与乡村振兴的耦合关系，探究数字经济赋能乡村振兴的机制路径，建立回归、ARIMA 等模型识别影响城乡收入差距的因素，为政府的科学决策提供参考依据。

针对问题一：建立乡村振兴-数字经济耦合模型。首先，基于题目给定的 8 个维度，选取 26 个指标构成乡村振兴和数字经济的评价指标体系，并采用熵权法得到各省区的乡村振兴发展指数和数字经济发展指数。然后，构建耦合协调度模型，得到我国及各省区历年的耦合度和耦合协调度，以地理方位为依据将我国划分为八个地区，以各地区所含省份的平均耦合度及耦合协调度表征该地区的耦合关系，分析各地区的耦合度变化趋势、比较地区之间的耦合差异。最后，根据实证研究结果分析当前乡村振兴-数字经济耦合系统中存在的问题，并针对性地提出政策性建议。

针对问题二：构建中国城乡收入差距影响因素的多元线性回归模型。首先，使用第一问的数据预处理后的数据作为自变量，城乡收入差距作为因变量，选取问题一的乡村振兴的 5 个维度和数字经济的 3 个维度，分别采用多项指标对上述维度进行量化，一方面通过熵权法对指标赋予权重，根据贡献率筛选出 18 个指标，进而构建中国城乡收入差距影响因素的多元线性回归模型，另一方面，采用随机森林方法衡量指标的重要性，建立中国城乡收入差距定量预测模型，然后以 MAE、MSE、RMSE 和拟合度 4 个评价指标对两种模型进行对比分析并可视化实际值和预测值的拟合图，最终发现回归模型的效果最佳。其次，从华东，华中等七个地区分别选取 2022 年农业现代化和乡村振兴示范区作为典型案例，进行描述性统计分析，并构建城乡收入差距定量预测模型，对上述结论进行验证。最后，以国家和各地区的视角，横向分析不同地区城乡收入差距的差异，纵向分析城乡收入差距随时间变化的特征，进而得出相应结论，提出有针对性的政策建议。

针对问题三：构建 ARIMA 差分时间自回归移动平均模型。首先，采集城乡收入差距与政府干预、对外开放、城市化水平、数字基础、经济发展水平、就业结构六项指标数据。其次，采用 ADF 检验来判断数据的平稳性。然后，建立 ARIMA 差分时间自回归移动平均模型来预测我国未来城乡人均可支配收入差距的变化趋势，通过相关性分析来检测不同指标对于城乡人均可支配收入差距的贡献度是正向还是反向的。通过灵敏度分析来判断不同的指标对于该目标的影响力度，以及不同指标在不同力度下对城乡收入差距的影响，并估计我国多久能达到 OECD 成员国的城乡收入差距的水平。结果表明，在政策、政府干预条件不变的情况下，城乡收入差距下降缓慢，需要超过 20 年以上的时间才能将该差距缩小到接近 OECD 成员国的基本水平。

最后，我们对提出的模型进行全面的评价：本文综合使用多种模型进行分析，流程完整规范，从实证研究结果来看模型精度较高，能有效解决所提出的问题，且模型简单易懂，可推广至其他应用场景，具有较强的通用性。

关键词：耦合度 随机森林 ARIMA 模型 熵权法 多元线性回归模型

目录

摘要	I
1 问题综述	1
1.1 问题背景	1
1.2 问题提出	1
2 模型假设与符号说明	2
2.1 模型基本假设	2
2.2 符号说明	2
3 数据预处理	3
3.1 指标选取	3
3.1.1 乡村振兴-数字经济耦合模型	3
3.1.2 城乡收入差距影响因素模型	5
3.1.3 城乡收入差距调控模型	7
3.2 数据清洗	7
3.3 数据规约	10
4 问题一分析与模型建立	12
4.1 问题分析	12
4.2 耦合关系模型建立	12
4.2.1 评价模型——熵值法	12
4.2.2 耦合模型——耦合协调度计算模型	13
4.3 模型求解	13
4.3.1 熵值法求解	13
4.3.2 耦合协调度计算模型求解	15
4.4 结果分析	15
5 问题二分析与模型建立	18
5.1 问题分析	18
5.2 模型建立	18
5.2.1 模型准备——k 折交叉验证	18
5.2.2 模型准备——评价指标	19
5.2.3 模型介绍——随机森林	20
5.2.4 模型介绍——多元线性回归	21
5.3 模型求解	21
5.3.1 RF 回归模型	21
5.3.2 线性回归模型	23
5.3.3 模型比较	26
5.4 典型案例分析	27
5.5 城乡收入差距分析	30

5.5.1 城乡收入差距变化分析	30
5.5.2 城乡收入差距地区差异分析	31
6 问题三分析与模型建立	31
6.1 问题分析	31
6.2 模型介绍	32
6.3 模型准备	32
6.4 模型求解	33
6.5 结果分析	37
7 模型评价与推广	41
7.1 模型评价	41
7.2 模型的推广	41
参考文献	42
附录	43

1 问题综述

1.1 问题背景

随着脱贫攻坚战的全面胜利，我国的“三农”工作重点已向推进乡村振兴转移。与此同时，当前物联网、云计算、5G 等信息技术快速发展，科技创新成果不断涌现，带动了新产业、新业态、新模式的发展，加快了经济的数字化转型进程，数字经济已成为推动经济社会发展的新引擎。如何借力数字经济赋能乡村振兴已成为需要切实解决的现实问题。因此，厘清数字经济与乡村振兴的耦合关系，探究数字经济赋能乡村振兴的机制路径具有重要的现实意义。另一方面，伴随着数字经济的高质量发展与乡村振兴工作的全面推进，我国城乡发展的协调性不断提高，城乡差距逐渐缩小，取得了积极的成效。但与西方国家和其他发展中国家相比，我国的城乡收入差仍处于较高的水平，城乡收入差距达依旧是阻碍我国经济社会发展的瓶颈。为此，需要识别影响城乡收入差距的因素，为政府的科学决策提供参考依据。

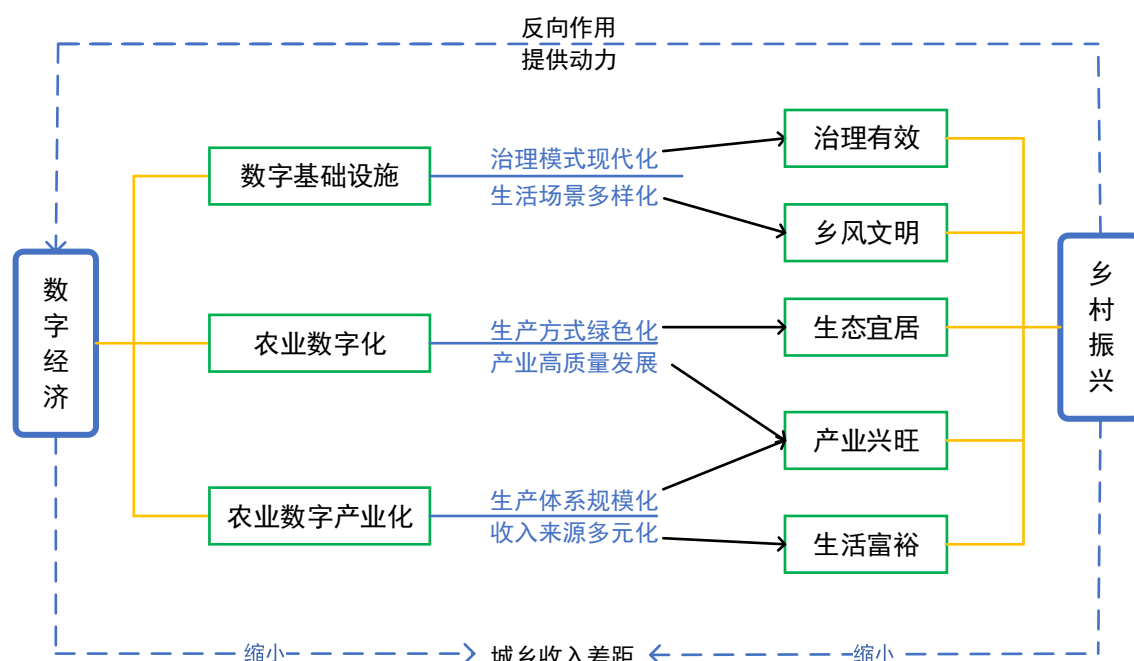


图 1-1 乡村振兴-数字经济耦合机制图

1.2 问题提出

本文拟从探析数字经济与乡村振兴的耦合关系、识别城乡收入差距影响因素、制定缩小城乡收入差距对策三个角度出发，解决以下三个问题：

一、通过查阅文献，分别从产业兴旺、生态宜居、乡风文明、治理有效、生活富裕五个维度及数字基础设施、农业数字化、农业数字产业化三个维度构建乡村振兴与数字经济的评价指标体系，并在此基础上建立数学模型，定量评价数字经济与乡村振兴的耦合关系。最后，结合实证研究结果，针对数字经济与乡村振兴的耦合中存在的问题提出相应的政策建议。

二、从乡村振兴和数字经济的角度出发，识别影响我国城乡收入差距的主要因素，并比较不同因素对缩小城乡收入差距的影响程度。从实际出发，选取典型案例进行实证研究，检验所得结论的合理性，并讨论我国城乡收入差距的变化特征以及不同地区之间的差异。

三、大部分 OECD 成员国的城乡居民收入比在 1.25 以下，而我国的城乡收入差距与美国等 OECD 成员国的城乡收入差距水平存在较大差距。因此，需要通过建立数学模型，探究缩短城乡收入差距的可行措施，明确以多大的调节力度、在什么条件下、经过多长时间才可能达到 OECD 成员国的城乡收入差距的基本水平。

2 模型假设与符号说明

2.1 模型基本假设

对于论文所研究的问题的统计数据与数据来源，本文做出如下假设与约束：

1. 所查阅的相关数据真实有效，且能准确反映指标所代表的含义。
2. 指标的计算方式均采用官方的计算公式，不考虑其他类别的定义。
3. 所采集到的数据在当年能代表地区或国家的一般情况，不考虑由自然灾害、恶劣天气、突发事件对指标所造成的影响。
4. 所有涉及指标的相关政策都以理性为前提，不存在因偏好发布的相关政策。

2.2 符号说明

表 2-1 符号说明

符号	含义
X	代表乡村振兴的变量
Y	代表数字经济的变量
x_{new}	效益型指标规约化后的表示
x_t	指标数据值的表示
x_{max}	指标数据中的最大值
x_{min}	指标数据中的最小值
X_{ij}	第 i 年第 j 个指标变量
x_{ij}	第 i 年第 j 个指标变量值
x_j	第 j 个指标变量值
p_{ij}	第 i 年第 j 个指标的占全部指标值的比重
e_j	第 j 个指标的熵值
g_j	第 j 个指标的差异系数
w_j	第 j 个指标的权重
F_{ij}	第 i 年第 j 个指标的评价值
F_i	第 i 年总的评价值
C	耦合度
T	耦合协调发展水平的综合评价系数
D	耦合协调度
α	表示乡村振兴子系统的权重
β	表示数字经济子系统的权重
MAE	平均绝对误差
MSE	均方误差
$RMSE$	均方根误差
R	拟合度
SSR	回归线的误差平方和
SST	均线的误差平方和

3 数据预处理

对于指标中的统计数据需要进行预处理操作，预处理的主要目的是为了剔除数据集中的存在的一些异常数据、以及空值。对于异常数据，在统计表中出现的一般情况是在一组数据中出现了个别异常大（异常小）的值，或是对于整型约束的条件中出现小数，或是在非负的数据集中出现小于 0 的数。对于空值，则是数据缺失。这两类数据集中存在的问题，在对数据进行统计分析时，具有干扰作用，容易影响数据分析的结果，因此需要对数据进行预处理。

3.1 指标选取

3.1.1 乡村振兴-数字经济耦合模型

为构建乡村振兴-数字经济耦合模型，本文从 8 个维度构建模型的指标体系，其中乡村振兴指标体系从产业兴旺、生活富裕、生态宜居、乡风文明、治理有效五个维度刻画，数字经济体系从数字基础设施、数字产业化、产业数字化三个维度刻画，参考相关文献选取了 27 个具体的指标。最终得到如表 3-1 所示的指标体系。

本文以 2002 年到 2021 年 20 年中全国 31 个省（自治区、直辖市）的面板数据为研究对象，考虑到数据的可获得性，暂未包括港澳台地区。所选数据均为年度数据，数据来源于国泰安数据库、国家统计局、北京大学数字金融研究中心。

表 3-1 耦合模型指标体系

变量	一级指标	二级指标	单位
乡村振兴 (X)	产业兴旺	有效灌溉面积	千公顷
		农作物播种面积	千公顷
		亩均粮食产量	吨/公顷
		单位耕地面积化肥施用量	吨/公顷
	生活富裕	农村居民人均可支配收入	元
		农村人均消费支出	元
		农村从业人口	万人
	生态宜居	森林资源覆盖率	%
		农村每万人医疗机构床位数	个
		单位面积农药投用量	千克/公顷
		单位面积塑料薄膜使用量	千克/公顷
	乡风文明	普通高中在校学生人数	万人
		农村电视节目人口覆盖率	%
	治理有效	每万人村民居委会单位数	个
		水土流失治理面积	千公顷
		除涝面积	千公顷
数字经济 (Y)	数字基础设施	网页数	万个
		光缆线路长度	公里
		互联网宽带接入端口数	万个
		农村居民每百户年移动电话拥有量	个
		农业气象观测站	个
	数字产业化	规模以上电子信息产业制造企业数	个
		软件业务收入	万元
		信息技术服务收入	万元
		农村电子商务销售额和采购额	亿元
	产业数字化	每百家企业用于网站数	个
		数字普惠金融指数	-

采集到的统计数据通常存在一定的冗余性。观察表 3-2、表 3-3 和表 3-4 的数据可以看出，表 3-2 中列举了最近 10 年全国各省份的粮食总产量，表 3-3 中列举了最近 10 年全国各省份的粮食种植面积，表 3-4 中列举了最近 10 年全国各省份平均每公顷土地上收获的粮食产量，这三个表格中的统计数据存在冗余情况，表 3-4 是表 3-2 和表 3-2 综合以后的成果，因此在指标选取阶段会考虑值选取其中两个指标的数据另外一个指标的数据直接舍弃，这里考虑表 3-2 和表 3-4 中的指标数据。

表 3-2 全国各省份粮食总产量

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	269.51	263.43	281.7	296.77	308.32	338.06	368.24	420.07	421.78	395.71	363.14
天津市	509.26	476.44	414.35	390.5	382.07	395.57	378.44	367.75	351.12	327.44	311.83
河北省	7018.67	6742.49	6061.46	5707	5373.38	5299.66	5291.68	5373.76	5284.43	4912.42	4570.27
山西省	2134.02	1935.84	1626.54	1460.64	1418.73	1429.91	1424.96	1440.6	1372.02	1265.98	1163.4
内蒙古	3815.12	3472.36	3176.34	2985.32	2813.54	2803.55	2761.56	2786.54	2702.69	2450.26	2205.24
辽宁省	4927.73	4582.56	4368.25	4061.93	3851.62	3764.09	4057.59	3949.39	3878.89	3679.53	3343.75
吉林省	2972.32	2976	2442.73	2184.34	2064.29	2167.89	2292.97	2302.04	2276.4	2181.92	2029.86
黑龙江省	6459.97	6438.11	5929.97	5624.29	5586.63	5202.87	5030.06	4865.8	4560.19	3841.96	3103.02
上海市	268.93	279.82	284.84	289.58	292.61	300.84	327.71	343.78	342.29	337.81	328.24
江苏省	8279.72	7952.59	7503.15	7192.46	7161.21	7178.96	6980.37	6402.75	6124.25	5781.5	5216.99
浙江省	3579.21	3496.94	3355.25	3157.25	3093.36	3038.49	2845.56	2773.7	2780.11	2615.08	2501.08
安徽省	6004.31	5680.91	5162.13	4672.71	4597.94	4432.32	4183.14	4024.2	3818.73	3552.27	3296.59
福建省	5200.97	4901.07	4636.56	4229.52	3947.16	3784.24	3399.3	3247.11	3057.36	2843.47	2614.57
江西省	3998.09	3820.74	3481.29	3148.57	3069.01	3019.87	2808.37	2670.18	2529.69	2359.96	2175.14
山东省	11468.01	10190.58	9671.67	9397.39	9140.36	9075.6	9283.92	8988.18	8577.06	7817.84	7311.11
河南省	10501.2	9956.35	8541.77	7757.94	7562.53	7405.42	7299.58	7244.34	6938.24	6473.7	6055.53
湖北省	8296.44	7303.64	6681.85	6207.83	6129.72	5863.98	5387.13	5162.94	4920.13	4542.16	4110.16
湖南省	7662.36	7511.96	6405.06	5361.62	5213.48	5057.52	4682.31	4577.08	4432.69	4390.34	4111.04
广东省	8305.84	7901.92	7175.89	6318.12	5969.87	5817.55	5303.63	5053.72	4802.01	4550.29	4301.86
广西	6524.39	5913.28	5498.81	4909.24	4698.71	4560.2	4197.12	3947.73	3755.19	3490.72	3323.37
海南省	2014.79	1821.02	1689.4	1535.73	1488.86	1433.88	1294	1227.14	1125.93	1067.33	991.53
重庆市	2935.65	2749.05	2337.81	2052.41	1902.47	1851.6	1609.05	1485.78	1418.27	1327.34	1204.16
四川省	9383.32	9216.4	7889.35	7195.65	6955.55	6816.92	6377.84	5888.09	5620.27	5433.12	4932.73
贵州省	4691.97	4358.62	3888.99	3619.52	3413.86	3123.13	2740.93	2120.27	1664.49	1437.97	1166.69
云南省	6351.82	5920.52	4935.73	4108.88	3872.93	3704.69	3438.73	3307.82	3097.5	2716.5	2334.48
西藏	255.34	233.53	212.81	195.47	178.16	172.97	149.46	138.72	128	118.33	109.37
陕西省	4313.44	4056.61	3536.8	3239.99	3077.62	2994.83	2821.56	2748.59	2569.78	2309.55	2063.84
甘肃省	2439.54	2103.61	1887.58	1659.36	1559.64	1443.12	1386.18	1307.31	1233.78	1116	999.69
青海省	528.53	507.1	454.35	405.93	364.1	338.8	319.27	327.49	310.3	263.86	230.82
宁夏	759.81	703.07	584.85	575.77	517.42	496.27	484.46	448.18	432.13	386.46	355.62
新疆	5143.12	4315.61	3850.65	3637.79	3326.59	3165.92	2968.39	2881.48	2648	2355.35	2011.88

表 3-3 全国各省份粮食种植面积

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	96.84	98.18	88.55	103.79	120.94	145.55	172.14	194.64	237.29	278.03	298.62
天津市	419.57	419.17	410.27	429.27	439.52	443.66	433.28	443.01	441.34	448.98	442.45
河北省	8139.75	8089.44	8132.69	8197.13	8381.65	8467.51	8457.76	8454.9	8416.01	8361.89	8278.32
山西省	3540.36	3541.51	3524.42	3555.16	3577.62	3591.54	3612.65	3664.57	3667.89	3696.23	3694.41
内蒙古	8863.97	8882.81	8885.02	8824.07	9014.22	8957.21	8423.68	8079.08	7823.05	7670.84	7539.51
辽宁省	4237.33	4287.78	4217.1	4207.11	4172.32	4242.73	4335.47	4219.8	4154.41	4095.55	3997.53
吉林省	6116.31	6150.99	6117.05	6080.89	6086.2	6063.25	5997.91	5890.5	5632.95	5431.98	5297.23
黑龙江省	14784.51	14910.13	14770.08	14673.33	14767.59	14829.46	14811.88	14497.68	14211.99	13929.68	13749.58
上海市	266.27	255.16	261.39	282.27	284.95	303.78	350.64	369.91	391.3	401.64	420.03
江苏省	7480.42	7478.4	7442.63	7520.23	7556.4	7639.9	7693.7	7615.79	7597.93	7569.22	7508.44
浙江省	1997.60	2014.5	1999.62	1978.68	1981.11	1946.46	1977.8	1986.77	2062.45	2095.67	2267.84
安徽省	8790.35	8817.99	8781.96	8771.11	8726.68	8790.1	9598.29	9500.18	9364.9	9336.3	9392.44
福建省	1602.62	1631.27	1599.29	1577.31	1549.34	1548.8	1617.17	1670.17	1740.96	1761.23	1878.97
江西省	5573.79	5644.37	5521.18	5555.81	5638.46	5668.88	5688.41	5667.34	5636.98	5596.83	5546.41
山东省	10966.32	10889.06	10933.09	11076.8	11107.79	11278.61	11381.01	11316.65	11136.63	10862.27	10782.86
河南省	14728.44	14687.99	14713.98	14783.35	14732.53	14902.72	14879.73	14731.54	14586.5	14386.89	14373.32
湖北省	7914.39	7974.39	7815.89	7952.9	7956.14	7908.5	7983.43	7791.42	7731.26	7684.4	7454.98
湖南省	8211.34	8400.13	8122.79	8111.09	8321.97	8341.5	8355.22	8398.63	8352.67	8299.62	8178.06
广东省	4362.85	4451.81	4357.38	4279.36	4227.51	4181.56	4194.55	4225.19	4178.17	4248.43	4244.73
广西	6022.97	6107.32	5989.21	5972.38	5969.88	5966.74	6078.81	5855.03	6076.46	6036.16	5899.88

海南省	688.68	676.86	676.24	712.94	709.44	731.9	757.5	779.4	772.48	786.57	782.4
重庆市	3355.59	3372.54	3345.74	3348.49	3339.56	3333.1	3311.32	3288.58	3318.49	3320.31	3225.78
四川省	9719.40	9849.89	9692.98	9615.32	9575.05	9493.82	9451.06	9377.7	9371.69	9319.58	9215.4
贵州省	5478.02	5475.35	5481.56	5477.16	5659.37	5604.8	5532.9	5510.5	5388.73	5201.03	4998.1
云南省	6939.77	6989.67	6938.86	6890.77	6790.8	6786.62	6818.64	6842.84	6848.56	6674.13	6484.91
西藏	271.33	272.08	271.53	270.38	254.05	263.15	229.69	227.1	225.77	220.24	219.04
陕西省	4127.98	4160.85	4132.09	4091.01	4063.88	4160.15	4050	4053.87	4108.22	4182.84	4193.6
甘肃省	3845.65	3931.82	3831.57	3773.55	3752.03	3749.2	3768.38	3775.83	3779.84	3770.7	3774.33
青海省	560.74	571.42	553.54	557.25	555.31	557.75	558.39	553.7	555.77	554.21	540.05
宁夏	1163.92	1174.23	1152.96	1164.57	1132.63	1118.85	1132.43	1126.76	1147.08	1176.46	1205.76
新疆	6172.96	6280.01	6169.99	6068.89	5886.96	5921.27	5175.47	5074.23	4744.47	4673.34	4475.41

表 3-4 全国各省份每公顷粮食产量（单位：吨/公顷）

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	27.83	26.83	31.81	28.59	25.49	23.23	21.39	21.58	17.77	14.23	12.16
天津市	12.14	11.37	10.10	9.10	8.69	8.92	8.73	8.30	7.96	7.29	7.05
河北省	8.62	8.33	7.45	6.96	6.41	6.26	6.26	6.36	6.28	5.87	5.52
山西省	6.03	5.47	4.62	4.11	3.97	3.98	3.94	3.93	3.74	3.43	3.15
内蒙古	4.30	3.91	3.57	3.38	3.12	3.13	3.28	3.45	3.45	3.19	2.92
辽宁省	11.63	10.69	10.36	9.65	9.23	8.87	9.36	9.36	9.34	8.98	8.36
吉林省	4.86	4.84	3.99	3.59	3.39	3.58	3.82	3.91	4.04	4.02	3.83
黑龙江省	4.37	4.32	4.01	3.83	3.78	3.51	3.40	3.36	3.21	2.76	2.26
上海市	10.10	10.97	10.90	10.26	10.27	9.90	9.35	9.29	8.75	8.41	7.81
江苏省	11.07	10.63	10.08	9.56	9.48	9.40	9.07	8.41	8.06	7.64	6.95
浙江省	17.92	17.36	16.78	15.96	15.61	15.61	14.39	13.96	13.48	12.48	11.03
安徽省	6.83	6.44	5.88	5.33	5.27	5.04	4.36	4.24	4.08	3.80	3.51
福建省	32.45	30.04	28.99	26.81	25.48	24.43	21.02	19.44	17.56	16.14	13.91
江西省	7.17	6.77	6.31	5.67	5.44	5.33	4.94	4.71	4.49	4.22	3.92
山东省	10.46	9.36	8.85	8.48	8.23	8.05	8.16	7.94	7.70	7.20	6.78
河南省	7.13	6.78	5.81	5.25	5.13	4.97	4.91	4.92	4.76	4.50	4.21
湖北省	10.48	9.16	8.55	7.81	7.70	7.41	6.75	6.63	6.36	5.91	5.51
湖南省	9.33	8.94	7.89	6.61	6.26	6.06	5.60	5.45	5.31	5.29	5.03
广东省	19.04	17.75	16.47	14.76	14.12	13.91	12.64	11.96	11.49	10.71	10.13
广西	10.83	9.68	9.18	8.22	7.87	7.64	6.90	6.74	6.18	5.78	5.63
海南省	29.26	26.90	24.98	21.54	20.99	19.59	17.08	15.74	14.58	13.57	12.67
重庆市	8.75	8.15	6.99	6.13	5.70	5.56	4.86	4.52	4.27	4.00	3.73
四川省	9.65	9.36	8.14	7.48	7.26	7.18	6.75	6.28	6.00	5.83	5.35
贵州省	8.57	7.96	7.09	6.61	6.03	5.57	4.95	3.85	3.09	2.76	2.33
云南省	9.15	8.47	7.11	5.96	5.70	5.46	5.04	4.83	4.52	4.07	3.60
西藏	9.41	8.58	7.84	7.23	7.01	6.57	6.51	6.11	5.67	5.37	4.99
陕西省	10.45	9.75	8.56	7.92	7.57	7.20	6.97	6.78	6.26	5.52	4.92
甘肃省	6.34	5.35	4.93	4.40	4.16	3.85	3.68	3.46	3.26	2.96	2.65
青海省	9.43	8.87	8.21	7.28	6.56	6.07	5.72	5.91	5.58	4.76	4.27
宁夏	6.53	5.99	5.07	4.94	4.57	4.44	4.28	3.98	3.77	3.28	2.95
新疆	8.33	6.87	6.24	5.99	5.65	5.35	5.74	5.68	5.58	5.04	4.50

3.1.2 城乡收入差距影响因素模型

我国的城乡收入差距影响因素模型依旧从上述八个维度进行构建，并选取各省区的均值作为全国性指标。此外，还补充了在“乡村振兴”的“生活富裕”维度新增了“农村居民恩格尔系数”，在“乡风文明”维度新增了“文娱教育支出占总支出比例”。数据来源为国泰安数据库。

下面针对各指标进行描述性分析。从表 3-5 可以看出：Y15 农业气象观测站,Y24 农村电子商务销售额和采购额两项的最大值超过平均值 3 个标准差，说明数据波动较大，需要对数据做预处理。

表 3-5 城乡收入差距影响因素模型描述性分析

名称	最小值	最大值	平均值	标准差	中位数
X11 有效灌溉面积	54014.230	69160.560	61830.033	5543.037	62358.990

名称	最小值	最大值	平均值	标准差	中位数
X12 农作物总播种面积	152149.470	167487.120	160490.749	5761.144	161215.820
X13 亩均粮食产量	67.420	348.460	191.186	88.305	190.355
X14 单位耕地面积化肥施用量	8.580	12.040	10.841	1.020	10.920
X21 农村居民人均可支配收入	83726.000	610654.000	284425.550	168492.158	254567.000
X22 农村人均消费支出	62038.000	499175.000	228154.850	137081.710	202547.500
X23 农村从业人口	48526.850	57435.530	53418.315	2746.319	53771.670
X24 农村居民恩格尔系数	30.000	45.300	36.900	5.072	36.500
X31 森林资源覆盖率	930.700	1066.800	1003.390	60.330	1003.700
X32 农村每万人医疗机构床位数	750.640	1548.720	1061.004	277.505	959.785
X33 单位面积农药投用量	26.770	40.440	35.309	4.509	36.170
X34 单位面积塑料薄膜使用量	364089.230	656009.710	548136.075	98870.000	567040.275
X41 普通高中在校学生人数	1683.820	2522.420	2364.935	201.902	2420.825
X42 农村电视节目人口覆盖率	2989.800	3088.050	3030.115	30.583	3015.870
X43 文娱教育支出占总支出比例	0.095	0.137	0.112	0.013	0.107
X51 每万人村民居委会单位数	287.750	355.120	311.768	19.696	308.245
X52 水土流失治理面积	85410.200	147664.070	111677.234	17801.369	108277.870
X53 除涝面积	21111.620	24624.000	22384.493	1243.935	21804.565
Y11 网页数	8658229.850	34169475.827	16975332.348	8855094.067	11370568.745
Y12 光缆线路长度	2728799.907	60017791.845	20587164.772	18241468.737	13456301.665
Y13 互联网宽带接入端口数	1394.292	97222.333	38863.400	35430.461	27674.000
Y14 农村居民每百户年移动电话拥有量	1021.438	3734.783	2395.590	916.257	2512.240
Y15 农业气象观测站	625.000	1922.477	805.666	298.268	722.000
Y21 规模以上电子信息产业制造业企业数	15054.000	25105.000	19625.181	2842.338	18184.631
Y22 软件业务收入	53975851.783	882566475.040	328926102.003	266979546.667	230706427.652
Y23 信息技术服务收入	132649501.448	660404490.314	259775720.082	146154843.402	206739323.279
Y24 农村电子商务销售额和采购额	64296.859	5737398.323	598664.723	1310968.361	186835.459
Y31 每百家企业拥有网站数	1401.000	1714.000	1635.377	79.041	1669.938
Y32 数字普惠金融指数	23.460	8918.369	3123.966	3224.704	1753.295
城乡收入差距	2.300	3.190	2.898	0.250	2.940

由于不同指标的量纲与跨度存在较大差异，因此对各指标进行线性归一化，归一化后的部分指标数据如表 3-6 所示。

表 3-6 指标数据归一化

名称		最小值	最大值	平均值	标准差	中位数
X43 文娱教育支出占总支出比例	处理前	0.095	0.137	0.112	0.013	0.107
	处理后	0	1	0.399	0.308	0.284
Y23 信息技术服务收入	处理前	132649501.448	660404490.314	259775720.082	146154843.402	206739323.279
	处理后	0	1	0.241	0.277	0.14

3.1.3 城乡收入差距调控模型

为构建我国城乡收入差距的调控模型，我们将从政府干预程度、对外开放程度、城市化水平、就业结构、数字基础设施、经济发展水平六个维度，探究不同调控手段对缩小城乡收入差距的影响。其中政府干预程度为农林水事务支出占财政总支出的比重；对外开放程度为进出口总额占国内生产总值的比重；城市化水平为城市人口比重；就业结构为第一产业占三大产业就业人数的百分比；数字基础设施为农村互联网端口接入数占全国互联网端口接入数的比重；经济发展水平为人均国内生产总值。我们选取 2002 年到 2020 年的全国年度数据作为样本，数据来源为国泰安数据库（宏观经济数据库、经济内循环数据库、进出口统计数据库、农林牧渔业数据库）。

表 3-7 城乡收入差距调控模型描述性统计

	城乡收入差距	政府干预程度	对外开放程度	城市化水平	数字基础设施	经济发展水平	就业结构
均值	2.93	0.08	0.46	49.80	0.21	37140.40	35.56
中位数	2.95	0.09	0.44	50.50	0.22	36301.57	34.80
极大值	3.19	0.10	0.64	61.40	0.30	71999.59	50.00
极小值	2.56	0.04	0.32	37.60	0.05	9506.20	23.60
方差	0.21	0.02	0.11	7.80	0.07	20999.22	8.48
样本数	19	19	19	19	19	19	19

表 3-7 为上述指标的描述性统计结果。可以发现，近 20 年里，我国的城乡收入差距的变化较为平稳，极大值与极小值之差和方差均较小，城乡收入差距的缩小速度较慢。而代表经济发展水平的人均国内生产总值的波动非常大，极小值仅为极大值的 13%左右，说明我国 20 年来的经济发展速度非常快，经济发展水平较 2002 年有了大幅提升。此外，政府干预程度和基础设施数字化水平的均值都与极大值十分接近，说明大多数情况下，政府对农林水事务的投资力度和基础设施数字化建设水平均处于相对较高水平，政府对乡村振兴和数字经济的重视程度高。我国的城市化水平和就业结构也有明显的变化，极大值大约为极小值的两倍，说明我国的城市人口数量显著提高，同时第一产业的就业人数不断增多，已经达到了总就业人数的一半。

3.2 数据清洗

数据清洗阶段，需要把上一阶段挑选出的指标做进一步处理，采集到的初始数据集存在缺失项，需要对数据进行填充，论文采用的方法是均值填充方法与移动平均法相结合的方式共同填补缺失数据。

均值填充方法是用缺失数据左右临近的一些数据项累加之后求平均值，再填充到缺失位置，如果缺失数据只有一侧有临近的数据那么可以使用单侧的均值填充，即使用缺失数据某一侧临近的数据项进行累加后求平均值进行填充，从而达到填补数据的效果。

假设有一组数据为 x ，缺失值的位置是 m ，那么双侧均值填充的计算公式为：

$$x_m = \frac{\sum_{i=1}^k (x_{m+i} + x_{m-i})}{k * 2}$$

单侧的均值填充计算公式为：

$$x_m = \frac{\sum_{i=1}^k x_{m \pm i}}{k}$$

如表 3-8，其中全国各省份 2021 年的有效灌溉面积缺失，因此使用单侧均值填充的方法进行数据填补，填补后的数据如表 3-9 所示。

表 3-8 全国各省份有效灌溉面积

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
----	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

北京市	109.35	109.24	109.67	115.48	128.47	137.35	143.11	153.02	207.54	209.33
天津市	299.12	304.80	304.66	306.62	306.62	308.87	308.87	308.87	337.04	338.00
河北省	4470.03	4482.16	4492.33	4474.67	4457.64	4447.98	4404.22	4349.03	4603.07	4596.61
山西省	1517.38	1519.34	1518.68	1511.21	1487.29	1460.28	1408.17	1382.79	1319.06	1319.85
内蒙古	3199.12	3199.19	3196.52	3174.83	3131.53	3086.90	3011.88	2957.76	3125.24	3072.39
辽宁省	1632.54	1629.18	1619.33	1610.55	1572.99	1520.31	1473.97	1407.84	1698.82	1588.38
吉林省	1905.44	1909.50	1893.05	1893.05	1832.17	1790.87	1628.43	1510.13	1851.87	1807.52
黑龙江省	6171.61	6177.59	6119.57	6030.97	5932.74	5530.84	5305.20	5342.12	4776.48	4332.65
上海市	165.05	190.76	190.76	190.76	189.81	188.21	184.09	184.09	199.02	199.61
江苏省	4224.73	4205.44	4179.83	4131.88	4054.07	3952.50	3890.53	3785.27	3929.72	3817.92
浙江省	1415.70	1405.43	1440.80	1444.70	1446.31	1432.15	1425.37	1409.39	1471.02	1456.80
安徽省	4608.83	4580.84	4538.29	4504.14	4437.46	4400.34	4331.69	4305.53	3585.09	3547.65
福建省	1110.42	1076.78	1085.18	1064.84	1055.37	1061.65	1116.12	1122.42	968.51	967.48
江西省	2038.49	2036.14	2032.02	2039.42	2036.83	2027.67	2001.57	1995.60	1907.06	1867.67
山东省	5293.56	5271.37	5235.99	5191.06	5161.16	4964.43	4901.95	4729.03	5058.11	4986.88
河南省	5463.07	5328.95	5288.69	5273.63	5242.92	5210.64	5101.15	4969.11	5205.63	5150.44
湖北省	3086.04	2968.99	2931.90	2919.17	2905.57	2899.15	2855.32	2791.41	2548.91	2455.69
湖南省	3192.90	3176.11	3164.00	3145.87	3132.37	3113.32	3101.70	3084.30	2715.78	2762.41
广东省	1776.46	1773.40	1775.21	1774.61	1771.71	1771.26	1770.99	1770.76	1874.44	1873.16
广西	1731.03	1713.08	1706.88	1669.87	1646.07	1618.79	1600.00	1586.37	1541.29	1529.24
海南省	292.21	290.63	290.48	289.25	289.95	263.99	259.92	260.93	256.75	247.51
重庆市	698.32	697.68	696.94	694.26	690.60	687.19	677.26	675.18	702.97	692.88
四川省	2992.24	2954.09	2932.54	2873.10	2813.55	2735.09	2666.32	2616.54	2662.65	2600.75
贵州省	1165.49	1153.99	1132.24	1114.12	1088.07	1065.43	981.83	926.90	1214.57	1201.19
云南省	1978.07	1922.49	1898.07	1851.42	1809.39	1757.71	1708.97	1660.27	1677.90	1634.24
西藏	282.82	275.94	264.53	261.23	251.53	247.80	244.03	239.27	251.04	245.31
陕西省	1336.81	1285.16	1274.99	1263.09	1251.39	1236.77	1226.49	1209.94	1277.18	1274.34
甘肃省	1338.60	1328.86	1337.54	1331.43	1317.51	1306.72	1297.06	1284.08	1297.58	1291.82
青海省	219.25	213.33	214.04	206.61	202.35	196.99	182.49	186.90	251.67	251.67
宁夏	552.53	538.29	523.45	511.45	515.15	506.53	498.91	498.56	491.35	477.59
新疆	4893.35	4959.89	4883.46	4952.29	4982.03	4944.92	4831.89	4769.89	4029.07	3884.57

表 3-9 国各省份有效灌溉面积（数据填充后）

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	109.42	109.35	109.24	109.67	115.48	128.47	137.35	143.11	153.02	207.54	209.33
天津市	302.86	299.12	304.8	304.66	306.62	306.62	308.87	308.87	308.87	337.04	338
河北省	4481.51	4470.03	4482.16	4492.33	4474.67	4457.64	4447.98	4404.22	4349.03	4603.07	4596.61
山西省	1518.47	1517.38	1519.34	1518.68	1511.21	1487.29	1460.28	1408.17	1382.79	1319.06	1319.85
内蒙古	3198.28	3199.12	3199.19	3196.52	3174.83	3131.53	3086.9	3011.88	2957.76	3125.24	3072.39
辽宁省	1627.02	1632.54	1629.18	1619.33	1610.55	1572.99	1520.31	1473.97	1407.84	1698.82	1588.38
吉林省	1902.66	1905.44	1909.5	1893.05	1893.05	1832.17	1790.87	1628.43	1510.13	1851.87	1807.52
黑龙江省	6156.26	6171.61	6177.59	6119.57	6030.97	5932.74	5530.84	5305.2	5342.12	4776.48	4332.65
上海市	182.19	165.05	190.76	190.76	190.76	189.81	188.21	184.09	184.09	199.02	199.61
江苏省	4203.33	4224.73	4205.44	4179.83	4131.88	4054.07	3952.5	3890.53	3785.27	3929.72	3817.92
浙江省	1420.64	1415.7	1405.43	1440.8	1444.7	1446.31	1432.15	1425.37	1409.39	1471.02	1456.8
安徽省	4575.99	4608.83	4580.84	4538.29	4504.14	4437.46	4400.34	4331.69	4305.53	3585.09	3547.65
福建省	1090.79	1110.42	1076.78	1085.18	1064.84	1055.37	1061.65	1116.12	1122.42	968.51	967.48
江西省	2035.55	2038.49	2036.14	2032.02	2039.42	2036.83	2027.67	2001.57	1995.6	1907.06	1867.67
山东省	5266.97	5293.56	5271.37	5235.99	5191.06	5161.16	4964.43	4901.95	4729.03	5058.11	4986.88
河南省	5360.24	5463.07	5328.95	5288.69	5273.63	5242.92	5210.64	5101.15	4969.11	5205.63	5150.44
湖北省	2995.64	3086.04	2968.99	2931.9	2919.17	2905.57	2899.15	2855.32	2791.41	2548.91	2455.69
湖南省	3177.67	3192.9	3176.11	3164	3145.87	3132.37	3113.32	3101.7	3084.3	2715.78	2762.41
广东省	1775.02	1776.46	1773.4	1775.21	1774.61	1771.71	1771.26	1770.99	1770.76	1874.44	1873.16
广西	1717.00	1731.03	1713.08	1706.88	1669.87	1646.07	1618.79	1600	1586.37	1541.29	1529.24
海南省	291.11	292.21	290.63	290.48	289.25	289.95	263.99	259.92	260.93	256.75	247.51
重庆市	697.65	698.32	697.68	696.94	694.26	690.6	687.19	677.26	675.18	702.97	692.88
四川省	2959.62	2992.24	2954.09	2932.54	2873.1	2813.55	2735.09	2666.32	2616.54	2662.65	2600.75
贵州省	1150.57	1165.49	1153.99	1132.24	1114.12	1088.07	1065.43	981.83	926.9	1214.57	1201.19
云南省	1932.88	1978.07	1922.49	1898.07	1851.42	1809.39	1757.71	1708.97	1660.27	1677.9	1634.24
西藏	274.43	282.82	275.94	264.53	261.23	251.53	247.8	244.03	239.27	251.04	245.31
陕西省	1298.99	1336.81	1285.16	1274.99	1263.09	1251.39	1236.77	1226.49	1209.94	1277.18	1274.34
甘肃省	1335.00	1338.6	1328.86	1337.54	1331.43	1317.51	1306.72	1297.06	1284.08	1297.58	1291.82
青海省	215.54	219.25	213.33	214.04	206.61	202.35	196.99	182.49	186.9	251.67	251.67
宁夏	538.09	552.53	538.29	523.45	511.45	515.15	506.53	498.91	498.56	491.35	477.59
新疆	4912.23	4893.35	4959.89	4883.46	4952.29	4982.03	4944.92	4831.89	4769.89	4029.07	3884.57

需要注意的是，均值填充的方法一般用于不存在上升/下降趋势的数据，对于有变化趋势的数据，使用均值填充的方法不能够很好的利用数据本身的变化特点，因此，使用线性回归的方式对数据进行拟合，再使用拟合的系数进行数据填充。

假设有一组数据 x 和 y ，先构造一个简单的一元线性回归方程：

$$y = a_0 + a_1 x_1$$

给定初始的系数 a_0 和 a_1 ，根据 x 计算每组数据中的 y 值（该 y 值为预测值），将预测值与真实数据进行比较计算误差，再通过下山法不断的减小误差，最终使得系数 a_0 和 a_1 能够拟合数据 x 和 y 。再将需要预测的 x 值传入回归方程，即可预测缺失的数据。

表 3-10 表示的是农村居民的人均可支配收入，可以明显看出数据的上升趋势，如果使用均值预测缺失值，则预测的数据与变化趋势有较大的差异，因此这里使用回归方程进行先拟合再预测。

表 3-10 互联网宽带接入端口数（万个）

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市		2084.1	2060.1	2059.9	1818	1784	1580.5	1159.9	1186.8	1076.8	801.8
天津市		1254.6	1092.6	897.4	795.3	724.3	470.2	397.7	353.9	461.7	367.9
河北省		4598.2	4345.8	4192.4	4126.9	3841.1	2948.5	2204.7	2049.3	1756	1150.4
山西省		2322.3	2148.2	1989.2	1840.1	1582.9	1345.9	996.8	881.7	737.4	639.2
内蒙古		1441.3	1372.7	1358.4	1294.1	1200.7	916	739.9	677.2	548.7	409.3
辽宁省		3305.2	3270.7	3240.3	3118.8	3239.5	2710.8	2083.5	1631.4	1267.5	981.3
吉林省		1651.2	1686.6	1526.7	1761.1	1560.7	987.2	806	695	635.6	498.8
黑龙江省		2127.5	2181.1	2113.2	1937.2	1964.9	1308.8	1044	865.2	771.7	640.6
上海市		2322	2028.6	1871.8	1810.2	1595.7	1464.8	1404.5	1374.6	1491.5	706.4
江苏省		7224.9	7249	7131.5	6531.7	5676.8	4697.3	3503.3	3063.2	2892.4	2064.2
浙江省		6031.5	6284.4	5971	5455.1	4720.6	4768.9	2632.6	2447.4	2250.4	1458.2
安徽省		3543.7	3481.3	3374	2872.2	2527.3	2211.5	1175.7	1124.6	1063.8	732.8
福建省		3370	3232.1	3245	2861.8	2482.3	2335.3	1618.7	1439.4	1439.8	910.3
江西省		2532.9	2369.5	2032.7	1985.9	2055.6	1693.2	975.1	868.6	719.3	562
山东省		6756.8	6915.2	6312.3	5596.9	4680	4003.4	2949.1	2537	2144.9	1756.4
河南省		4934.5	4752.8	4780.8	4475.8	4345.8	3241.9	2016.4	1747	1435.7	1104.1
湖北省		3221.4	3062.3	2961.3	2605.5	2594.7	2061.1	1266.1	1153.2	1023.5	746.1
湖南省		3242.4	2997.9	2821.1	2436	2395.3	1931.8	1498.3	1219.6	1070.2	780.3
广东省		8653.2	8538	8149.1	6482.3	6515.6	4765.5	3597.7	3325.2	3158.1	2519.4
广西		3356.2	3023	2760.1	2216.4	2094.9	1530.9	1126.2	978.8	848	590.2
海南省		852.3	794.1	726.1	571.6	522.9	398.4	223.9	198.6	168.9	127.6
重庆市		2368.8	2318.2	2245.7	1935.2	1643.6	1349.4	963.4	778.8	648.8	482.7
四川省		6284.7	5864	5400.5	4702.8	3709.6	3117.9	2200.4	1775.7	1316.3	979.1
贵州省		1744.5	1759.9	1535.4	1325.6	1113.9	875.8	580.6	494.2	466.8	427
云南省		2218	2091.1	1962.6	1661.8	1674.4	1151.1	741.3	736.9	663.7	443.8
西藏		219	209.1	194.3	154.6	107.2	51	48.1	44.1	41	26.7
陕西省		2576.5	2322.7	2239.6	1993.2	2083.1	1539.3	1070.4	940.9	803.6	621.2
甘肃省		1460.6	1405.7	1128	1099.9	946	820.5	473.2	460.3	430.8	251.4
青海省		412.8	382.3	355.4	310.5	262.2	208.1	134.2	117.8	105.6	62
宁夏		550	520	498.2	415	307.1	200.8	180.3	150.3	122.1	79
新疆		1943.8	1819	1678.6	1407.4	1323.9	1023.8	734.3	628.6	548.1	319.1

对表 3-10 中的数据进行线性回归拟合，再将 2021 年的预测值填充进缺失部分，填充后的数据如表 3-11，利用回归方程拟合的系数进行数据填充的时候，可以很好的反应数据本身的变化趋势。

表 3-11 互联网接入端口数（数据填充后）

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	2092.2	2084.1	2060.1	2059.9	1818	1784	1580.5	1159.9	1186.8	1076.8	801.8
天津市	1373.7	1254.6	1092.6	897.4	795.3	724.3	470.2	397.7	353.9	461.7	367.9
河北省	4733.5	4598.2	4345.8	4192.4	4126.9	3841.1	2948.5	2204.7	2049.3	1756	1150.4
山西省	2433.3	2322.3	2148.2	1989.2	1840.1	1582.9	1345.9	996.8	881.7	737.4	639.2
内蒙古	1468.9	1441.3	1372.7	1358.4	1294.1	1200.7	916	739.9	677.2	548.7	409.3
辽宁省	3326.8	3305.2	3270.7	3240.3	3118.8	3239.5	2710.8	2083.5	1631.4	1267.5	981.3
吉林省	1692.7	1651.2	1686.6	1526.7	1761.1	1560.7	987.2	806	695	635.6	498.8
黑龙江省	2132.3	2127.5	2181.1	2113.2	1937.2	1964.9	1308.8	1044	865.2	771.7	640.6

上海市	2472.1	2322	2028.6	1871.8	1810.2	1595.7	1464.8	1404.5	1374.6	1491.5	706.4
江苏省	7256.0	7224.9	7249	7131.5	6531.7	5676.8	4697.3	3503.3	3063.2	2892.4	2064.2
浙江省	6051.7	6031.5	6284.4	5971	5455.1	4720.6	4768.9	2632.6	2447.4	2250.4	1458.2
安徽省	3600.3	3543.7	3481.3	3374	2872.2	2527.3	2211.5	1175.7	1124.6	1063.8	732.8
福建省	3411.7	3370	3232.1	3245	2861.8	2482.3	2335.3	1618.7	1439.4	1439.8	910.3
江西省	2699.6	2532.9	2369.5	2032.7	1985.9	2055.6	1693.2	975.1	868.6	719.3	562
山东省	6905.0	6756.8	6915.2	6312.3	5596.9	4680	4003.4	2949.1	2537	2144.9	1756.4
河南省	4985.7	4934.5	4752.8	4780.8	4475.8	4345.8	3241.9	2016.4	1747	1435.7	1104.1
湖北省	3308.1	3221.4	3062.3	2961.3	2605.5	2594.7	2061.1	1266.1	1153.2	1023.5	746.1
湖南省	3382.8	3242.4	2997.9	2821.1	2436	2395.3	1931.8	1498.3	1219.6	1070.2	780.3
广东省	8821.2	8653.2	8538	8149.1	6482.3	6515.6	4765.5	3597.7	3325.2	3158.1	2519.4
广西区	3554.9	3356.2	3023	2760.1	2216.4	2094.9	1530.9	1126.2	978.8	848	590.2
海南省	894.4	852.3	794.1	726.1	571.6	522.9	398.4	223.9	198.6	168.9	127.6
重庆市	2409.8	2368.8	2318.2	2245.7	1935.2	1643.6	1349.4	963.4	778.8	648.8	482.7
四川省	6579.4	6284.7	5864	5400.5	4702.8	3709.6	3117.9	2200.4	1775.7	1316.3	979.1
贵州省	1814.2	1744.5	1759.9	1535.4	1325.6	1113.9	875.8	580.6	494.2	466.8	427
云南省	2303.1	2218	2091.1	1962.6	1661.8	1674.4	1151.1	741.3	736.9	663.7	443.8
西藏	227.2	219	209.1	194.3	154.6	107.2	51	48.1	44.1	41	26.7
陕西省	2688.8	2576.5	2322.7	2239.6	1993.2	2083.1	1539.3	1070.4	940.9	803.6	621.2
甘肃省	1571.5	1460.6	1405.7	1128	1099.9	946	820.5	473.2	460.3	430.8	251.4
青海省	431.9	412.8	382.3	355.4	310.5	262.2	208.1	134.2	117.8	105.6	62
宁夏回族自治区	567.3	550	520	498.2	415	307.1	200.8	180.3	150.3	122.1	79
新疆	2032.2	1943.8	1819	1678.6	1407.4	1323.9	1023.8	734.3	628.6	548.1	319.1

3.3 数据规约

对数据进行清洗和填充之后，下一步需要对采集到的数据进行规约化，可以看出采集到的数据具有不同的量纲，不同的范围，为了提高分析数据的准确性，需要把所有的数据规约到一个确定的大小，因此对采集到的所有数据进行数据规约化，论文采用的方式是 $MaxMin$ 的规约方式，即通过每一组数据的最大值和最小值对该组数据进行归一化。

对于任意一组数据 x ， $MaxMin$ 规约方式的对于效益型指标的计算方式为：

$$x_{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

对于成本型指标的计算方式为：

$$x_{new} = \frac{x_{max} - x_i}{x_{max} - x_{min}}$$

使用 $MaxMin$ 的规约方式，可以确保规约后的数据与规约之前的数据具有同样的变化趋势，并且可以让数据均匀的分布在 0-1 的空间之内。

原始数据的范围从几十到上万不等，如果进行数据分析的话，相关性会依托有限的几个大数据上，使得分析不够合理，因此需要采用规约化的操作，来对数据进行缩放。规约化后的结果如表 3-12 所示：

表 3-12 全国各省份粮食总产量（规约化后）

地区	2021 年	2020 年	2019 年	2018 年	2017 年	2016 年	2015 年	2014 年	2013 年	2012 年	2011 年
北京市	0.03	0.04	0.00	0.06	0.13	0.22	0.33	0.42	0.59	0.75	0.83
天津市	0.08	0.08	0.00	0.17	0.26	0.30	0.20	0.29	0.28	0.34	0.29
河北省	0.06	0.00	0.05	0.13	0.35	0.45	0.44	0.43	0.39	0.32	0.22
山西省	0.16	0.16	0.12	0.20	0.25	0.28	0.33	0.45	0.46	0.52	0.52
内蒙古	0.95	0.96	0.96	0.94	1.00	0.98	0.82	0.71	0.63	0.59	0.55
辽宁省	0.86	0.93	0.83	0.82	0.77	0.87	1.00	0.84	0.74	0.66	0.52
吉林省	0.98	1.00	0.98	0.95	0.96	0.94	0.90	0.82	0.65	0.51	0.42
黑龙江省	0.98	1.00	0.97	0.95	0.97	0.98	0.98	0.92	0.86	0.81	0.77
上海市	0.05	0.00	0.03	0.12	0.13	0.22	0.43	0.52	0.61	0.66	0.74
江苏省	0.41	0.41	0.34	0.49	0.55	0.71	0.81	0.66	0.63	0.58	0.47
浙江省	0.05	0.06	0.05	0.03	0.03	0.00	0.03	0.04	0.10	0.13	0.29
安徽省	0.07	0.10	0.06	0.05	0.00	0.07	1.00	0.89	0.73	0.70	0.76
福建省	0.05	0.07	0.05	0.03	0.00	0.00	0.06	0.11	0.17	0.19	0.30
江西省	0.83	0.94	0.76	0.81	0.93	0.97	1.00	0.97	0.93	0.87	0.79

山东省	0.44	0.34	0.40	0.59	0.63	0.86	1.00	0.91	0.67	0.30	0.19
河南省	0.89	0.86	0.88	0.92	0.89	1.00	0.99	0.89	0.80	0.67	0.66
湖北省	0.94	0.99	0.85	0.97	0.97	0.93	1.00	0.82	0.77	0.72	0.51
湖南省	0.83	1.00	0.75	0.74	0.93	0.95	0.96	1.00	0.96	0.91	0.80
广东省	0.24	0.36	0.24	0.13	0.07	0.00	0.02	0.06	0.00	0.09	0.09
广西	0.50	0.59	0.46	0.45	0.44	0.44	0.56	0.32	0.56	0.51	0.37
海南省	0.05	0.00	0.00	0.16	0.14	0.24	0.35	0.45	0.42	0.48	0.46
重庆市	0.72	0.76	0.69	0.70	0.68	0.66	0.60	0.55	0.62	0.63	0.39
四川省	0.81	1.00	0.77	0.66	0.60	0.49	0.42	0.32	0.31	0.23	0.08
贵州省	0.85	0.85	0.85	0.85	1.00	0.95	0.90	0.88	0.78	0.62	0.45
云南省	0.96	1.00	0.96	0.92	0.84	0.84	0.86	0.88	0.89	0.74	0.59
西藏	0.99	1.00	0.99	0.97	0.69	0.85	0.28	0.24	0.21	0.12	0.10
陕西省	0.57	0.70	0.58	0.42	0.32	0.70	0.26	0.28	0.49	0.78	0.83
甘肃省	0.72	1.00	0.68	0.49	0.42	0.41	0.47	0.50	0.51	0.48	0.49
青海省	0.90	1.00	0.83	0.86	0.85	0.87	0.88	0.83	0.85	0.84	0.70
宁夏	0.61	0.70	0.50	0.61	0.31	0.18	0.31	0.26	0.45	0.72	1.00
新疆	0.96	1.00	0.96	0.92	0.86	0.87	0.61	0.57	0.45	0.43	0.36

4 问题一分析与模型建立

4.1 问题分析

题目一由三个子问题构成，一是要求从指定维度出发，分别构建乡村振兴与数字经济系统的评价模型；二是建立乡村振兴-数字经济耦合模型；三是根据实证研究结果分析当前乡村振兴-数字经济耦合系统中存在的问题，并针对性地提出政策性建议。

针对第一个子问题，我们首先需要客观地量化各省区的乡村振兴与数字经济发展水平，建立乡村振兴和数字经济的评价指标体系。因此，我们将从 8 个维度筛选了刻画乡村振兴和数字经济发展水平的 27 个指标，并采用熵权法得到各省区的乡村振兴发展指数和数字经济发展指数。

针对第二个子问题，我们将基于子问题一中得到的发展指数，利用耦合协调度计算模型计算各省区历年的耦合度和耦合协调度。为分区域研究我国的耦合差异，我们以地理方位为依据将我国划分为八个地区，以各地区所含省份的平均耦合度及耦合协调度表征该地区的耦合关系。

最后，通过分析各地区的耦合度变化趋势、比较不同地区的耦合度差异，剖析我国乡村振兴和数字经济耦合的主要问题，并提出改进建议。

4.2 耦合关系模型建立

4.2.1 评价模型——熵值法

熵值法是综合评价中的一种常用的评价方法。其中，熵是度量系统无序程度的指标，熵越大意味着系统的无序度越低，所能提供的信息越多。熵值法则是以各指标所含信息的多少为依据计算指标权值的一种客观的赋权方法。对于指标来说，熵越大，表示该指标的权值越小，变异程度越小，对评价整体的影响越小。

具体步骤如下：

假设 X_{ij} 为第 i 年第 j 个指标的值， X'_{ij} 是指标标准化后的结果， m 为年数， n 为指标数。本文中 m 为 20， n 为 27。

①指标标准化

$$\text{对于正向指标: } X'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

$$\text{对于负向指标: } X'_{ij} = \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)}$$

②计算第 i 年第 j 项指标的比重 p_{ij}

$$p_{ij} = \frac{X'_{ij}}{\sum_{i=1}^m X'_{ij}}$$

③计算第 j 项指标的熵值 e_j

$$e_j = -k \sum_{i=1}^m p_{ij} \ln p_{ij} \quad , \quad \text{其中 } k = \frac{1}{\ln m}$$

④计算第 j 项指标的差异系数 g_j

$$g_j = 1 - e_j$$

⑤给指标赋权，定义权重 w_j

$$w_j = \frac{g_j}{\sum_{j=1}^n g_j}$$

⑥通过权重计算样本评价值，第 i 年第 j 项指标的评价值为：

$$F_{ij} = w_j X_{ij}'$$

则第 i 年的总体评价值为：

$$F_i = \sum_{j=1}^n F_{ij}$$

4.2.2 耦合模型——耦合协调度计算模型

耦合度 C ：

耦合度测量的是两个及以上的物体或系统运动协调的程度。耦合度 C 的值介于 0 到 1 之间。 C 越接近 0，表示各系统间的耦合度越小，反之趋向 1，意味着系统间的耦合度越大。

耦合协调度 D ：

耦合协调度是用于评价系统之间相互影响关系。耦合协调度 D 的取值范围同样为 0 到 1。

考虑到乡村振兴与数字经济之间存在的是非线性的耦合关系，将乡村振兴-数字经济耦合系统的协调度模型构建为：

$$D = \sqrt{C \times T}，其中：$$

$$C = \left\{ \frac{f(x) \cdot g(y)}{\left[\frac{f(x) + g(y)}{2} \right]^2} \right\}^{1/2}$$

$$T = \alpha f(x) + \beta g(y)$$

其中： D 代表系统之间的耦合协调度， C 代表系统之间的耦合度， T 代表综合评价指数， α 和 β 分别代表乡村振兴子系统与数字经济子系统的权重，均设为常数 0.5。

4.3 模型求解

4.3.1 熵值法求解

基于熵值法构建的乡村振兴-数字经济子系统评价模型如图 4-1 所示。

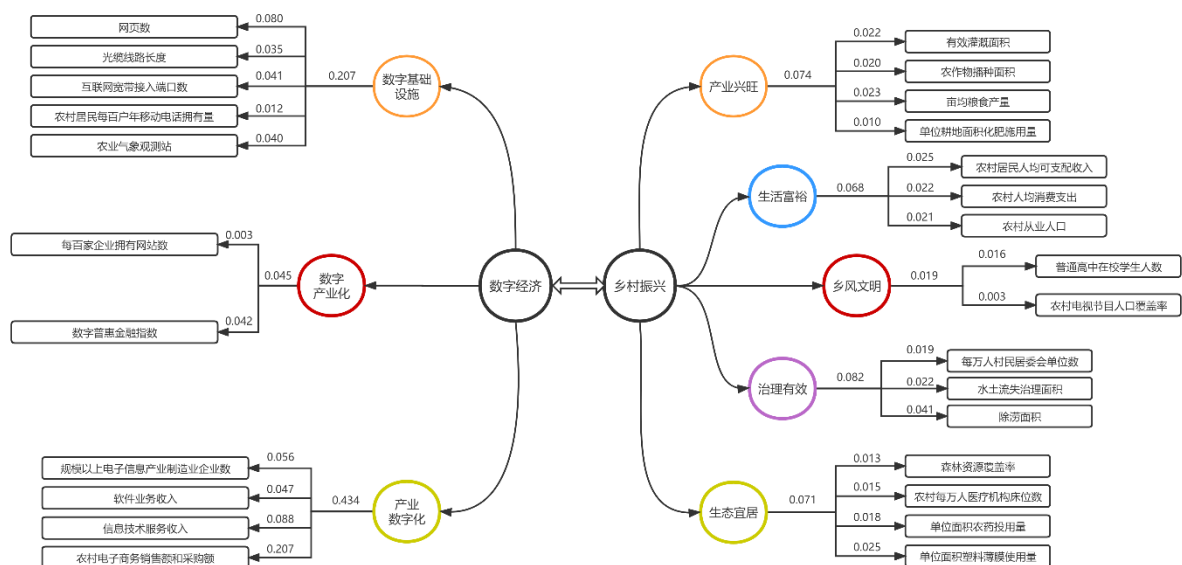


图 4-1 乡村振兴-数字经济评价体系

继而依据所求得的权重分别计算得到乡村振兴子系统发展指数 $f(x)$ 与数字经济子系统发展指数 $g(y)$ 。由于篇幅有限，表 4-1 仅展示了 2015 年-2021 年的计算结果，完整结果见附件二。

表 4-1 乡村振兴-数字经济发展指数

年份 地区	2015	2016	2017	2018	2019	2020	2021
北京	0.233/0.585	0.232/0.583	0.234/0.582	0.232/0.577	0.231/0.582	0.225/0.581	0.197/0.547
天津	0.288/0.552	0.289/0.533	0.280/0.536	0.275/0.557	0.274/0.561	0.275/0.572	0.173/0.517
河北	0.170/0.254	0.178/0.345	0.186/0.446	0.174/0.445	0.181/0.482	0.203/0.446	0.219/0.501
山西	0.185/0.152	0.194/0.227	0.200/0.255	0.203/0.346	0.214/0.435	0.228/0.487	0.240/0.607
内蒙古	0.238/0.334	0.242/0.292	0.254/0.281	0.258/0.281	0.265/0.347	0.277/0.389	0.289/0.430
辽宁	0.268/0.335	0.288/0.310	0.301/0.341	0.315/0.349	0.318/0.370	0.320/0.391	0.350/0.421
吉林	0.316/0.308	0.331/0.339	0.355/0.389	0.358/0.421	0.365/0.385	0.409/0.426	0.418/0.451
黑龙江	0.308/0.257	0.337/0.300	0.357/0.330	0.363/0.291	0.366/0.296	0.391/0.332	0.406/0.349
上海	0.232/0.571	0.233/0.583	0.231/0.582	0.228/0.582	0.227/0.582	0.225/0.582	0.134/0.548
江苏	0.242/0.333	0.258/0.357	0.295/0.398	0.312/0.441	0.328/0.463	0.345/0.459	0.341/0.487
浙江	0.331/0.334	0.356/0.344	0.357/0.376	0.356/0.416	0.330/0.454	0.354/0.423	0.366/0.445
安徽	0.307/0.269	0.323/0.291	0.335/0.334	0.349/0.395	0.363/0.435	0.396/0.453	0.394/0.500
福建	0.318/0.362	0.348/0.383	0.354/0.416	0.356/0.446	0.368/0.474	0.404/0.448	0.410/0.468
江西	0.345/0.288	0.361/0.289	0.371/0.334	0.378/0.385	0.386/0.414	0.412/0.445	0.426/0.493
山东	0.275/0.308	0.294/0.354	0.307/0.381	0.312/0.408	0.324/0.381	0.352/0.418	0.367/0.434
河南	0.299/0.643	0.300/0.643	0.299/0.643	0.298/0.639	0.297/0.614	0.298/0.625	0.274/0.631
湖北	0.301/0.302	0.317/0.333	0.329/0.363	0.332/0.403	0.342/0.442	0.351/0.465	0.371/0.506
湖南	0.297/0.643	0.290/0.622	0.291/0.605	0.291/0.571	0.290/0.601	0.289/0.605	0.254/0.595
广东	0.348/0.336	0.368/0.354	0.379/0.381	0.363/0.451	0.372/0.469	0.399/0.418	0.410/0.437
广西	0.325/0.234	0.358/0.261	0.370/0.302	0.389/0.358	0.413/0.417	0.437/0.440	0.454/0.500
海南	0.281/0.332	0.327/0.342	0.367/0.339	0.375/0.383	0.396/0.414	0.425/0.392	0.425/0.417
重庆	0.297/0.324	0.305/0.359	0.314/0.399	0.330/0.433	0.323/0.451	0.333/0.455	0.356/0.483
四川	0.343/0.284	0.360/0.305	0.373/0.346	0.367/0.410	0.366/0.455	0.400/0.457	0.412/0.503
贵州	0.349/0.317	0.356/0.341	0.371/0.384	0.380/0.440	0.371/0.466	0.393/0.450	0.406/0.478
云南	0.336/0.263	0.361/0.271	0.386/0.325	0.398/0.381	0.421/0.418	0.445/0.442	0.451/0.491
西藏	0.289/0.242	0.276/0.265	0.245/0.246	0.243/0.260	0.233/0.197	0.242/0.190	0.235/0.176

山西	0.319/0.317	0.334/0.362	0.336/0.399	0.346/0.448	0.295/0.467	0.325/0.466	0.359/0.497
甘肃	0.341/0.300	0.340/0.308	0.336/0.351	0.346/0.382	0.372/0.396	0.398/0.413	0.400/0.446
青海	0.251/0.537	0.252/0.523	0.252/0.338	0.257/0.387	0.243/0.438	0.245/0.459	0.237/0.477
宁夏	0.315/0.272	0.331/0.294	0.345/0.346	0.349/0.390	0.365/0.391	0.388/0.383	0.398/0.410
新疆	0.312/0.333	0.310/0.349	0.330/0.348	0.338/0.444	0.349/0.434	0.358/0.430	0.383/0.460

4.3.2 耦合协调度计算模型求解

表 4-2 为依据耦合协调度计算模型得到的乡村振兴-数字经济耦合系统的耦合度和耦合协调度结果，由于篇幅有限仅展示了 2015 年到 2021 年的计算结果，完整结果见附件二。

表 4-2 乡村振兴-数字经济耦合系统的耦合度与耦合协调度

年份 地区	2015	2016	2017	2018	2019	2020	2021
北京	0.902/0.744	0.902/0.742	0.904/0.743	0.904/0.741	0.902/0.742	0.897/0.739	0.882/0.715
天津	0.949/0.750	0.955/0.744	0.949/0.743	0.941/0.749	0.939/0.750	0.936/0.754	0.867/0.692
河北	0.980/0.617	0.948/0.646	0.911/0.678	0.899/0.672	0.890/0.686	0.927/0.687	0.919/0.710
山西	0.995/0.573	0.996/0.612	0.992/0.624	0.965/0.656	0.940/0.687	0.932/0.709	0.901/0.754
内蒙古	0.985/0.660	0.995/0.644	0.998/0.640	0.999/0.640	0.990/0.669	0.985/0.687	0.980/0.705
辽宁	0.993/0.664	0.999/0.654	0.998/0.669	0.998/0.673	0.997/0.683	0.995/0.693	0.995/0.708
吉林	0.999/0.654	0.999/0.669	0.998/0.694	0.996/0.709	0.999/0.692	0.999/0.713	0.999/0.725
黑龙江	0.995/0.626	0.998/0.649	0.999/0.665	0.994/0.642	0.994/0.645	0.996/0.664	0.997/0.673
上海	0.906/0.739	0.903/0.743	0.902/0.742	0.899/0.740	0.898/0.740	0.897/0.739	0.796/0.672
江苏	0.987/0.660	0.987/0.672	0.988/0.693	0.985/0.713	0.985/0.724	0.989/0.724	0.984/0.735
浙江	0.999/0.667	0.999/0.672	0.999/0.688	0.996/0.706	0.987/0.720	0.996/0.709	0.995/0.720
安徽	0.997/0.633	0.998/0.645	0.999/0.667	0.998/0.696	0.996/0.715	0.997/0.725	0.993/0.746
福建	0.997/0.680	0.998/0.690	0.996/0.706	0.993/0.720	0.992/0.733	0.998/0.723	0.997/0.733
江西	0.996/0.642	0.993/0.641	0.998/0.666	0.999/0.692	0.999/0.707	0.999/0.722	0.997/0.745
山东	0.998/0.653	0.995/0.675	0.994/0.687	0.991/0.699	0.996/0.689	0.996/0.707	0.996/0.715
河南	0.930/0.787	0.931/0.787	0.931/0.787	0.931/0.785	0.937/0.776	0.935/0.780	0.918/0.775
湖北	0.999/0.651	0.999/0.666	0.998/0.681	0.995/0.699	0.991/0.717	0.990/0.727	0.988/0.747
湖南	0.929/0.786	0.931/0.776	0.936/0.770	0.945/0.758	0.936/0.769	0.935/0.770	0.915/0.755
广东	0.999/0.668	0.999/0.677	0.999/0.690	0.994/0.722	0.993/0.731	0.999/0.708	0.999/0.718
广西	0.986/0.610	0.987/0.624	0.994/0.648	0.999/0.678	0.999/0.708	0.999/0.720	0.998/0.749
海南	0.996/0.664	0.999/0.671	0.999/0.669	0.999/0.691	0.999/0.707	0.999/0.696	0.999/0.708
重庆	0.999/0.661	0.996/0.678	0.992/0.696	0.990/0.712	0.986/0.718	0.987/0.721	0.988/0.735
四川	0.995/0.640	0.996/0.651	0.999/0.673	0.998/0.704	0.994/0.724	0.997/0.727	0.995/0.749
贵州	0.998/0.658	0.999/0.670	0.999/0.692	0.997/0.718	0.993/0.730	0.997/0.724	0.996/0.737
云南	0.992/0.627	0.989/0.630	0.996/0.661	0.999/0.690	0.999/0.709	0.999/0.721	0.999/0.745
西藏	0.996/0.619	0.999/0.632	0.999/0.623	0.999/0.629	0.996/0.596	0.992/0.591	0.989/0.583
山西	0.999/0.658	0.999/0.680	0.996/0.697	0.991/0.720	0.974/0.720	0.983/0.725	0.986/0.742
甘肃	0.998/0.649	0.998/0.653	0.999/0.675	0.998/0.690	0.999/0.698	0.999/0.706	0.998/0.722
青海	0.932/0.734	0.936/0.730	0.989/0.664	0.979/0.683	0.958/0.698	0.952/0.706	0.941/0.709
宁夏	0.997/0.634	0.998/0.646	0.999/0.673	0.998/0.694	0.999/0.695	0.999/0.691	0.999/0.705
新疆	0.999/0.666	0.998/0.673	0.999/0.674	0.990/0.717	0.994/0.714	0.995/0.713	0.995/0.728

4.4 结果分析

为比较我国不同地区乡村振兴、数字经济发展指数以及二者之间的耦合关系，本文按照区域地理方位将我国划分为华东、华南、华北、华中、西南、西北、东北七个地区，

以各地区所包含省（直辖市、自治区）的均值作为地区的指标，并将所有省份的均值作为全国性指标。

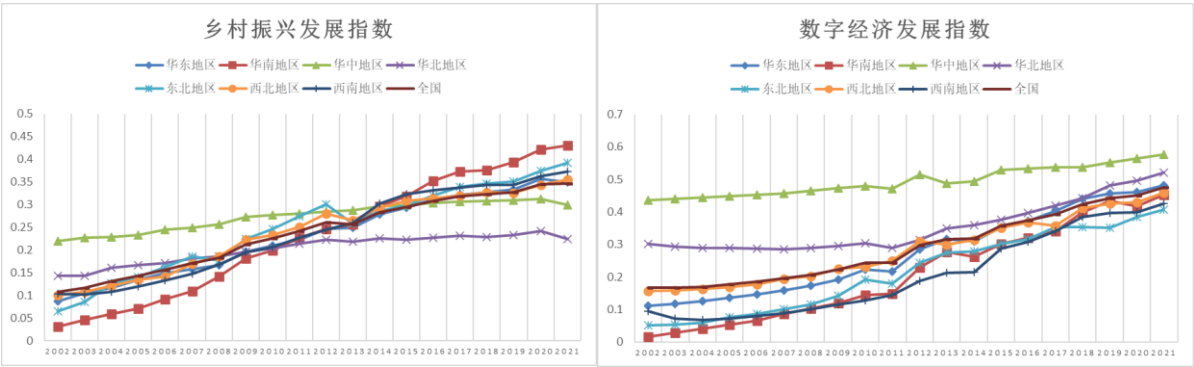


图 4-2 乡村振兴-数字经济发展指数

图 4-2 反应了不同地区乡村振兴发展指数与数字经济发展指数的变化趋势，总体而言，各地区的乡村振兴发展指数和数字经济发展指数均呈上升态势，且在数值上乡村振兴发展指数的值普遍小于数字经济发展指数，这表明各地区的数字经济发展水平优于乡村振兴发展水平，但二者的差距整体呈缩小的趋势。从不同地区来看，华北地区的乡村振兴发展指数变化较为平缓，近 20 年的涨幅不大，从领先水平退至落后位置，与其他地区存在较大差距，因此需要关注华北地区的乡村振兴发展状况，加强政府的宏观调控。此外，华南地区的乡村振兴发展指数与数字经济发展指数的增速明显高于其他地区，从垫底位置逐步升至全国领先的位置，其余地区可以借鉴华南地区的发展经验，结合各地实际，调整政策方向、力度。

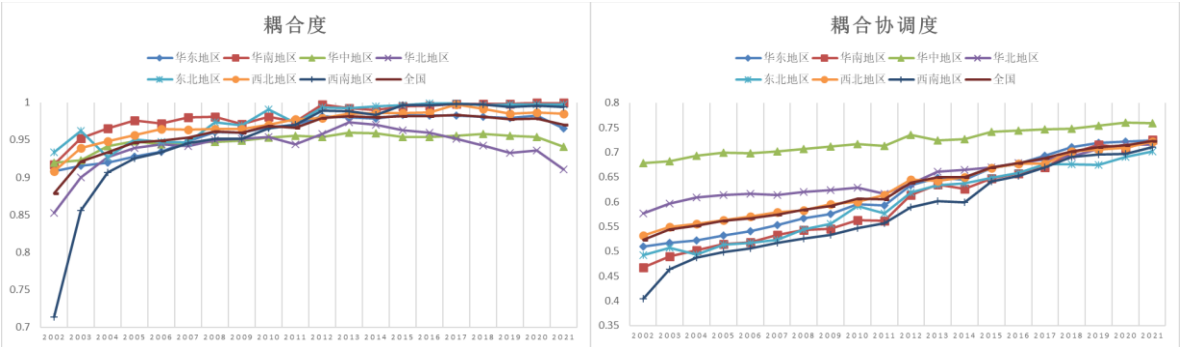


图 4-3 耦合关系图

图 4-3 展示了各地区乡村振兴子系统和数字经济子系统的耦合关系的演变趋势。从变化趋势来看，除西南地区在 2002 年到 2006 年间出现大幅增长外，大部分地区的耦合度变化较小，基本能维持在 0.9 以上，说明在我国乡村振兴与数字经济之间的相互作用强度较大，二者之间存在明显且稳定的影响关系。这一变化趋势与耦合协调度相反，从图中可以看到，除华中地区的耦合协调度变化平稳，且始终保持领先优势外，其余地区的耦合协调度均呈现波动上升的趋势，在过去 20 年里耦合协调度有着明显的提高。这表明我国乡村振兴--数字经济耦合系统的演化方向较为稳定，两个子系统的演化速度、方向的同步性稳态提升，且未来持续向好。从地区来看，西南地区的耦合度和耦合协调度均落后于其他地区，始终位于全国平均水平以下，因此政府需要加强对西南地区的政策引导，采取科学有效的措施来提高西南地区的耦合协调度。

表 4-3 西南地区结果

指标	省	2002	2004	2006	2008	2010	2012	2014	2016	2018	2020
	重庆	0.131	0.081	0.113	0.169	0.213	0.250	0.282	0.306	0.330	0.333
	四川	0.040	0.085	0.125	0.157	0.205	0.246	0.320	0.360	0.368	0.401

乡村振兴发展指数	贵州	0.041	0.053	0.092	0.139	0.179	0.235	0.315	0.357	0.381	0.394
	云南	0.013	0.046	0.081	0.117	0.175	0.238	0.312	0.362	0.398	0.446
	西藏	0.291	0.275	0.253	0.255	0.257	0.255	0.278	0.277	0.243	0.242
数字经济发展指数	重庆	0.009	0.026	0.046	0.074	0.105	0.151	0.227	0.359	0.434	0.456
	四川	0.007	0.023	0.043	0.072	0.109	0.209	0.219	0.306	0.410	0.458
	贵州	0.002	0.017	0.038	0.067	0.106	0.211	0.238	0.341	0.440	0.450
	云南	0.030	0.041	0.057	0.078	0.112	0.184	0.238	0.272	0.381	0.443
	西藏	0.427	0.236	0.213	0.209	0.204	0.188	0.154	0.265	0.260	0.191
耦合度	重庆	0.502	0.863	0.908	0.919	0.941	0.969	0.994	0.997	0.991	0.988
	四川	0.713	0.816	0.872	0.927	0.952	0.997	0.982	0.997	0.998	0.998
	贵州	0.447	0.859	0.908	0.937	0.967	0.999	0.990	1.000	0.997	0.998
	云南	0.923	0.999	0.984	0.980	0.975	0.992	0.991	0.990	1.000	1.000
	西藏	0.982	0.997	0.996	0.995	0.993	0.989	0.958	1.000	0.999	0.993
耦合协调度	重庆	0.256	0.445	0.477	0.496	0.523	0.560	0.610	0.678	0.712	0.722
	四川	0.360	0.419	0.457	0.499	0.530	0.603	0.601	0.651	0.704	0.728
	贵州	0.224	0.438	0.473	0.502	0.536	0.605	0.614	0.671	0.719	0.724
	云南	0.477	0.520	0.520	0.529	0.543	0.588	0.615	0.631	0.691	0.721
	西藏	0.704	0.616	0.604	0.602	0.599	0.588	0.556	0.632	0.630	0.592

表 4-3 列出了西南地区各指标的具体数值，可见，西藏的发展趋势与其余四省存在较大差异，具体而言，其余四省四个指标在近 20 年里都涨幅明显，而西藏却呈现下降趋势。从具体指标来看，西藏的乡村振兴发展指数变化较小，与 2002 年相比略有降低，但是在地区的发展排名明显下滑，由 2002 年的大幅领先降到如今的末位水平，说明西藏在乡村振兴方面缺乏增长动力。数字经济发展指数与乡村振兴发展指数类似，也是逐渐失去领先优势，不同之处在于数字经济发展指数下降得非常快，从 0.427 降至 0.191，而其余省均保持稳定增速，因此迅速拉开了差距。从耦合关系来看，西藏的耦合度很高，稳定在 0.95 以上，说明西藏的乡村振兴发展与数字经济发展关系密切，二者之间存在显著的相互作用。但西藏的耦合协调度逐年降低，从 0.704 降到了 0.592，这表明乡村振兴子系统和数字经济子系统的协调状况较差，未能实现高水平的相互促进。

为提高西藏的耦合协调度，促进西藏乡村振兴和数字经济的协调发展，可从以下角度进行改进。一方面，可以利用时下热门的数字电商来推动西藏的产业发展，比如采用直播等形式宣传西藏的特色产品、旅游景点等，拓宽西藏的产品销售渠道，带动旅游业的发展，提高当地人的生产生活质量，用数字经济这一新引擎有效助力乡村振兴。另一方面，要加强西藏的数字基础设施建设，加大财政投资力度，夯实数字化基础条件，不仅能为数字经济的发展提供有力支撑、实现高质量发展，还能丰富居民的日常生活、改善生态环境，实现数字经济与乡村振兴的协同发展。总的来说，采取的政策需要以乡村振兴与数字经济的作用机理为依据，从二者相互作用的路径出发，选择科学有效的措施来提高西藏的耦合协调度，进而实现西南地区的协调发展。

5 问题二分析与模型建立

5.1 问题分析

该问题包含多个方面，本文将其拆分成三个子问题，逐个分析解答，具体如下：

首先，结合问题一，本文选取乡村振兴的 5 个维度：产业兴旺、生态宜居、乡风文明、治理有效、生活富裕，数字经济的 3 个维度：数字基础设施、农业数字化、农业数字产业化，分别采用多项指标对上述维度进行量化，并通过熵权法对指标赋予权重，根据贡献率筛选出 18 个指标,进而构建中国城乡收入差距影响因素的多元线性回归模型。另外，本文采用随机森林方法筛选指标，并建立乡村振兴和数字经济对中国城乡收入差距定量预测模型。采用统一的评价指标对两种模型进行综合对比分析，以确定最佳预测模型。最后对结果进行分析并可视化，厘清各影响因素与中国城乡收入差距的相关关系。

其次，考虑到各省份之间发展不均衡，城乡收入差距明显，本文从华东，华中等七个地区分别选取 2022 年农业现代化和乡村振兴示范区作为典型案例，由于指标数量众多，时间跨度长，且市级及其以下城市或地区缺少详细资料及官方数据，本文采用该乡村振兴示范区所在省份数据，进行描述性统计分析，并采用随机森林方法剔除指标，构建城乡收入差距模型，对上述结论进行验证。

最后，以国家的视角，横向分析不同地区城乡收入差距的差异，纵向分析城乡收入差距随时间变化的特征，进而得出相应结论，提出有针对性的政策建议。

问题二的思路流程图如下所示：

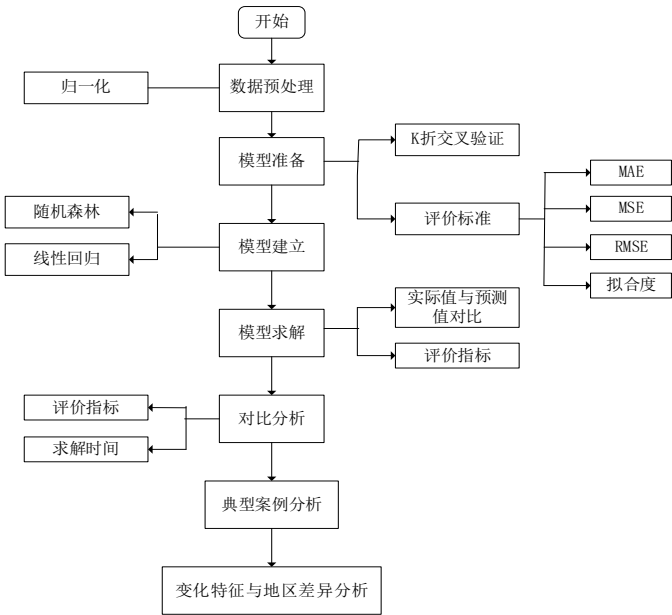


图 5-1 问题二的思路流程图

5.2 模型建立

5.2.1 模型准备——k 折交叉验证

K 折交叉是把原始数据划分为 K 组，其中每次选取一组为测试集，剩余组(K-1 组)为训练集，因此共得到 K 个模型，其该分类器的性能评估指标即为 K 个模型准确率平均值。在实验中通常会使 K 大于等于 2，当数据量较小的时候会取 2，数据量较大时

K 取大于 2, K 要足够大这样训练样本数够多, 一般而言 K 取 10 足够。实验共建立 k 个模型, 并分别计算测试集的识别率, 取平均识别率。

比如对数据集 D 做交叉验证, 将其分成 K 组 (K=10), 依次将其中一个包作为验证集, 其他 K-1 个包作为训练集, 如图 5-2 所示。



10折交叉验证示意图

图 5-2 十折交叉验证法

5.2.2 模型准备——评价指标

1) 平均绝对误差 (MAE)

平均绝对误差 (MAE) 计算所有实际值和预测值的差, 然后求和, 与观察次数或样本量相除。平均绝对误差是一种回归度量。最终结果中, MAE 值越小, 代表该回归模型越好。

$$MAE = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n}$$

2) 均方误差 (MSE)

均方误差计算所有数据和原始数据对应点误差, 并求平方和, 取其均值。均方误差的计算公式如下:

$$MSE = \frac{SSE}{n} = \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{n}$$

3) 均方根误差 (RMSE)

也叫回归系统的拟合标准差。与均方误差相关, 均方根误差 (RMSE) 则取所有数据的实际值和预测值之差, 然后进行平方, 相加, 除以样本数量或观测次数, 最后取结果的平方根。可以看出, RMSE 是对 MSE 进行开方。为了使回归模型被认为是一个好的模型, RMSE 应该尽可能小。计算公式如下:

$$RMSE = \sqrt{MSE}$$

4) 拟合度

R^2 的取值介于 [0 1], 越接近 1, 说明模型选取的自变量对其自身的因变量解释能力越强, 拟合效果越好。计算公式如下:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^n w_i (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n w_i (y_i - \bar{y}_i)^2}$$

其中，SSR 是回归线的误差平方和，SSM 是均线误差的平方和，SSE 为残差平方和。我们将回归线与平均线进行比较。

假设模型的 R^2 得分为 0.6，这意味着可以说模型能够解释 60% 的输出方差。也就是说，60%的城乡收入差距变化可以用输入（指标）来解释，但剩下的 40%是未知的。如果我们的模型有 18 个指标，那么我们的模型能够使用这 18 个输入指标解释 60%的城乡收入差距变化。

5.2.3 模型介绍——随机森林

1)随机森林原理

随机森林（Random Forest，简称 RF）是一种监督式机器学习算法，该方法基于树且高度灵活。顾名思义，该算法利用多棵树（多个相同或不同类型的算法）的力量进行决策，也就是在每个随机训练集上对决策数进行训练，总体而言，整个森林的方差都很小。集成算法大致分成三类，其中，Boosting 指在样本中多次有放回的重复抽样，每抽一次形成一个样本，反复抽取，从而形成多个样本。随机森林的重点是“随机”，包括训练数据的选取，指标的筛选。

随机森林由决策树构成，采用所有决策树的预测值取平均，从而进行整体预测。随机森林对每棵树进行训练，比单个决策树要强大得多，是一个实用的建模方法。该算法在最终输出结果时并不依赖于各个决策树，这样会导致结果方差很低，而是取决于多个决策树的组合。对于回归问题，最终输出是所有决策树输出的平均值。

2)随机森林步骤

Step 1:假设有 N 个样本数据，有放回的随机选择 N 个样本子集，基于这 N 个样本子集训练决策树，作为决策树根节点的样本。

Step 2:假设每个样本有 M 个属性，决策树在分裂时随机选取 m 个属性（ $m \leq M$ ），然后基于这些属性，采用信息增益等策略选择一个分裂属性。

Step 3:直到属性无法分裂，即到达叶子节点。

Step 4:重复上述三个步骤，建立大量的决策树，即构成随机森林。

随机森林算法示意图如图 5-3 所示。

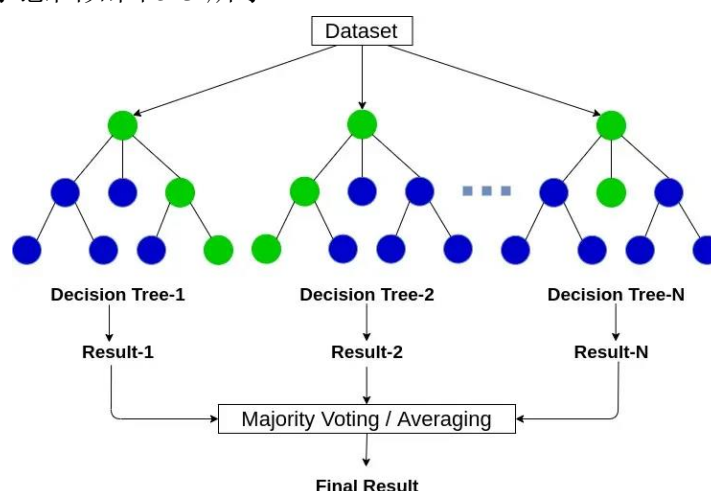


图 5-3 随机森林示意图

3)变量重要性

该方法能够求得变量的相对重要性，从而量化所有变量对模型的贡献。通过变量重要性评价，可以删除那些不重要的变量，并且性能不会受到影响。另外，如果我们使用

不同的机器学习方法（例如支持向量机），则可以将随机森林特征重要性用作一种特征选择方法。

5.2.4 模型介绍——多元线性回归

线性回归方程中变量之间的相关关系多样，其中最常见的是线性。如果随机变量与变量之间是线性的相关关系，那么实验后获得的实际值将分布在某一条直线附近。由此，该回归函数的类型可判定为线性函数。多元线性回归是指因变量和自变量之间是线性关系，变量的数量不再是一个，而是两个或两个以上。线性回归方程式：

$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n + \varepsilon$$

线性回归方程这个统计方法应用十分广泛，其原理是数理统计中的回归分析。线性回归方程可以确定两种或两种以上变量间相互关系。

5.3 模型求解

本文首先基于熵权法和 RF 方法计算各指标对城乡收入差距的贡献率，保留前 85% 的指标，剔除贡献率较低的部分指标，采用五折交叉验证验证模型的鲁棒性并分别对测试集中数据预测，使用 RMSE、MAE、RMAE 和拟合度作为模型评价指标，可视化模型的预测情况，进一步对比分析模型的优劣，确定最合适的回归预测模型。

5.3.1 RF 回归模型

1) RF 降维法

采用随机森林降维法对附件二的指标数据进行处理，根据各指标变量对中国城乡收入差距影响的贡献率（重要性）进行排序，如图 5-4 所示。

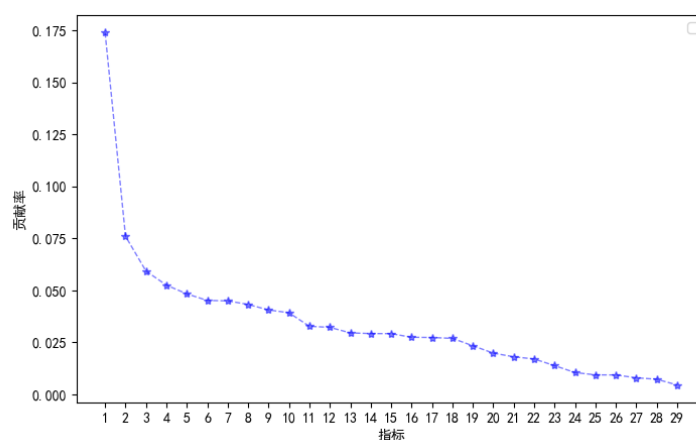


图 5-4 排序后的各指标贡献率

根据图 5-4，可以看出，前 18 个指标对中国城乡收入差距的贡献率较大，具有显著影响，后面特征的贡献率小于 0.024，影响因子较小。

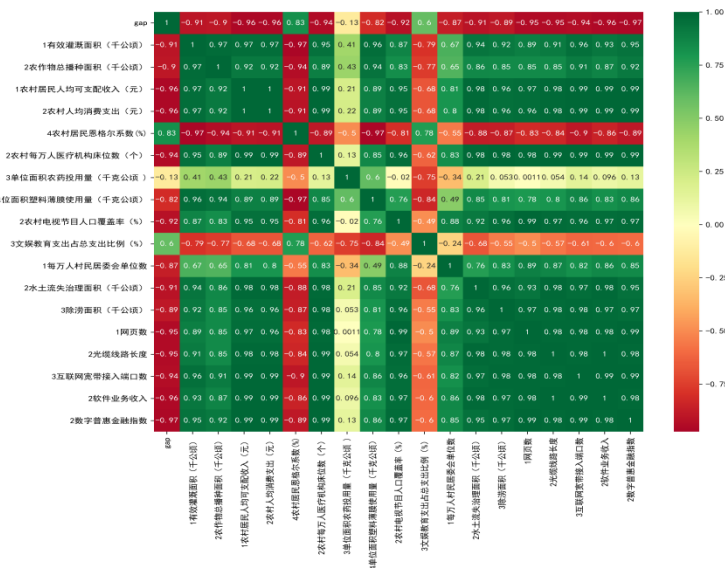


图 5-5 前 18 个指标之间的相关性（换）

根据图 5-5 可以看出 RF 得出的城乡人均收入差距与其他指标之间的相关性，与有效灌溉面积、农作物总播种面积、农村居民人均可支配收入、农村人均消费支出、农村居民恩格尔系数、农村每万人医疗机构床位数、单位面积塑料薄膜使用量、农村电视节目人口覆盖率、每万人村民居委会单位数、水土流失治理面积、除涝面积、网页数、光缆线路长度、互联网宽带接入端口数、软件业务收入、数字普惠金融指数这些指标存在相关关系，在这些指标中有效灌溉面积、农作物总播种面积、农村居民人均可支配收入、农村人均消费支出、农村电视节目人口覆盖率、水土流失治理面积、除涝面积、网页数、光缆线路长度、互联网宽带接入端口数、软件业务收入、数字普惠金融指数与城乡居民人均收入差距存在强相关，其他指标相关性较弱，同样农村居民恩格尔系数呈现正相关关系，这与使用回归方程方法统计的结果并无太大差异。同样对于随机森林算法统计指标结果中的 P 值可知，上述提及的所有指标显著性 P 值都远远小于 0.05，可以认为统计结果显著，可以被用来描述城乡人均收入差距与其他指标间的相关关系。

通过 RF 方法降维，最终筛选的指标及其贡献率如下表 5-1 所示。

表 5-1 对城乡收入差距具有显著影响的 18 个指标

指标名称	贡献率	指标名称	贡献率
农村居民恩格尔系数 (%)	0.17399	农村每万人医疗机构床位数 (个)	0.03928
农作物总播种面积 (千公顷)	0.07604	互联网宽带接入端口数	0.03276
网页数	0.05933	单位面积塑料薄膜使用量 (千克公顷)	0.03226
农村电视节目人口覆盖率 (%)	0.05256	有效灌溉面积 (千公顷)	0.02960
光缆线路长度	0.04836	每万人村民居委会单位数	0.02926
除涝面积 (千公顷)	0.04520	农村人均消费支出 (元)	0.02924
数字普惠金融指数	0.04501	农村居民人均可支配收入 (元)	0.02749
水土流失治理面积 (千公顷)	0.04326	软件业务收入	0.02738
文娱教育支出占总支出比例 (%)	0.04063	单位面积农药投用量 (千克公顷)	0.02695
软件业务收入	0.17399	农村每万人医疗机构床位数 (个)	0.03928
农业气象观测站	0.07604	互联网宽带接入端口数	0.03276
光缆线路长度	0.05933	单位面积塑料薄膜使用量 (千克公顷)	0.03226

2)RF 回归模型求解

画出训练集与测试集数据的中国城乡收入差距实际值和预测值对比图，如图 5-6 所示。可以直观的看到，RF 预测值与实际值差值较小，拟合效果良好，拟合度达到了 65%，提升很大，模型训练时间较短，因此该模型针对本问题较为适用，可作为预测模型。

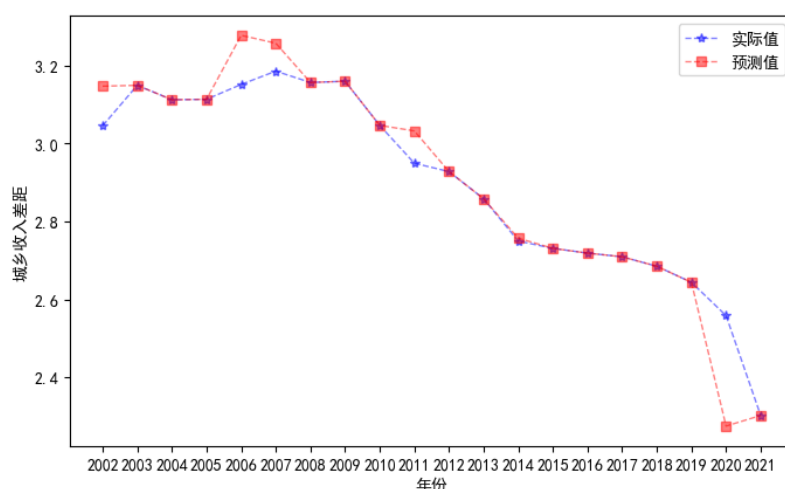


图 5-6 RF 测试数据的中国城乡收入差距实际值和预测值对比图

5.3.2 线性回归模型

1) 熵权法降维

采用熵权法对附件二的指标数据进行处理，根据各指标变量对中国城乡收入差距影响的贡献率（重要性）进行排序，如图 5-7 所示。

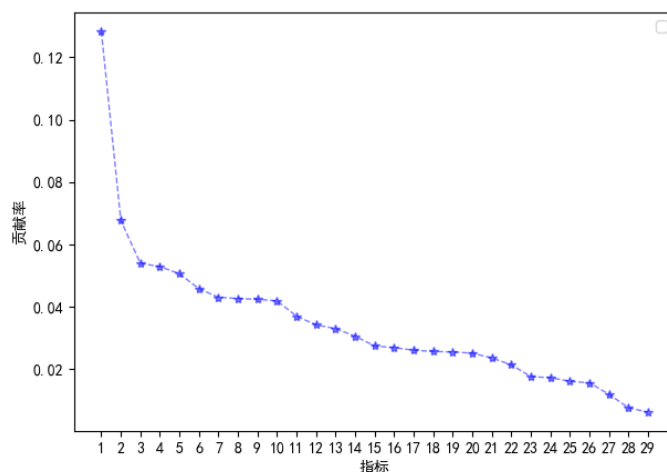


图 5-7 排序后的指标累计贡献率

根据图 5-7，可以看出，前 20 个指标对中国城乡收入差距的贡献率较大，具有显著影响，后面指标的贡献率小于 0.024，影响因子较小。

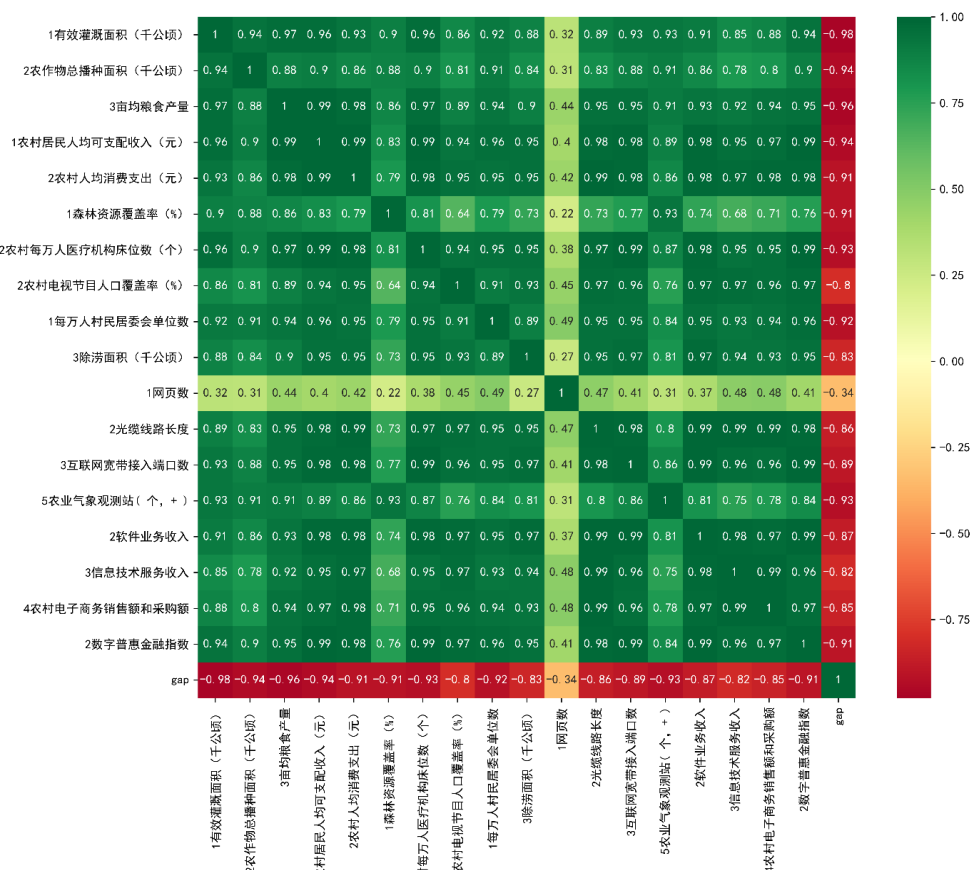


图 5-8 前 20 个指标之间的相关性

根据图 5-8，可以看出通过城乡人均收入差距与其他指标之间的相关信息，与有效灌溉面积、农作物总播种面积、亩均粮食产量、农村居民人均可支配收入、农村人均消费支出、农村居民恩格尔系数、森林资源覆盖率、农村每万人医疗机构床位数、农村电视节目人口覆盖率、每万人村民居委会单位数、除涝面积、网页数、光缆线路长度、互联网宽带接入端口数、软件业务收入、数字普惠金融指数存在相关关系。这些指标中有效灌溉面积、农作物总播种面积、亩均粮食产量、农村居民人均可支配收入、农村人均消费支出、森林资源覆盖率、农村每万人医疗机构床位数、农村电视节目人口覆盖率、每万人村民居委会单位数、除涝面积、网页数、光缆线路长度、互联网宽带接入端口数、软件业务收入、数字普惠金融指数与城乡人均收入差距存在强相关关系，其余指标相关性较弱。这些相关数据源中城乡人均收入差距与农村居民的恩格尔系数呈现正相关，与其他指标则均呈现负相关。从表格第三列可知上述存在强相关的指标中显著性 P 值的统计量远远小于 0.05，因此可以认为上表的统计结果显著，可以被用来描述城乡人均收入差距与其他指标间的相关性。

恩格尔系数表示农村人均食物类产品消费支出与农村人均可支配收入的占比关系。如果恩格尔系数较大，则说明居民的可支配收入不多，并且消费支出多用来购买食物满足生活需要。间接的反应了城乡的人均收入差距。

综合考虑乡村振兴和数字经济两个维度的指标及其下级指标的贡献度和相关性，可以得出结论：中国城乡收入差距主要受到有效灌溉面积、农作物总播种面积、农村居民人均可支配收入、农村人均消费支出、农村居民恩格尔系数、农村电视节目人口覆盖率、除涝面积、网页数、光缆线路长度、互联网宽带接入端口数、软件业务收入、数字普惠金融指数这些指标的影响。其中农村居民恩格尔系数为消极影响，恩格尔系数的提高会

导致城乡收入差距的增大。其他指标都对城乡人均收入差距有积极影响，增加指标值会减少这个差距。

因此，如果需要缩小城乡收入差距，政府可以选择的策略包括：1.增加对农业科技的资金投入，这样能够有效提高灌溉面积，从而提高单位面积的粮食产量；2.增加对自然灾害的处理面积，减少因洪涝灾害导致的农民受益降低或者耕地面积的减少；3.增加互联网宽带接入端口数，推动农产品在互联网的方面的宣传；4.增加软件业务量，通过网上销售的方式提高农民的收入。

通过熵权法降维，最终筛选的指标及其贡献率如下表 5-2 所示。

表 5-2 对城乡收入差距具有显著影响的 18 个指标

指标名称	贡献率	指标名称	贡献率
农村电子商务销售额和采购额	0.128377638	农村每万人医疗机构床位数（个）	0.037035066
农业气象观测站（个，+）	0.067703805	农村居民人均可支配收入（元）	0.034279989
数字普惠金融指数	0.053993024	农村人均消费支出（元）	0.033078506
信息技术服务收入	0.052966155	每万人村民居委会单位数	0.030637141
网页数	0.050647228	有效灌溉面积（千公顷）	0.027489647
光缆线路长度	0.045857564	农村居民恩格尔系数（%）	0.02697446
森林资源覆盖率（%）	0.04310441	农作物总播种面积（千公顷）	0.026139556
互联网宽带接入端口数	0.042704719	亩均粮食产量	0.025758281
软件业务收入	0.042453143	农村电视节目人口覆盖率（%）	0.025586151
除涝面积（千公顷）	0.041962397	文娱教育支出占总支出比例（%）	0.025201945
农村电子商务销售额和采购额	0.128377638	农村每万人医疗机构床位数（个）	0.037035066
农业气象观测站（个，+）	0.067703805	农村居民人均可支配收入（元）	0.034279989
数字普惠金融指数	0.053993024	农村人均消费支出（元）	0.033078506

熵权法与 RF 方法遴选出来的指标有些许差异，这是因为两种方法计算权重的方式不同，但整体上来看选出的指标基本保持一致，共同选出的指标数量超过 70%，可以认为本文选取的指标具有较高的可靠性。

2)线性回归模型求解

基于 2002-2021 年相关数据，采用多元线性回归方法，计算得到各变量的回归系数及截距项，如表 5-3 所示。

表 5-3 各变量回归系数

自变量名称	系数	自变量名称	系数
有效灌溉面积（千公顷）	5.05E-05	文娱教育支出占总支出比例（%）	-8.51E-08
农作物总播种面积（千公顷）	-0.001168052	每万人村民居委会单位数	-4.56E-05
亩均粮食产量	-8.53E-05	除涝面积（千公顷）	1.13E-03
农村居民人均可支配收入（元）	-3.06E-04	网页数	-7.24E-07
农村人均消费支出（元）	3.07E-04	光缆线路长度	1.28E-07
农村居民恩格尔系数（%）	-3.49E-05	互联网宽带接入端口数	6.91E-05
森林资源覆盖率（%）	0.000867837	农业气象观测站（个，+）	-2.40E-05
农村每万人医疗机构床位数（个）	0.000957031	软件业务收入	3.20E-08
单位面积农药投用量（千克公顷）	5.48E-06	信息技术服务收入	-4.71E-08
农村电视节目人口覆盖率（%）	-0.000111575	农村电子商务销售额和采购额	1.71E-06
文娱教育支出占总支出比例（%）	8.65750218		

画出训练集与测试集数据的中国城乡收入差距实际值和预测值对比图，如图 5-9 所示。可以直观的看到，线性回归方法预测值与实际值差值较大，与 RF 方法求解结果相

比拟合效果欠佳，拟合度为 60%，提升空间很大，求解时间较短，因此该模型可作为备选预测模型。

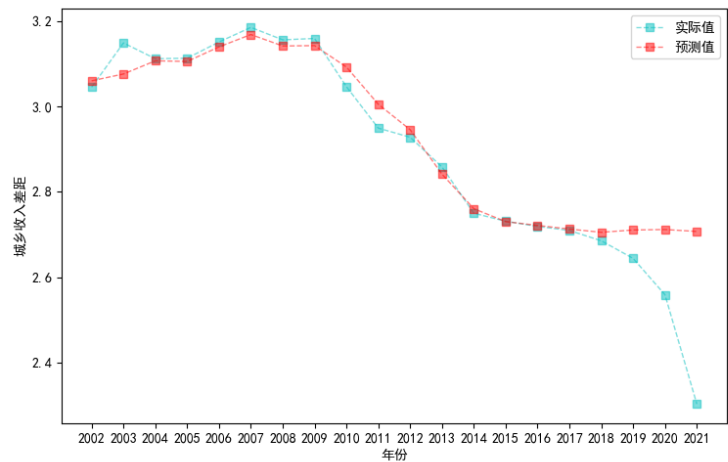


图 5-9 回归测试数据的中国城乡收入差距实际值和预测值对比图

5.3.3 模型比较

根据上述评价指标 RMSE、MAE 以及拟合度对两种模型进行计算，得到结果如表 5-4 所示。根据两种模型的评价指标绘制统计图，从而进行清晰可观的对比分析，对比图如图 5-10 所示。

表 5-4 中国城乡收入差距在两种模型下的评价指标结果

	MAE	MSE	RMSE	拟合度
随机森林	0.126870174	0.033255463	0.182360803	0.619872348
线性回归	0.112127104	0.019763627	0.140583168	0.602343659

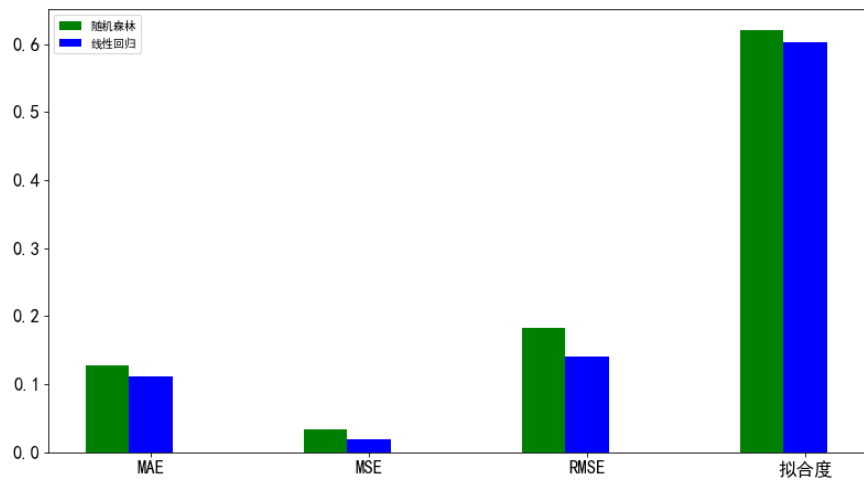


图 5-10 中国城乡收入差距在两种模型下的评价指标对比图

指标 RMSE 和 MAE 值越小，模型预测值越接近实际值，而拟合度 R 方越大，模型拟合效果越好。

综合上图对模型进行对比分析，可以看出，基于线性回归的预测模型在测试集上的 MAE、MSE、RMSE 值比较小，而基于随机森林的预测模型的拟合度较优。在求解时间上，线性回归方法为 10.1s，RF 需 13.4s，两种方法相比，回归的求解时间占优势。

5.4 典型案例分析

本文以 2022 年农业现代化和乡村振兴示范区为选择依据，分别从华北等地区选择一个城市，构建模型并进行验证。具体如下表 5-5 所示。本文以湖北为例进行分析。

表 5-5 典型案例

地区	省、市/县
华东	浙江省常山县
华南	广西壮族自治区合浦县
华中	湖北省云梦县
华北	山西省高平市
东北	吉林省公主岭市
西北	甘肃省永昌县
西南	重庆市秀山县

湖北省指标和城乡收入差距数据情况如表 5-6 所示。

表 5-6 湖北省描述性统计

名称	最小值	最大值	平均值	标准差	中位数
城乡收入差距	2.206	2.73	2.458	0.18	2.407
有效灌溉面积	2006.49	3086.04	2537.784	392.93	2502.3
农作物总播种面积	6900.59	7983.43	7541.394	367.479	7569.69
单位耕地面积化肥施用量	0.34	0.48	0.412	0.046	0.415
农村居民人均可支配收入	2525	18259	8861.55	5218.098	8061
农村人均消费支出	1893	17647	7683.75	4914.594	6711.5
农村每万人医疗机构床位数	16.77	55.2	33.389	13.856	27.9
单位面积农药投用量	1.24	1.94	1.599	0.25	1.54
单位面积塑料薄膜使用量	7130.29	9473.04	8254.993	565.348	8323.08
农村电视节目人口覆盖率	97.5	99.85	98.439	0.744	98.18
每万人村居委会单位数	8.16	11.48	9.435	1.099	9.345
水土流失治理面积	4033.6	6509.13	5072.074	868.264	4796.55
除涝面积	1126.89	1511.48	1288.317	121.462	1232.625
网页数	156806.08	321288.166	212608.669	40251.126	202027.784
光缆线路长度	86098.37	2234455.549	768281.436	693464.116	492231.35
互联网宽带接入端口数	48.46	3308.1	1311.099	1218.352	884.8
数字普惠金融指数	0.455	397.068	136.789	144.991	70.62

沿用 RF 方法对湖北的数据集进行训练并测试，各指标间的相关性见图 5-11。观察各指标对城乡收入差距的影响关系，经检验，湖北省的指标与城乡收入差距正负关系与前一章节中国整体的指标与城乡收入差距正负关系保持一致，此处不再赘述。

画出 2002-2021 湖北城乡收入差距实际值和预测值对比图和评估指标，如图 5-12，表 5-7 所示。可以直观的看到，RF 预测值与实际值差值较小，拟合效果良好，拟合度达到了 85%，模型训练时间较短。

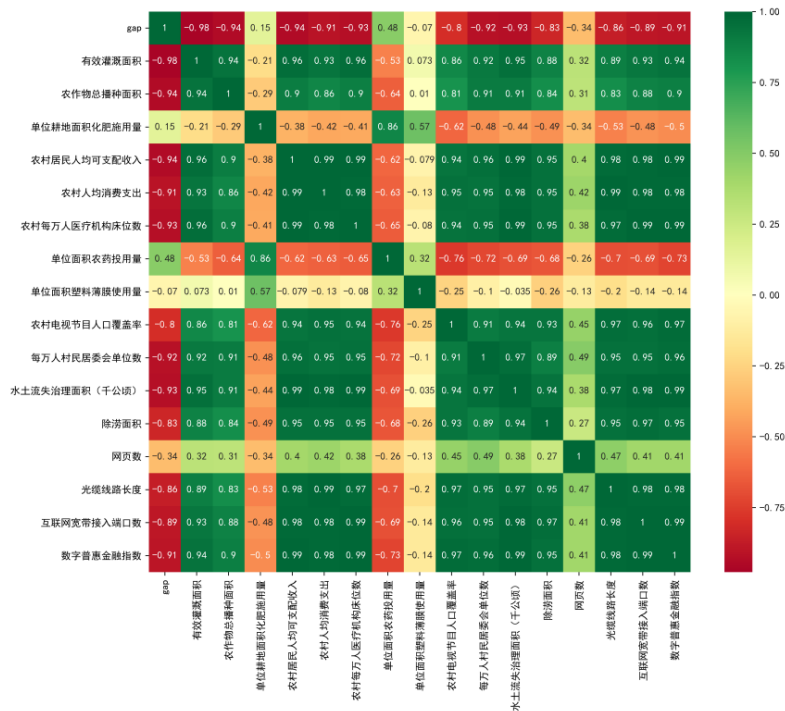


图 5-11 各指标之间相关性

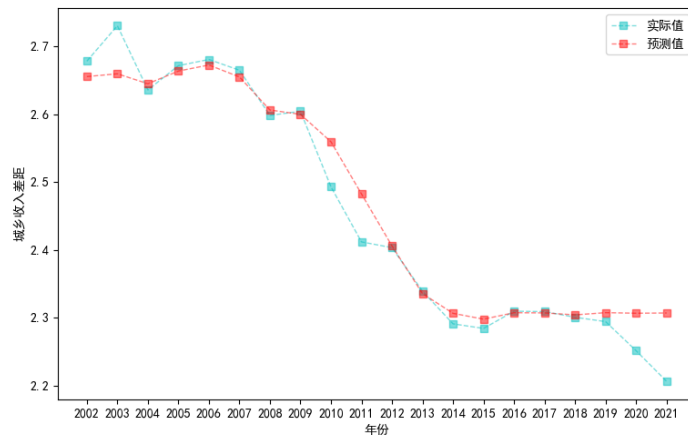


图 5-12 RF 测试数据的湖北城乡收入差距实际值和预测值对比图

表 5-7 湖北省城乡收入差距 RF 模型的评估指标

MAE	MSE	RMSE	拟合度	运行时间(s)
0.051508463	0.003804807	0.061683118	0.878988842	9.468704939

其他地区的城乡收入差距实际值和预测值对比图如下所示，图中蓝色的线表示各个地区的城乡人均收入差距的实际值，橙色的线表示预测值，可以看出图中的预测数据基本与真实值保持一致。根据表 5-8 评估指标统计，可以看出各地区及全国整体的预测值与真实值之间的误差在合理范围内，拟合度均达到 0.6 以上，拟合效果较好，因此可以认为论文建立的预测模型准确有效。完整测试数据见附件二。

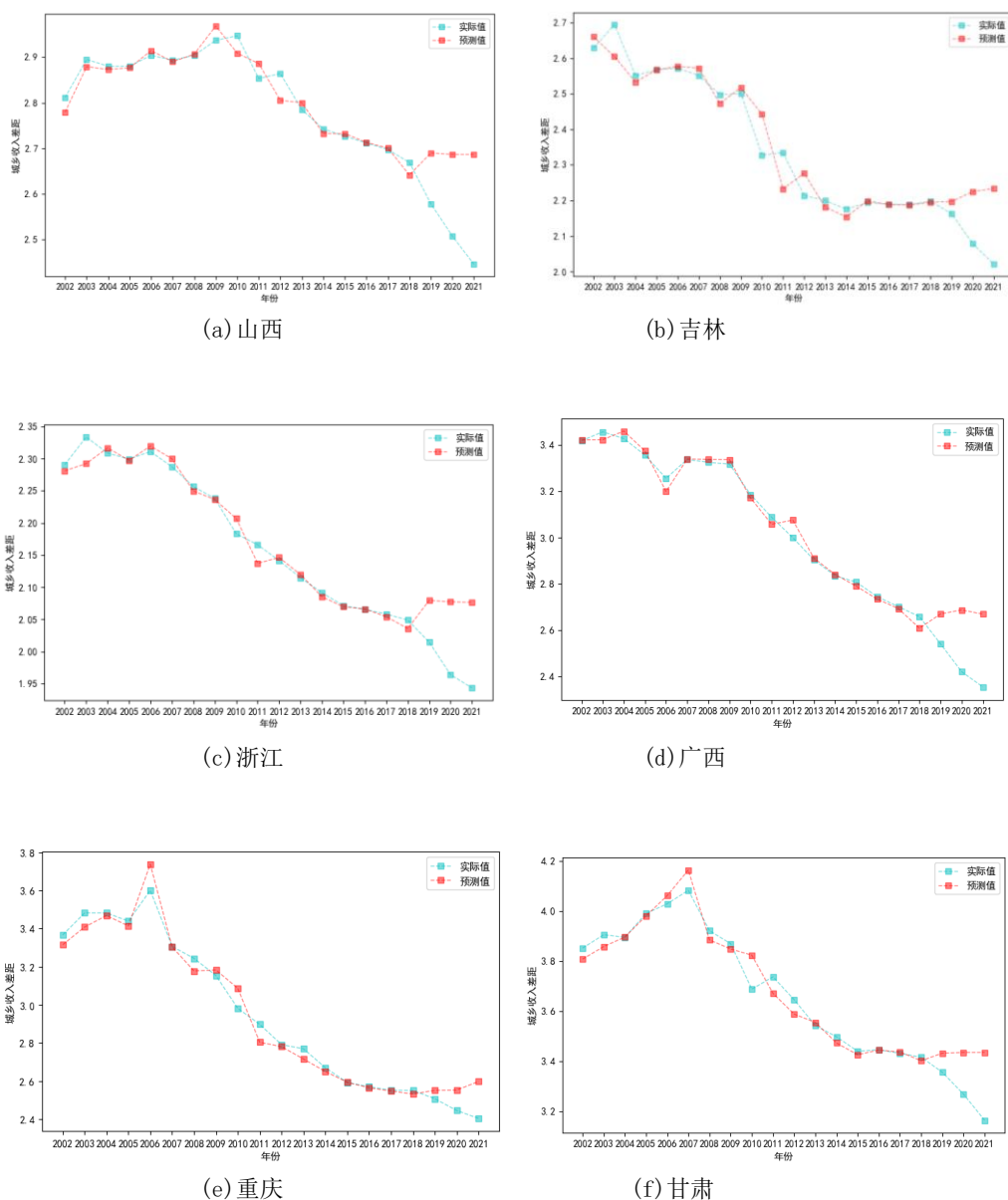


图 5-13 各地区城乡收入差距实际值和预测值对比图

表 5-8 其他省份城乡收入差距 RF 模型的评估指标

省份	MAE	MSE	RMSE	拟合度	运行时间(s)
山西	0.106565208	0.017524629	0.132380622	0.766758415	9.425700426
吉林	0.110301415	0.015539016	0.12465559	0.683802843	9.042675495
浙江	0.053247922	0.004074195	0.063829424	0.785072262	11.80688524
广西	0.139276617	0.032374201	0.179928321	0.810733115	11.23783779
重庆	0.089509325	0.011260096	0.106113601	0.92138584	10.13975596
甘肃	0.075637105	0.009338758	0.09663725	0.81164677	10.67534256

5.5 城乡收入差距分析

5.5.1 城乡收入差距变化分析

本文从纵向的角度分析中国及其各省份的城乡收入差距自 2002 年到 2021 年的发展变化，探讨导致变化的原因，从而为未来继续缩小城乡收入差距提供合理的建议。中国城乡收入差距的变化如图 5-14 所示。由于地区内部收入差距变化较为相似，本文求取其城乡收入差距的平均值作为该地区的城乡收入差距值，七个地区的城乡收入差距变化如图 5-15 所示。

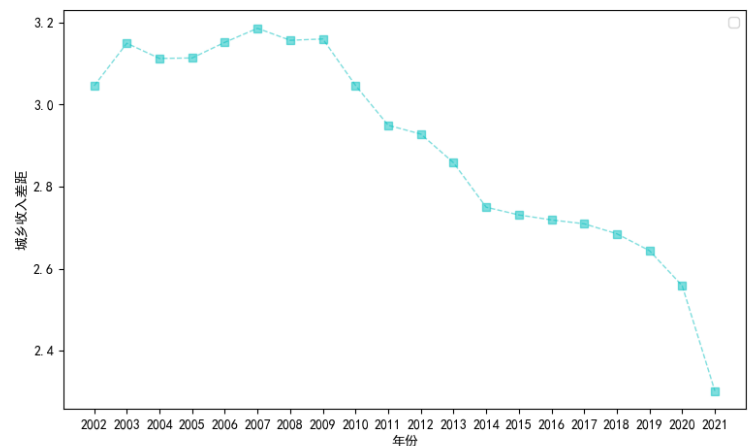


图 5-14 中国城乡人均收入差距变化

图 5-14 中展示了全国整体的城乡人均收入差距，从图中可以看出，近 20 年以来城乡人均收入差距总体上呈稳定下降趋势，部分年份存在一定的波动。

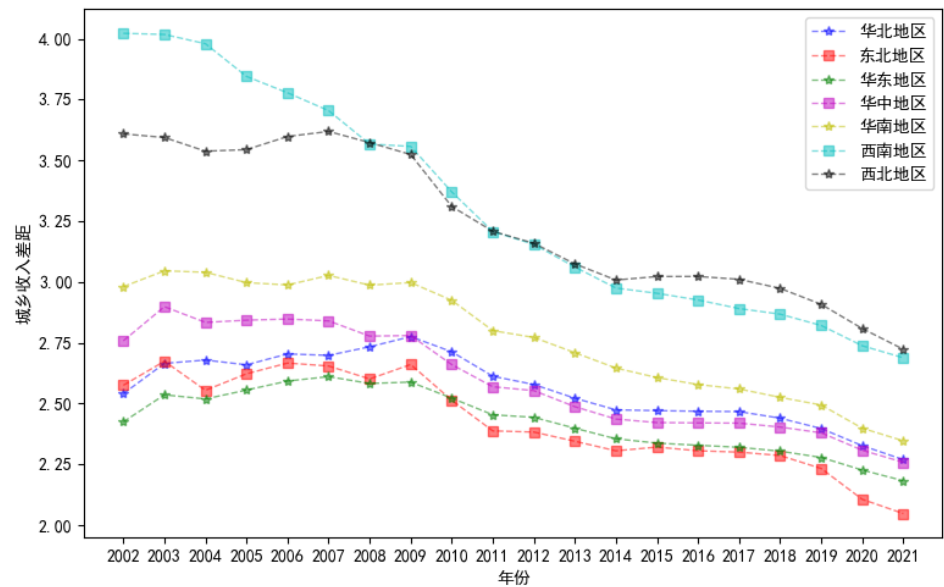


图 5-15 各地区城乡人均收入差距变化

图 5-15 中展示了华东、华中、华南、华北、西南、西北、东北七个地区的城乡收入差距，从图中可以看出，近 20 年以来城乡人均收入差距总体上呈稳定下降趋势，部分

年份存在一定的波动。其中，西北和西南地区城乡收入差距下降趋势明显，逐步赶上其它地区，发展态势良好。

5.5.2 城乡收入差距地区差异分析

本文从空间角度分析各地区的城乡收入差距的差异。由于地区内部收入差距变化较为相似，本文求取其城乡收入差距的平均值作为该地区的城乡收入差距值，七个地区的城乡收入差距如图 5-16 所示。

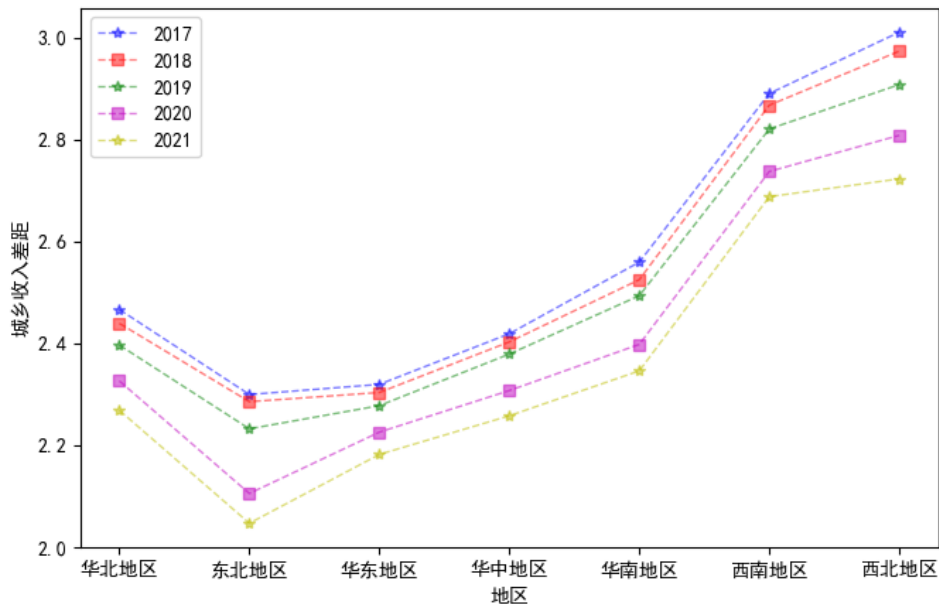


图 5-16 各地区的城乡收入差距差异

根据上图 5-16 分析，得出如下结论：

在 2017-2019 这三年，各地区城乡收入差距相对较小且之间的差距始终保持不变，2020 和 2021 年，东北地区城乡收入差距拉开距离，远远优于其它地区，其次是华东地区，同时结合近 20 年各个地区的城乡收入差距，可以看出东北和华东地区的城乡收入差距始终优于其他地区且呈现加速下降趋势。原因在于这两个地区，地理位置上靠近中国东部沿海地区，科技发达，对于一些能够提高粮食产量或者农民收入的科技创新最先被用到这些地区，同时互联网在这些地区发展迅猛，能够有效推动农产品的宣传和农产品的贸易，因此这些地区的城乡收入差距比其他地区要小。

西南和西北地区城乡收入差距始终居高不下，远高于其他五个地区，因此仍需要国家及当地政府加大宏观调控力度，采取合理措施，加快缩小该地区的城乡收入差距，追赶上其他地区。

华东地区、华南地区、华中地区、华北地区的城乡收入差距较为平衡，且都呈逐年下降的趋势，发展势头良好，应继续保持。

6 问题三分析与模型建立

6.1 问题分析

问题三要求建立数学模型来预测我国未来城乡人均可支配收入差距的变化趋势，并且建立模型考虑在采取行动或者措施情况下（即考虑不同指标敏感度对目标的影响）城乡人均可支配收入差距的变化趋势。通过 arima 模型预测的结果可以看出，在政策、政

府干预条件不变的情况下，城乡收入差距下降缓慢，需要超过 20 年以上的时间才能将该差距缩小到接近 OECD 成员国的基本水平。通过查阅文献，我们了解到城乡收入差距与政府干预、对外开放、城市化水平、数字基础、经济发展水平、就业结构之间存在关系，因此采集关于这六项指标的统计数据进行分析。

我们将利用 arima 差分时间自回归移动平均模型来预测中国的城乡人均可支配收入差距未来的变化趋势，同时通过相关性分析来检测不同指标对于城乡人均可支配收入差距的贡献度是正向还是反向的。通过灵敏度分析来判断不同的指标对于该目标的影响力度，最终得出结论，不同指标在不同力度下对城乡收入差距的影响，并估计在这样的影响下我国多久能达到 OECD 成员国的城乡收入差距的水平。

6.2 模型介绍

ARIMA(p, d, q)模型全称为差分自回归移动平均模型(Autoregressive Integrated Moving Average Model,简记 ARIMA)， p 为自回归项， q 为移动平均项数， d 为时间序列成为平稳时所做的差分次数。所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型。ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，包括移动平均过程（MA）、自回归过程（AR）、自回归移动平均过程（ARMA）以及 ARIMA 过程。

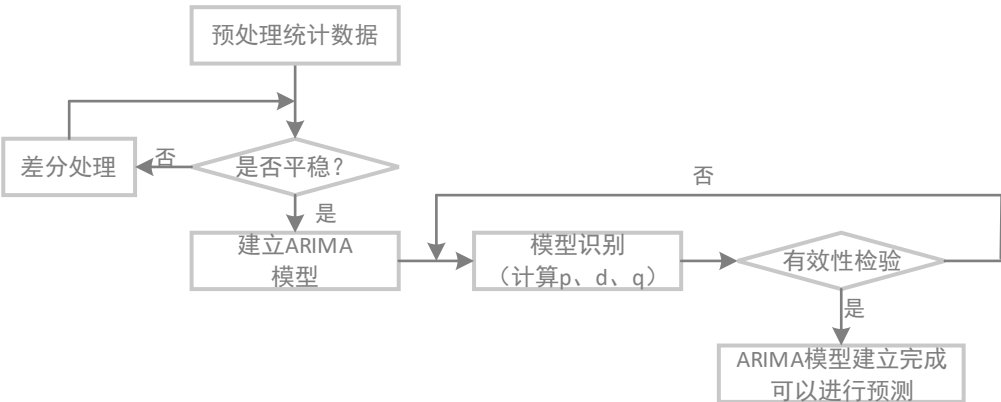


图 6-1 问题三流程图

6.3 模型准备

在建立 ARIMA 模型前，需要检验时间序列数据的平稳性，因此我们首先对所有指标进行 ADF 检验，即单位根检验，检验结果见表 6-1。

表 6-1ADF 检验结果

指标	t-Statistic	Prob.*	是否通过检验
城乡收入差距	0.865835	0.9923	否
政府干预程度	-3.36513	0.0268	是
对外开放程度	-0.53812	0.8617	否
城市化水平	-0.97122	0.7368	否
数字基础设施	-3.32983	0.0307	是
经济发展水平	1.885991	0.9995	否
就业结构	-2.26176	0.1936	否

根据检验结果，城乡收入差距、对外开放程度、城市化水平、经济发展水平、就业结构所对应的指标均未通过平稳性检验。

6.4 模型求解

由于所选指标未能全部通过平稳性检验，本文首先对原始数据进行差分处理，进而对差分后的数据建立 ARIMA 模型，即 ARIMA(p,d,q)模型，其中 d 为对不平稳数据进行差分的次数，p 为自回归项，由 PACF（即偏自相关函数）来确定。q 为移动平均项，由 ACF（即自相关函数）来确定。

ACF 映了同一序列在不同时序的取值之间的相关性，公式如下：

ACF(k) = ρ = Cov(y_t, y_{t-k}) / Var(y_t)

其中分子用于求协方差矩阵，分母用于计算样本方差。求出的 ACF 值为[-1, 1]。但对于一个平稳的 AR 模型，求出其滞后值的自相关系数后，得到的并不是 x(t) 与 x(t-k) 之间单纯的相关关系，其中掺杂了其他随机变量的影响。而偏自相关系数 PACF 是严格这两个变量(现值和滞后值)之间的相关性，剔除了中间随机变量的干扰。

AR公式: y_t = μ + ∑_{i=1}^p γ_i y_{t-i} + ε_t

MA公式: y_t = μ + ε_t + ∑_{i=1}^q θ_i ε_{t-i}

其中ε是白噪声，均值为 0，方差为 1，θ和γ表示自回归系数。

ARIMA 模型本质上由三部分组成：AR（p 阶自回归模型）+I（i 阶差分）+MA（q 阶移动平均模型）。

ARIMA 首先要确定差分阶数 i，以确保数据在 i 阶差分后平稳；接着，再确定是 AR（q=0）、MA（p=0）还是 ARMA（p、q 均不为 0）。

表 6-2ARIMA 模型结果

ARIMA(p,d,q)模型	BIC	R 方	标准误差
ARIMA(1,1,1)	-5.648	0.882	0.114
ARIMA(1,1,2)	-6.183	0.958	0.051
ARIMA(1,1,3)	-5.999	0.96	0.048
ARIMA(2,1,1)	-6.19	0.959	0.047
ARIMA(2,1,2)	-6.006	0.96	0.047
ARIMA(2,1,3)	-5.752	0.958	0.052
ARIMA(3,1,1)	-5.98	0.959	0.046
ARIMA(3,1,2)	-5.778	0.96	0.047
ARIMA(3,1,3)	-5.51	0.96	0.047

文中遍历了 1-3 的所有 p 和 q，BIC 为贝叶斯信息准则，在选取 p 和 q 项系数的时候，优先选择选取其中 BIC 值最小的模型作为 ARIMA 模型的参数，从 R 方和标准误差可以看出，(p=2,q=1)的 R 方足够大标准误差足够小，拟合效果最好，是所有模型中最优的一项。因此论文选择了 ARIMA(2,1,1)模型作为时间序列预测的模型。

如图 6-2，可以看出所有指标一阶和二阶的残差项都在置信区间内部，又如图 6-3，模型的拟合结果位于置信区间内部，因此可以认为拟合效果良好，该模型可以用来执行时间序列预测任务。

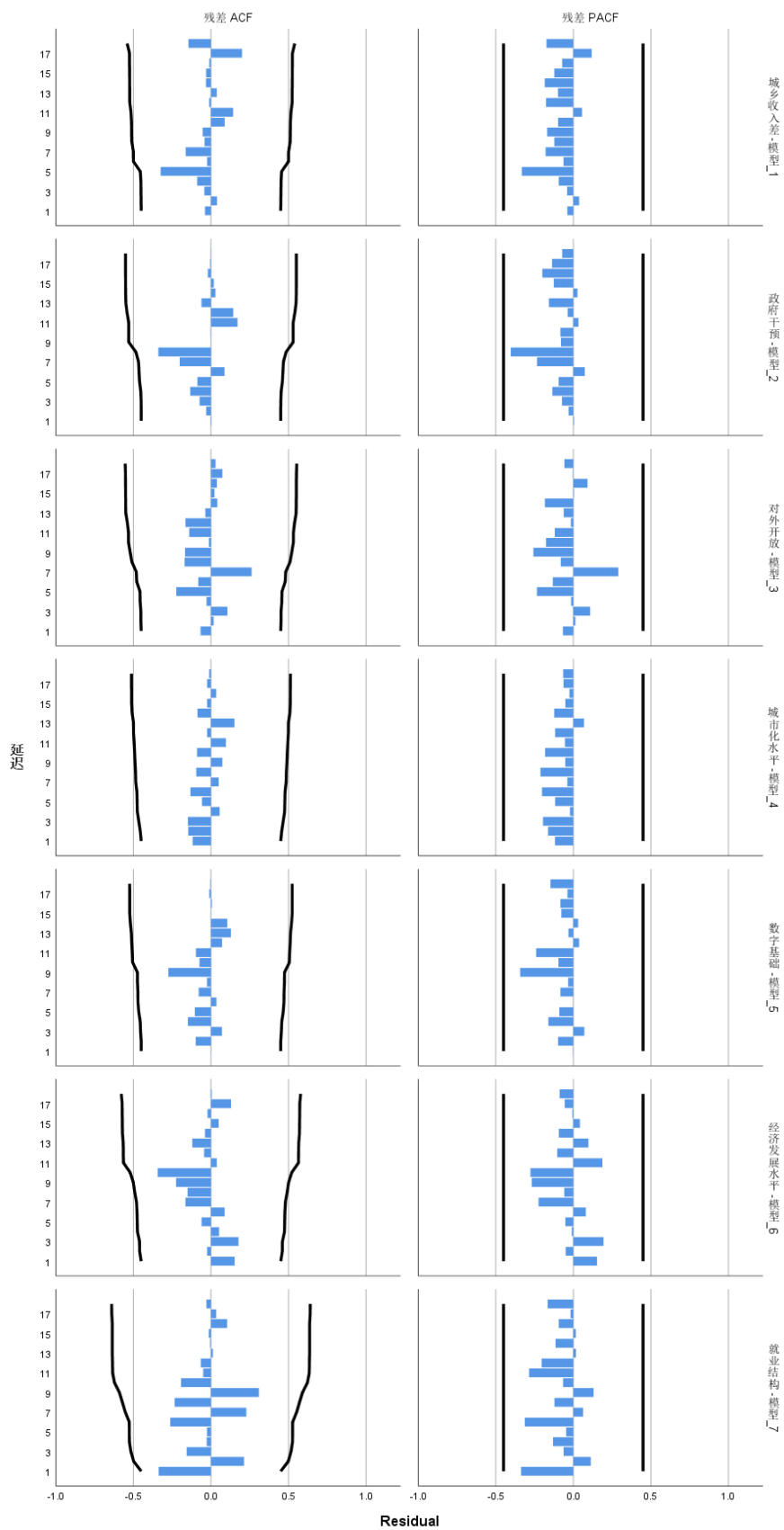


图 6-2 所有指标一阶延迟和二阶延迟的残差项

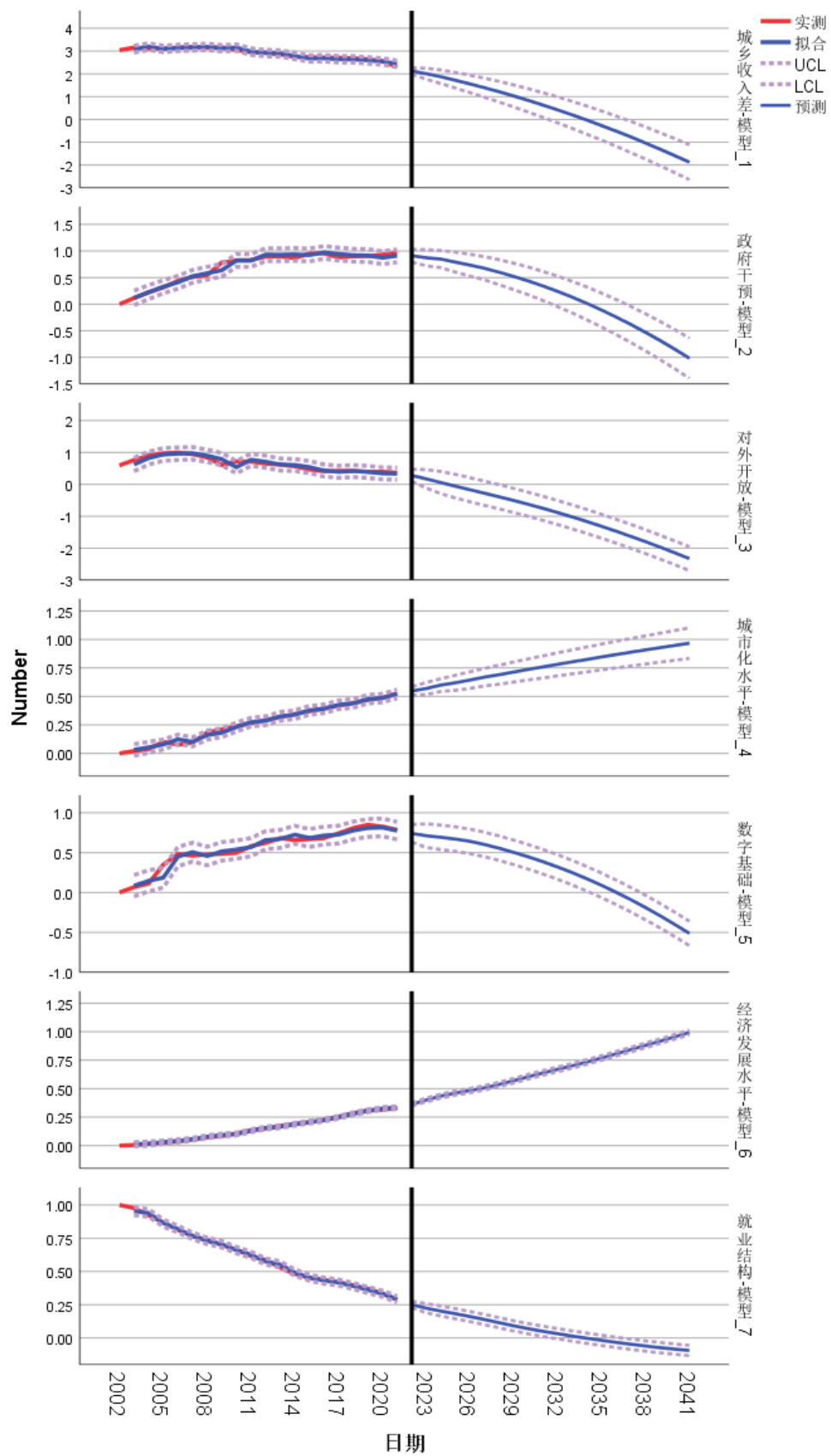


图 6-3 所有指标的拟合、预测情况以及置信区间

6.5 结果分析

表 6-3 是通过 ARIMA 模型预测的时间序列数据，分别对政府干预程度、对外开放程度、城市化水平、数字基础设施、经济发展水平、就业结构进行了未来 20 年的预测，通过数学模型的预测结果可以看出在国家不主动干扰指标值的情况下，根据各项指标自然发展的趋势，我国想要达到 OECD 成员国的基本水平需要花费 23 以上的时间。

表 6-3ARIMA 预测结果

年度标识	城乡收入差距	政府干预程度	对外开放程度	城市化水平	数字基础设施	经济发展水平	就业结构
2002	3.045788849	0	0.595383	0	0	0	1
2003	3.149516729	0.115942	0.762358	0.020547	0.068618	0.005843	0.977261
2004	3.112523951	0.223952	0.903587	0.039038	0.113073	0.015016	0.921678
2005	3.113658754	0.313617	0.972016	0.100678	0.349231	0.024492	0.868621
2006	3.151822568	0.440198	1	0.076022	0.483385	0.03643	0.813037
2007	3.185997227	0.512377	0.949687	0.094514	0.461571	0.055353	0.767559
2008	3.156783357	0.543277	0.844381	0.182864	0.477627	0.073517	0.737241
2009	3.1600092	0.784851	0.607555	0.211629	0.484262	0.083992	0.699343
2010	3.046785714	0.821929	0.71673	0.23834	0.497558	0.107308	0.663972
2011	2.949659183	0.829859	0.712009	0.26505	0.580558	0.134982	0.615968
2012	2.928205984	0.893877	0.647456	0.29135	0.621393	0.15298	0.585649
2013	2.858440668	0.896115	0.618695	0.317444	0.682593	0.172173	0.530066
2014	2.749945337	0.867576	0.564882	0.345182	0.656659	0.189745	0.482062
2015	2.731187991	0.952469	0.467397	0.369838	0.66849	0.205181	0.451743
2016	2.719011034	0.955222	0.410857	0.394494	0.682705	0.224838	0.436584
2017	2.709577184	0.87716	0.424102	0.417095	0.744369	0.254435	0.418898
2018	2.68528136	0.899885	0.422397	0.443805	0.809197	0.284617	0.39616
2019	2.644009816	0.90403	0.395762	0.466406	0.84893	0.30923	0.370894
2020	2.558668893	0.931518	0.395985	0.489008	0.826859	0.314811	0.332996
2021	2.30232679	0.965038	0.366375	0.527841	0.785097	0.330874	0.288782
2022	2.286684	0.979037	0.323607	0.554757	0.78093	0.366587	0.250126
2023	2.205856	0.94808	0.291444	0.577769	0.794837	0.403344	0.222335
2024	2.200436	0.98548	0.270519	0.604274	0.829861	0.43218	0.205912
2025	2.193099	0.97555	0.253824	0.629546	0.871806	0.456362	0.195806
2026	2.182163	0.970285	0.235514	0.654407	0.906523	0.482045	0.185953
2027	2.156896	0.986906	0.214768	0.679063	0.926152	0.511424	0.171046
2028	2.094683	0.984711	0.194047	0.703719	0.930775	0.542948	0.149823
2029	2.013226	0.97741	0.17516	0.727758	0.926493	0.57466	0.124558
2030	1.943844	0.994386	0.157944	0.751798	0.921454	0.606101	0.099798
2031	1.877072	0.986538	0.141416	0.775632	0.921978	0.637905	0.079586
2032	1.835295	0.986728	0.125058	0.79926	0.930343	0.670673	0.065942
2033	1.815448	0.99628	0.109012	0.822478	0.944769	0.704468	0.057858
2034	1.795773	0.990788	0.093576	0.84549	0.961104	0.73905	0.052552
2035	1.776451	0.991746	0.078823	0.868297	0.975123	0.774244	0.046993
2036	1.751266	0.998763	0.064633	0.890898	0.984355	0.810056	0.039414

2037	1.709654	0.993412	0.05088	0.913088	0.988765	0.846573	0.029055
2038	1.667953	0.996053	0.037528	0.935278	0.990255	0.883841	0.018191
2039	1.629312	1	0.024597	0.957058	0.991455	0.921852	0.00859
2040	1.595386	0.996077	0.012095	0.978632	0.994448	0.960576	0.002274
2041	1.57563	0.99886	0	1	1	1	0

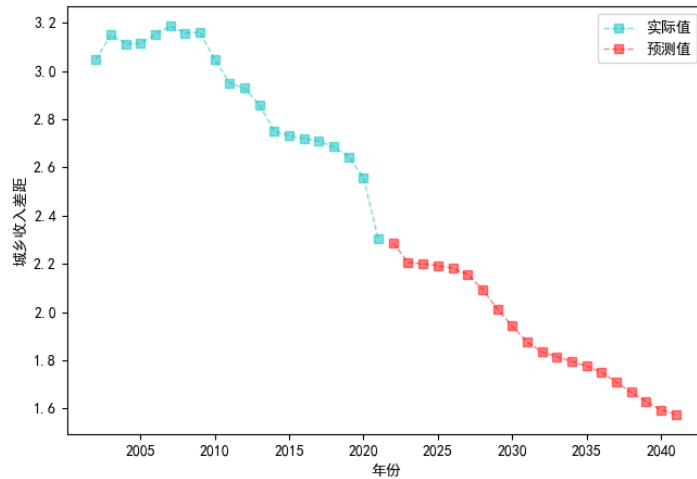


图 6-4ARIMA 预测结果

因此论文使用相关性分析的方法计算城乡人均可支配收入差距与各项指标之间的相关性，并通过熵权法计算出所有指标的影响力，该影响力与相关性对应，表示在指标发生变动的情况下对城乡人均可支配收入差距的影响，其中相关性为正的情况下，为积极影响，相关性为负的情况下为消极影响。

表 6-4 影响力结果

指标名称	影响力	城乡人均可支配收入差距	P值
政府干预	0.0430	-0.715	0
对外开放	0.1843	0.961	0
城市化水平	0.1192	-0.982	0
数字基础	0.0528	-0.872	0
经济发展水平	0.1847	-0.968	0
就业结构	0.2337	0.962	0

从表 6-4 可以看出，城乡人均可支配收入差距与政府干预、对外开放、城市化水平、数字基础、经济发展水平之间存在相关性，与对外开放、城市化水平、经济发展水平和就业结构存在强相关性，对于其他指标则相关性较弱，并且对外开放和就业结构两个指标与城乡人均可支配收入差距之间呈正相关关系，其他指标则为负相关关系。

因此建立城乡人均可支配收入差距的影响力模型：

$$F = -0.0430 * x_1 + 0.1843 * x_2 - 0.1192 * 3x_3 - 0.0528 * x_4 - 0.1847 * x_5 + 1.2337 * x_6$$

其中， $x_1\sim x_6$ 分别代表政府干预程度、对外开放程度、城市化水平、数字基础设施、经济发展水平、就业结构这些指标，前面的系数代表各自的贡献度。F 代表在不同指标变动情况下城乡人均可支配收入差距收到的影响。通过灵敏度分析发现，在其他条件不变的情况下，政府干预力度每增加 1，将导致城乡人均可支配收入差距下降 0.0430；同样，对外开放力度每增加 1，将导致城乡人均可支配收入差距增长 0.1843；城市化水平每增加 1，城乡人均可支配收入差距减少 0.1192；数字基础水平每增加 1，城乡人均可支配收入差距下降 0.0528,经济发展水平每增加 1,城乡人均可支配收入差距减少 0.1847；就业结构每增加 1，城乡人均可支配收入差距则增加 0.2337。

由于城市化水平、经济发展水平和就业结构和这三个指标国家难以出台政策直接影响或控制，因此这里只考虑政府干预、对外开放、数字基础这三项指标的调节力度，以最近一年的城乡人均可支配收入差距作为初值,分别估计在上述 3 个指标每年增长 10%，20%，30%，40%，50%的不同条件下，未来城乡人均可支配收入差距的变化趋势。

表 6-5 调控效果

年份	调控力度 0.1	调控力度 0.2	调控力度 0.3	调控力度 0.4	调控力度 0.5
2022 年	2.2926	2.2646	2.2366	2.2086	2.1806
2023 年	2.2655	2.2121	2.1605	2.1107	2.0627
2024 年	2.2390	2.1609	2.0849	2.0091	1.9318
2025 年	2.2128	2.1090	2.0028	1.8880	1.7587
2026 年	2.1867	2.0541	1.9074	1.7312	1.5104
2027 年	2.1604	1.9944	1.7914	1.5194	1.1438
2028 年	2.1336	1.9275	1.6462	1.2274	
2029 年	2.1061	1.8511	1.4613		
2030 年	2.0776	1.7626			
2031 年	2.0479	1.6587			
2032 年	2.0166	1.5362			
2033 年	1.9835	1.3906			
2034 年	1.9482				
2035 年	1.9105				
2036 年	1.8699				
2037 年	1.8261				
2038 年	1.7786				
2039 年	1.7271				
2040 年	1.6711				
2041 年	1.6100				
2042 年	1.5433				
2043 年	1.4704				

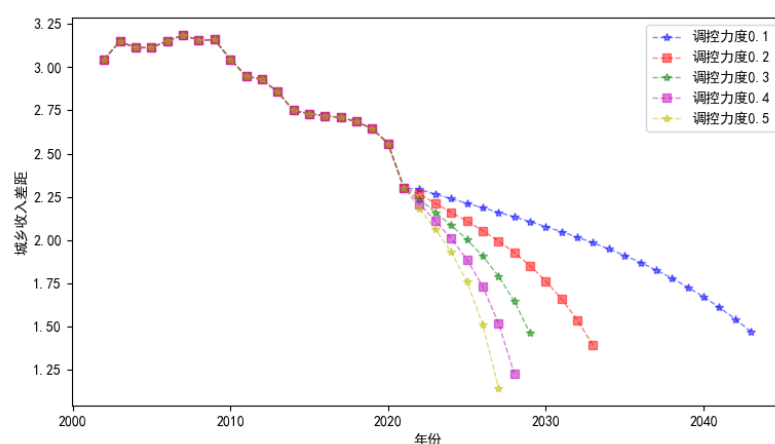


图 6-5 调控效果

根据表 6-5 可以看出，在保持其他条件不变的情况下，政府干预、数字基础和对外开放程度的调控力度在不同程度上具有不同的效果，在保持其他条件不变的情况下，调控力度为 10% 时中国城乡人均可支配收入差距在 2043 年可以达到 OECD 成员国的基本水平，需要 22 年时间；调控力度为 20% 时中国城乡人均可支配收入差距在 2033 年可以达到 OECD 成员国的基本水平，需要 12 年时间；调控力度为 30% 时中国城乡人均可支配收入差距在 2033 年可以达到 OECD 成员国的基本水平，需要 8 年时间；调控力度为 40% 时中国城乡人均可支配收入差距在 2033 年可以达到 OECD 成员国的基本水平，需要 7 年时间；调控力度为 50% 时中国城乡人均可支配收入差距在 2033 年可以达到 OECD 成员国的基本水平，需要 6 年时间。

7 模型评价与推广

7.1 模型评价

优点：

1.论文中使用的模型对现实问题进行简化，充分体现“用简单的方法解决复杂的问题”这一数学建模核心思想。模型简单，容易被大多数人理解，易于应用到生活中的其他场景。

2.模型使用不同的方法对数据进行分析，在模型准备阶段做了数据校验、数据预处理，模型结果输出后做了可靠性检验，从检验结果可以看出所建立的模型性能良好，可以使用。

3.论文中建立的耦合度模型能够与实际紧密联系，结合实际情况对问题进行求解，使得模型的现实作用高于一般的其他模型。

不足：

1.论文中使用的指标数较少，没有考虑到与问题相关的所有指标项，在分析问题时可能存在一些误差。

2.收集到的最近 20 年的年数据样本量较少，对于未来长期的变化趋势预测不够精确。

3.在假设中我们做出了“以理性为前提”的政策约束，但现实情况下，为了拉动中西部地区的经济发展，颁布的经济政策总会给与这些地区一些政策支持。

7.2 模型的推广

本文在考虑收集指标数据时，主要考虑与乡村振兴和数字经济相关的指标，通过表格可以看出收集指标的内容，因此如果需要对其他的现实情况进行分析，只需要变换数据分析和评价使用的指标，因此论文提出的模型可以被用于其他领域，像销量预测、经济发展情况、水资源污染、火势蔓延等多个场景。

参考文献

- [1] 刘蕾,王轶.数字化经营何以促进农民增收?——基于全国返乡创业企业的调查数据[J].中国流通经济,2022,36(01):9-19.DOI:10.14089/j.cnki.cn11-3664/f.2022.01.002.F.
Galton .Personal indentification and description[J].Nature,1888: 173-177.
- [2] 徐华伟.基于时间序列分析对河南省城乡收入差距的预测分析[J].科技信息,2012(30):65-66.张良均. Python 数据分析与挖掘实战[M]. 机械工业出版社, 2016.
- [3] 张蕴萍,栾菁.数字经济赋能乡村振兴:理论机制、制约因素与推进路径[J].改革,2022(05):79-89.《运筹学》教材编写组. 运筹学.第4版[M]. 清华大学出版社, 2012.
- [4] 陈一明.数字经济与乡村产业融合发展的机制创新[J].农业经济问题,2021(12):81-91.
- [5] 张旺,白永秀.数字经济与乡村振兴耦合的理论构建、实证分析及优化路径[J].中国软科学,2022(01):132-146.
- [6] 杨玉敬.数字经济与乡村振兴耦合协调发展水平研究[J].技术经济与管理研究,2022(07):14-19.
- [7] 卢云峰,陈红宇.乡村文化振兴与共同体重建:基于浙江省诸暨市的案例分析[J].清华大学学报(哲学社会科学版),2022,37(03):205-214+220.DOI:10.13613/j.cnki.qhdz.003134.
- [8] 张斌,曾文清,山岚.乡村振兴背景下甘肃省乡村人口—经济—社会耦合协调与空间分异研究[J].云南农业大学学报(社会科学),2022,16(04):1-8.
- [9] 李蕊.乡村振兴地方立法的逻辑进路[J].地方立法研究,2022,7(01):1-11.
- [10] 徐希瑞,王华书.中国数字普惠金融与乡村振兴耦合关系及空间效应分析[J/OL].新疆农垦经济:1-13[2022-08-27].<http://kns.cnki.net/kcms/detail/65.1048.F.20220822.1136.004.html>.
- [11] 龚承刚,王梦,谢航.基于 ARIMA 模型的湖北省城乡居民收入差距的预测[J].统计与决策,2014(15):4.
- [12] 李豫新,李枝轩.乡村振兴背景下数字经济发展与城乡收入差距[J].金融与经济,2022,(6): 60-67,96.
- [13] 何雷华,王凤,王长明.数字经济如何驱动中国乡村振兴?[J].经济问题探索,2022,(4): 1-18.
- [14] 杨江华,刘亚辉.数字乡村建设激活乡村产业振兴的路径机制研究[J].福建论坛(人文社会科学版),2022,(2): 190-200.

附录

论文主要使用的软件为 matlab、pycharm、spss、excel 软件，使用的变成语言是 python 语言，下面是 python 部分代码，详细代码见压缩文件 Python 代码文件夹。

Python 部分代码

问题一代码：#####

#####数据预处理#####

#####熵权法#####

```
import pandas as pd
import numpy as np
import math
from numpy import array
import matplotlib.pyplot as plt
# 加载数据集
df = pd.read_excel('zhongguo.xlsx')
# 定义熵值法函数
def cal_weight(x):
    # 数据的归一化处理
    x = x.apply(lambda x: ((x - np.min(x)) / (np.max(x) - np.min(x))))
    # 求 k
    rows = x.index.size # 行
    cols = x.columns.size # 列
    k = 1.0 / math.log(rows)
    lnf = [[None] * cols for i in range(rows)]
    # 信息熵
    x = array(x)
    lnf = [[None] * cols for i in range(rows)]
    lnf = array(lnf)
    for i in range(0, rows):
        for j in range(0, cols):
            if x[i][j] == 0:
                lnfij = 0.0
            else:
                p = x[i][j] / x.sum(axis=0)[j]
                lnfij = math.log(p) * p * (-k)
            lnf[i][j] = lnfij
    lnf = pd.DataFrame(lnf)
    E = lnf
    # 计算冗余度
    d = 1 - E.sum(axis=0)
```

```

# 计算各指标的权重
w = [[None] * 1 for i in range(cols)]
for j in range(0, cols):
    wj = d[j] / sum(d)
    w[j] = wj
w = pd.DataFrame(w)
return w

if __name__ == '__main__':
    #调用 cal_weight
    w = cal_weight(df)
    w.index = df.columns
    w.columns = ['weight']
    w.to_excel('权重.xlsx')
#####耦合协调度模型#####
import pandas as pd
import numpy as np
import math
from numpy import array
import matplotlib.pyplot as plt
#计算综合评价指数
def Comprehensive_evaluation_index(w, x):
    score = np.zeros(len(x[:,0]))
    for i in range(0,len(x[:,0])):
        for j in range(0,len(w)):
            score[i] = score[i] + w[j] * x[i][j];
        #print(score[i])
    return score
if __name__ == '__main__':
    #加载数据集
    df = pd.read_excel('18hunan.xlsx')
    #数据的归一化处理
    df = np.array(df)
    for j in range(0, len(df[0,:])):
        if np.max(df[:,j]) - np.min(df[:,j]) == 0:
            for i in range(0, len(df[:,0])):
                df[i][j] = 0
        else:
            for i in range(0, len(df[:,0])):
                df[i][j] = (df[i][j] - np.min(df[:,j])) / (np.max(df[:,j]) - np.min(df[:,j]))
    #加载由熵权法计算得的各指标的权重

```

```

w = pd.read_excel('权重.xlsx')
#为了让图形显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.figure(figsize=(10, 5.5))
x_data = w.index
plt.bar(x_data, w['weight'], color='b')
plt.xlabel('指标', labelpad=19)
plt.ylabel('贡献率', labelpad=10)
plt.title('各个指标的贡献率', fontsize=22)
plt.savefig('各个指标的贡献率.png', dpi=500, bbox_inches='tight')
#计算综合评价指数
w = np.array(w)
df = np.array(df)
s1 = Comprehensive_evaluation_index(w[0:16,1], df[:, 0:16])
print("乡村振兴的综合评价指数")
print(s1)
data_output = np.zeros((20,0))
data_output = pd.DataFrame(data_output)
data_output['乡村振兴的综合评价指数'] = s1
s2 = Comprehensive_evaluation_index(w[16:26,1], df[:, 16:26])
print("数字经济的综合评价指数")
print(s2)
data_output['数字经济的综合评价指数'] = s2
#计算耦合度
C = s1
for i in range(0,len(s1)):
    C[i] = np.sqrt((s1[i] * s2[i]) / (((s1[i] + s2[i])/2)**2))
print("耦合度")
print(C)
data_output['耦合度'] = C
#计算耦合协调度
T = s1
D = s1
a = 0.5
b = 0.5
for i in range(0,len(s1)):
    T[i] = a * s1[i] + b * s2[i]
    D[i] = np.sqrt(C[i] * T[i])
print("耦合协调度")
data_output['耦合协调度'] = D

```



```

print(D)
#将数据输出
data_output = pd.DataFrame(data_output)
writer2 = pd.ExcelWriter('输出结果' + '.xlsx')
data_output.to_excel(writer2, 'page_1')
writer2.save()

```

问题二代码：#####

#####随机森林#####

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
#加载自变量和因变量数据集
dataset = pd.read_excel(r'zhongguo 剔除后.xlsx')
# 准备训练数据，自变量，因变量
X = dataset.iloc[:, 1:19].values
y = dataset.iloc[:, 0].values
# 将数据分为训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
# 训练随机森林解决问题
regressor = RandomForestRegressor(n_estimators=200, random_state=0)
regressor.fit(X_train, y_train)
y_train_pred = regressor.predict(X_train)
y_pred = regressor.predict(X_test)
p = pd.DataFrame(y_train_pred)
p.to_excel('训练集测试结果.xlsx')
p = pd.DataFrame(y_pred)
p.to_excel('测试集测试结果.xlsx')
# 评估指标
from sklearn import metrics
from sklearn.metrics import r2_score
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test,
y_pred)))
print('R^2 Error:', metrics.r2_score(y_test, y_pred))
#各指标的相对重要性
importances = list(regressor.feature_importances_)
a = pd.DataFrame(importances)
a.to_excel('各指标的相对重要性.xlsx')

```

```

# 相关性分析
import seaborn as sns
q = dataset.corr()
q = pd.DataFrame(q)
q.to_excel('相关性分析.xlsx')
#画图，特征相关性图
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
plt.figure(figsize = (17,14),dpi = 80)
fig, ax = plt.subplots(figsize = (12, 10))
p1 = sns.heatmap(q,
                  xticklabels = q.columns,
                  yticklabels = q.columns, cmap = 'RdYlGn',center = 0,annot = True)
s1 = p1.get_figure()
s1.savefig('特征相关性图.png', dpi = 500, bbox_inches = 'tight')

```

#####多元线性回归#####

```

import pandas as pd
import seaborn as sns
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import pickle
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus']=False
#加载数据集
adv_data = pd.read_excel("17hubei.xlsx")
#相关性分析
import seaborn as sns
q = adv_data.corr()
q = pd.DataFrame(q)
q.to_excel('相关性分析.xlsx')
#为了显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus']=False
#画图
plt.figure(figsize=(17,14),dpi=80)
fig, ax = plt.subplots(figsize=(12, 10))
p1 = sns.heatmap(q,
                  xticklabels=q.columns,
                  yticklabels=q.columns, cmap='RdYlGn',center=0,annot=True)
s1 = p1.get_figure()
s1.savefig('各指标间的相关性分析.png',dpi=500,bbox_inches='tight')

```

```

#数据拆分
X_train, X_test, Y_train, Y_test = train_test_split(adv_data.iloc[:, 0:18], adv_data.gap,
train_size=0.7, random_state=2)
print("X: 原始数据特征:", adv_data.iloc[:, 0:18].shape,
      ",训练数据特征:", X_train.shape,
      ",测试数据特征:", X_test.shape)
print("Y: 原始数据标签:", adv_data.gap.shape,
      ",训练数据标签:", Y_train.shape,
      ",测试数据标签:", Y_test.shape)

#数据建模
model = LinearRegression()
model.fit(X_train, Y_train)
a = model.intercept_ # 截距
b = model.coef_ # 回归系数
print("最佳拟合线:截距=", a, ",回归系数=", b)
#b.to_excel('模型系数.xlsx')
for i in range(len(X_train.keys())):
    print(str(X_train.keys()[i])+'!'+str(b[i]))
print('常数:'+str(a))

#模型预测训练集
Y_train_pred = model.predict(X_train)
df_empty = pd.DataFrame(Y_train)
df_empty['Y_train_pred'] = Y_train_pred
Y_train_pred = pd.DataFrame(Y_train_pred)
Y_train_pred.to_excel('测试结果.xlsx')

#显示图像，训练集预测结果
plt.figure()
plt.plot(range(len(df_empty['Y_train_pred'])), Y_train, 'b', label="test data")
plt.plot(range(len(df_empty['Y_train_pred'])), df_empty['Y_train_pred'], 'r', label="pred
data")
plt.legend(loc=2)
plt.title("训练集预测结果")
plt.savefig("训练集预测结果.jpg")

#模型预测测试集
Y_pred = model.predict(X_test)
df_empty = pd.DataFrame(Y_test)
df_empty['Y_pred'] = Y_pred
print(df_empty)
Y_pred = pd.DataFrame(Y_pred)
Y_pred.to_excel('测试结果 1.xlsx')

```

```

# 显示图像，测试集预测结果
plt.figure()
plt.plot(range(len(Y_test)), Y_test, 'b', label="test_y")
plt.plot(range(len(Y_test)), Y_pred, 'r', label="pred_y")
plt.legend(loc=2)
plt.title("测试集预测结果")
plt.savefig("测试集预测结果.jpg")
# 评估指标
from sklearn import metrics
from sklearn.metrics import r2_score
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_test, Y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(Y_test, Y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(Y_test, Y_pred)))
print('R^2 Error:', metrics.r2_score(Y_test, Y_pred))

```

问题三代码：#####
#####随机森林#####