Использование скрытых марковских моделей при анализе заболеваемости COVID-19

Введение

Прогнозирование заболеваемости различными вирусными инфекциями, наряду с моделированием указанных процессов, является важной задачей, особенно в условиях пандемии и эпидемий. Обоснованный прогноз позволяет оптимально использовать наличные ресурсы и спланировать ограничительные меры таким образом, чтобы минимизировать негативные последствия для экономики и здоровья населения региона.

Объектом исследования является временной ряд, отражающий заболеваемость COVID-19 в регионе, содержащий даты, количество заболевших (накопленным итогом), и посуточная заболеваемость.

Скрытые марковские модели

Марковский процесс, также известный как процесс Маркова, является стохастическим процессом, в котором будущее состояние системы зависит только от текущего состояния и не зависит от предшествующей истории. Это свойство называется "свойство Маркова" или "свойство отсутствия памяти". Важными концепциями, связанными с марковскими процессами, являются матрица переходных вероятностей, описывающая вероятность перехода из одного состояния в другое.

При анализе эпидемиологических процессов такими состояниями могут быть состояние «волны» заболеваний и условного «плато», либо номер волны.

Скрытые марковские модели (hidden Markov model – HMM) позволяют оценить состояние системы, скрытое от наблюдателя, по другой, уже наблюдаемой переменной. Результатом анализа являются матрица переходов и, для временных рядов – периоды, в которых система находится в одном из состояний. Число состояний должно быть задано априорно. Обучение HMM включает в себя оценку параметров модели на основе наблюдаемых данных. Существуют алгоритмы, такие как алгоритм Витерби, которые позволяют получить наилучшие оценки параметров HMM.

Исходные данные

Для анализа используются данные \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ на \_\_\_\_\_ . Для их загрузки и первичной подготовки используется скрипт на языке Python ([www.python.org](http://www.python.org)) (см. приложение 1). Результаты импорта представлены на рисунке 1. Здесь и далее в качестве времени (Date) используется время от начала пандемии в днях, Infections – количество заболевших в регионе накопленным итогом, Inf\_day – количество заболевших COVID-19 в регионе за сутки.

В настоящей работе не исследуется влияние способа сбора данных и другие особенности статистического учета на результат.

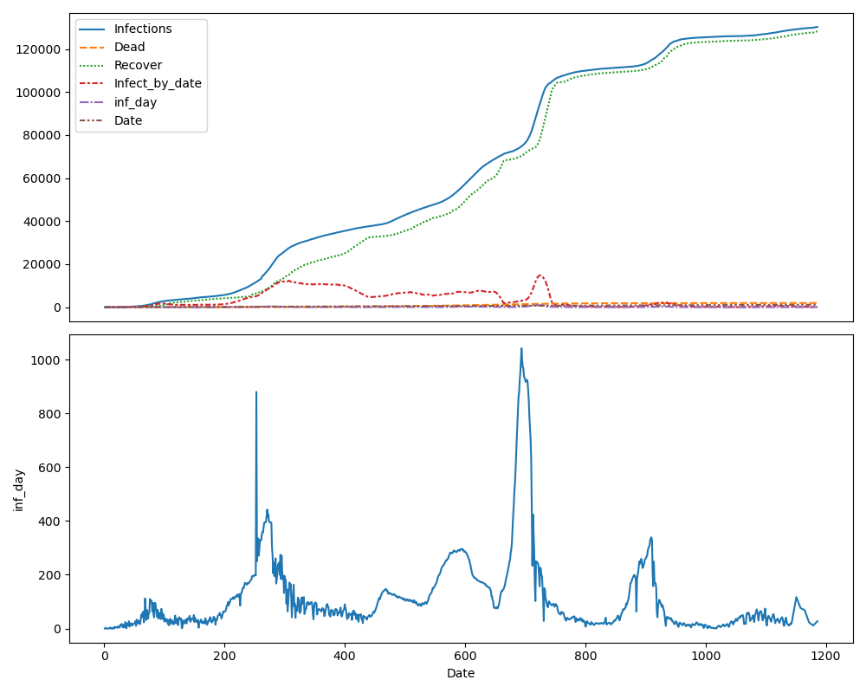


Рисунок Основные сведения о заболеваемости COVID-19 в регионе

Первичный анализ

Наблюдается статистически значимая линейная регрессия между общим количеством заболевших и временем , таким образом среднее количество заболевших в день за период пандемии составляет . Коэффициент детерминации модели при этом составляет .

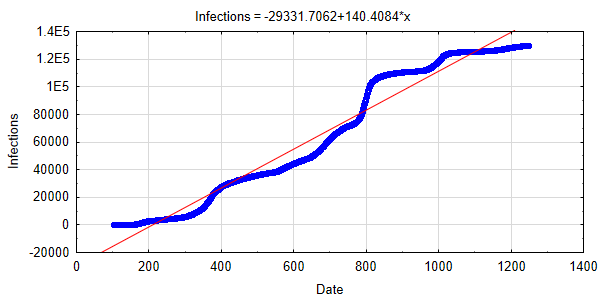


Рисунок Линейная модель

Очевидно, что линейная модель обладает существенными недостатками и не отражает динамики изменения заболеваемости.

Скрытая марковская модель процесса

Для анализа используем библиотеку Statsmodels языка Python (существуют и другие модули, например, HMMLearn).

Алгоритм применения модуля (при условии, что необходимые данные загружены) включает следующие этапы:

1. Выбор количества состояний модели
2. Обучение модели на имеющихся данных
3. Формирование матрицы переходов
4. Сопоставление наблюдаемой переменной найденным состояниям

Здесь (см. таблица 1) создаются датасеты с заболеваемостью (накопленным итогом и посуточно), задается число состояний системы (k\_reg = 2) и происходит обучение модели.

Таблица 1. Создание модели и ее обучение

|  |
| --- |
| import statsmodels.api as sm  df1 = df[['Date','Infections']]  df2 = df[['Date','Inf\_day']]  k\_reg = 2  # Fit the model  # (a switching mean is the default of the MarkovRegession model)  mod = sm.tsa.MarkovRegression(df1['Infections'], k\_regimes=k\_reg)  res = mod.fit()  mod1 = sm.tsa.MarkovRegression(df1['Inf\_day'], k\_regimes=k\_reg)  res1 = mod1.fit() |

Таблица 2. Результаты

|  |
| --- |
| HMM для переменной Infections |
| Markov Switching Model Results  ==============================================================================  Dep. Variable: Infections No. Observations: 1149  Model: MarkovRegression Log Likelihood -12969.218  Date: Wed, 12 Jul 2023 AIC 25948.436  Time: 09:45:32 BIC 25973.669  Sample: 0 HQIC 25957.962  - 1149  Covariance Type: approx  Regime 0 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 2.394e+04 906.539 26.408 0.000 2.22e+04 2.57e+04  Regime 1 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 1.103e+05 997.925 110.507 0.000 1.08e+05 1.12e+05  Non-switching parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  sigma2 3.865e+08 0.011 3.48e+10 0.000 3.87e+08 3.87e+08  Regime transition parameters |
| HMM для переменной Inf\_day |
| Markov Switching Model Results  ==============================================================================  Dep. Variable: Inf\_day No. Observations: 1149  Model: MarkovRegression Log Likelihood -6851.957  Date: Wed, 12 Jul 2023 AIC 13713.914  Time: 09:56:45 BIC 13739.147  Sample: 0 HQIC 13723.439  - 1149  Covariance Type: approx  Regime 0 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 93.7357 2.775 33.778 0.000 88.297 99.175  Regime 1 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 829.9154 19.285 43.033 0.000 792.117 867.714  Non-switching parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  sigma2 8541.9266 356.553 23.957 0.000 7843.095 9240.759  Regime transition parameters  ... |

Интерпретация результатов. Для случая двух режимов система существует в двух состояниях, с условно низкой заболеваемостью (93,7 чел/сутки) и условно высокой (829,9 чел/сутки). Переключение происходит на 618 день пандемии и соответствует началу 5-ой волны пандемии (см. рисунок 3).

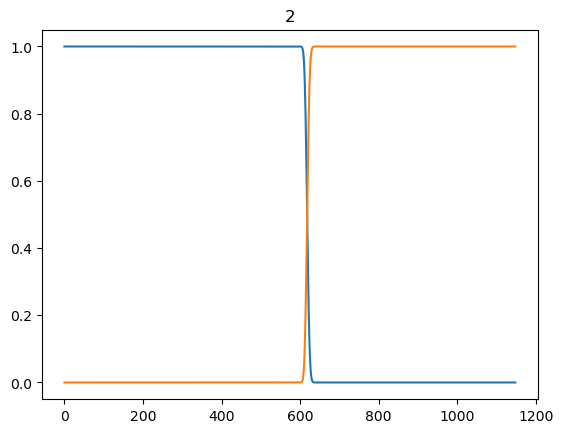


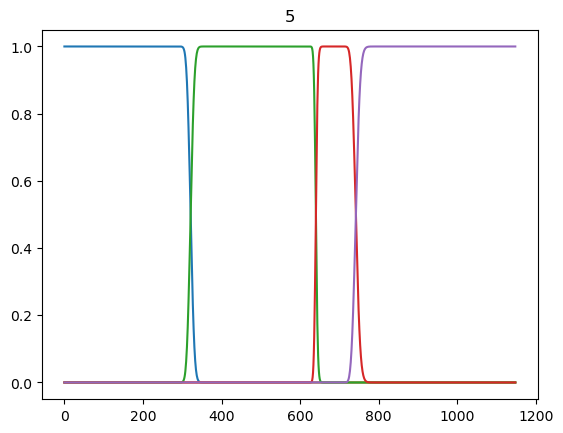
Рисунок Переключение состояний

Если принять в качестве количества состояний количество волн пандемии (5), то аналогичным образом, получим (см. приложение 2):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Режим | 1 | 2 | 3 | 4 | 5 |
| Заболеваемость (чел/сутки) | 76 | 45,5 | 6,8 | 61 | 455 |
| Период с/по (дни) | 0-310 | Не реализуется | 310-640 | 640-760 | 760- |

Заметим, что период 2 не реализуется, хотя и выявлен, периоды 1-4 соответствуют периодам волн пандемии, а 5-ый период соответствует этапу фактического снятия эпидемических ограничений и мер социального дистанцирования.

Аналогичные результаты (с точностью до номеров состояний) были получены с применением библиотеки HMMLearn (см.



1

3

4

5

Рисунок Переключение режимов с 1-го по 5-ый (режим 2 не реализуется)

Выводы

В исследовании показана возможность использования скрытых марковских моделей для оценки заболеваемости COVID-19. Отмечается, что пандемия состояла, фактически, из четырех состояний. Первые три из них соответствуют волнам пандемии и различной интенсивности посуточной заболеваемости (76, 6.8 и 61 чел/сутки), четвертый режим реализовался в условиях отказа от ограничительных мер и появления нового доминирующего штамма с повышенной контагиозностью и относительно легкой переносимостью пациентами.

Последний результат показывает высокую эффективность мер, принятых государством в период пандемии.

Данные, код на языке Python и рисунки к статье доступны в виде публичного репозитория по адресу: [Alex-Samarkin/TSRevenko (github.com)](https://github.com/Alex-Samarkin/TSRevenko)

Литература

Приложение 1. Загрузка данных

# %%

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

file = "COVID\_PSK.csv"

df = pd.read\_csv(file,sep=';',encoding='utf-8',index\_col='Date')

df["inf\_day"] = df["Infections"].diff(periods=1)

print(df.head())

print(df.columns.tolist())

#%%

# Assuming your DataFrame is called 'df'

df['Date'] = df.index

#%%

# Convert Date column to numerical format

df['Date'] = pd.to\_datetime(df['Date'],format='%d.%m.%Y',dayfirst=True)

print(df['Date'])

#%%

# Convert Date column to numerical format and convert from nanoseconds to days

df['Date'] = pd.to\_numeric(df['Date']-df['Date'].min()) / (24\*60\*60\*1000000000)

print(df['Date'])

#%%

if \_\_name\_\_ == '\_\_main\_\_':

    #plot the data using seaborn versus Date

    fig, axes = plt.subplots(2,1, figsize=(10, 8))

    sns.lineplot(data= df, ax =axes[0])

    sns.lineplot(data= df, x="Date", y="inf\_day", ax =axes[1])

    plt.show()

    print("Save to csv for using next time1")

    df.to\_csv("df.csv")

Приложение 2 Результаты анализа при 5 состояниях системы

|  |
| --- |
| Markov Switching Model Results  ==============================================================================  Dep. Variable: Infections No. Observations: 1149  Model: MarkovRegression Log Likelihood -12440.463  Date: Wed, 12 Jul 2023 AIC 24932.927  Time: 10:17:17 BIC 25064.140  Sample: 0 HQIC 24982.460  - 1149  Covariance Type: approx  Regime 0 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 1.85e+04 719.454 25.719 0.000 1.71e+04 1.99e+04  Regime 1 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 2.521e+04 nan nan nan nan nan  Regime 2 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 4.254e+04 623.332 68.254 0.000 4.13e+04 4.38e+04  Regime 3 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 9.91e+04 nan nan nan nan nan  Regime 4 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 1.169e+05 578.924 201.852 0.000 1.16e+05 1.18e+05  Non-switching parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  sigma2 1.261e+08 nan nan nan nan nan  Regime transition parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  p[0->0] 0.9988 nan nan nan nan nan  p[1->0] 1.0000 135.602 0.007 0.994 -264.774 266.774  p[2->0] 0 nan nan nan nan nan  p[3->0] 0 nan nan nan nan nan  p[4->0] 0.0011 0.000 2.237 0.025 0.000 0.002  p[0->1] 0 1.79e-06 0 1.000 -3.51e-06 3.51e-06  p[1->1] 0 115.424 0 1.000 -226.226 226.226  p[2->1] 0 nan nan nan nan nan  p[3->1] 0 nan nan nan nan nan  p[4->1] 0 nan nan nan nan nan  p[0->2] 0.0012 0.001 0.999 0.318 -0.001 0.004  p[1->2] 0 47.465 0 1.000 -93.030 93.030  p[2->2] 0.9965 nan nan nan nan nan  p[3->2] 0 nan nan nan nan nan  p[4->2] 0 nan nan nan nan nan  p[0->3] 0 0.008 0 1.000 -0.017 0.017  p[1->3] 0 11.630 0 1.000 -22.794 22.794  p[2->3] 0.0035 0.004 0.998 0.318 -0.003 0.010  p[3->3] 0.9897 0.009 105.236 0.000 0.971 1.008  p[4->3] 0 nan nan nan nan nan  ============================================================================== |
| Markov Switching Model Results  ==============================================================================  Dep. Variable: Inf\_day No. Observations: 1149  Model: MarkovRegression Log Likelihood -6942.504  Date: Wed, 12 Jul 2023 AIC 13937.008  Time: 10:21:05 BIC 14068.221  Sample: 0 HQIC 13986.541  - 1149  Covariance Type: approx  Regime 0 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 76.0057 3.361 22.616 0.000 69.419 82.593  Regime 1 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 45.5778 121.947 0.374 0.709 -193.435 284.590  Regime 2 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 6.8495 53.856 0.127 0.899 -98.707 112.406  Regime 3 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 61.4616 46.514 1.321 0.186 -29.703 152.627  Regime 4 parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  const 455.7064 16.145 28.225 0.000 424.062 487.351  Non-switching parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  sigma2 9778.5900 417.190 23.439 0.000 8960.912 1.06e+04  Regime transition parameters  ==============================================================================  coef std err z P>|z| [0.025 0.975]  ------------------------------------------------------------------------------  p[0->0] 0.9962 0.002 519.833 0.000 0.992 1.000  p[1->0] 0.0033 22.024 0.000 1.000 -43.164 43.170  p[2->0] 1.0000 0.005 183.080 0.000 0.989 1.011  p[3->0] 0.9717 285.843 0.003 0.997 -559.271 561.214  p[4->0] 1.468e-24 nan nan nan nan nan  p[0->1] 4.361e-07 nan nan nan nan nan  p[1->1] 3.293e-13 nan nan nan nan nan  p[2->1] 5.802e-15 nan nan nan nan nan  p[3->1] 0.0283 nan nan nan nan nan  p[4->1] 1.685e-14 nan nan nan nan nan  p[0->2] 1.6e-08 nan nan nan nan nan  p[1->2] 0.9963 83.034 0.012 0.990 -161.747 163.740  p[2->2] 1.811e-29 nan nan nan nan nan  p[3->2] 6.383e-06 712.574 8.96e-09 1.000 -1396.619 1396.619  p[4->2] 0.0346 0.017 2.062 0.039 0.002 0.068  p[0->3] 1.536e-10 2.42e-06 6.34e-05 1.000 -4.75e-06 4.75e-06  p[1->3] 1.749e-08 nan nan nan nan nan  p[2->3] 7.793e-38 nan nan nan nan nan  p[3->3] 3.072e-07 nan nan nan nan nan  p[4->3] 1.245e-07 nan nan nan nan nan  ============================================================================== |

Приложение 3. Состояния системы, рассчитанные в модуле HMMLearns

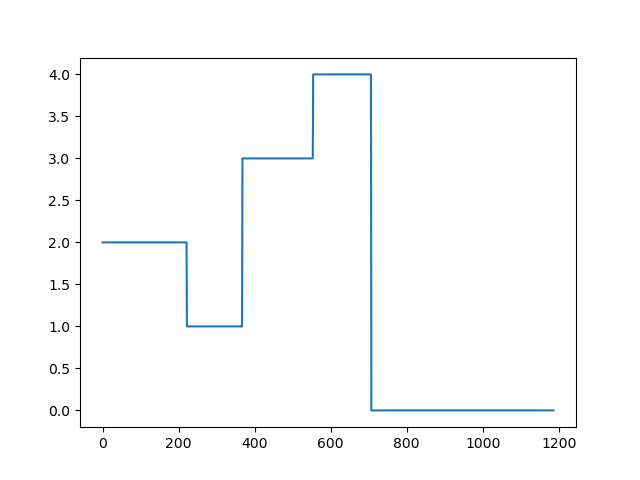
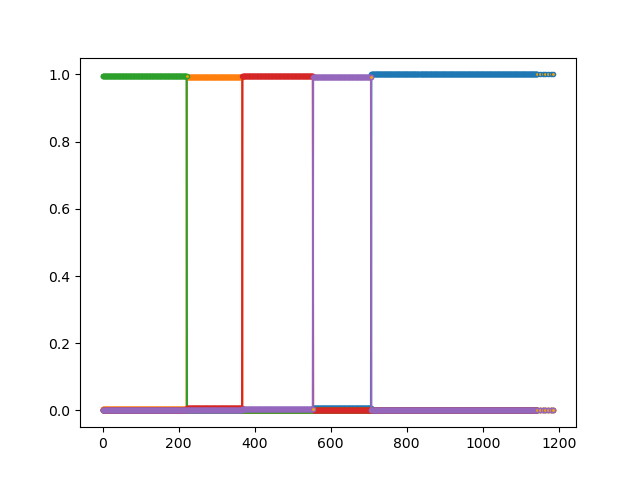
 

Рисунок Состояния системы, рассчитанные ы HMMLearns