УДК 57.032;519.688

# ПРОГНОЗИРОВАНИЕ РАСПРОСТРАНЕНИЯ COVID-19 НА РЕГИОНАЛЬНОМ УРОВНЕ НА ПРИМЕРЕ ХАНТЫ-МАНСИЙСКОГО АВТОНОМНОГО ОКРУГА - ЮГРЫ

***Балуев В.А., Бурлуцкий В.В., Керамов Н.Д., Мельников А.В.***
*Югорский НИИ информационных технологий, г. Ханты-Мансийск, Россия*

**Введение.** Пандемия COVID-19 потребовала от общества и органов государственной власти определить адекватные ответы на вопросы о необходимом уровне карантинных мер, степени мобилизации системы здравоохранения, обоснованной модели социального поведения, допустимого уровня экономических потерь. Для решения этих задач необходимо прогнозирование развития ситуации с определением основных показателей пандемии. В настоящее время такие прогнозы строятся для отдельных стран. Однако для такой большой страны как Россия с различными климатическими условиями, разными моделями региональной экономики, различными социальными моделями поведения такой усредненный подход дает неизбежную ошибку. Задачей данного исследования было провести анализ существующих математических моделей и практических результатов их применения для решения задачи прогнозирования пандемии на уровне региона с последующей их валидацией на основе данных по предыдущим временным периодам.

**Ключевые слова:** *COVID-19, coronavirus, prediction model, mathematical modeling, curve fitting, regression analysis, SEIR model, forecasting.*

## Обзор подходов

Математические модели стали важным инструментом теоретической эпидемиологии. Первые математические модели для прогнозирования развития инфекционных заболеваний появились в начале XX века. В 1927 году Кермак и Маккендрик предложили для проведения расчетов использовать компартментальную модель SIR[1-2], которая разделяет человеческую популяцию на людей, восприимчивых к болезни (S), и тех, что уже переболели (R). Восприимчивые заражались с некоторой скоростью

передачи болезни (R0), становясь инфицированными (I), которые в свою очередь поправлялись с некоторой скоростью.

Эта модель послужила основой для развития последующих моделей посредством вносимых модификаций, которые заключались в изменении уравнений или добавлении к расчету других лиц, не относящихся к трем указанным базовым категориям, что позволяло учитывать особенности тех или иных заболеваний. Были разработаны модели, принимающие во внимание возможность повторного инфицирования (модель SIS), смерти (SIRD), наличия у заболевания инкубационного периода (SEIR), временного иммунитета детей благодаря антителам матери (MSIR), повторного инфицирования и наличия инкубационного периода (SEIS) и т.д.

Недавно было предложено несколько модификаций этой модели для прогнозирования тенденций эпидемии COVID-19 [3-5]. Хотя эти работы и улучшили классическую модель SIR в некоторой степени, все еще существуют некоторые ограничения. Например, такие модели не учитывают влияние домашнего карантина и изоляции пациентов в обсерваторах (ковидариях).

Кроме того, уровень контакта инфекции в этих моделях рассматривается как фиксированное значение, что не соответствуют реальной эпидемической ситуации [6].

**Анализ исходных данных**

Прогнозирование распространения COVID-19 осложняется целым рядом проблем. Перечислим основные из них:

1) неизвестный характер нового вируса;
2) отчетные данные могут быть неполными и непоследовательными (встречаются пропуски данных);
3) проблема тестирования, влияющая количественные и качественные характеристики данных (задержки в тестировании приводят к искажению картины распространения эпидемии).

Для прогнозирования распространения COVID-19 на региональном уровне необходимы актуальные данные, характеризующие течение пандемии в каждом отдельно взятом регионе. Изучив имеющиеся открытые источники данных о распространении COVID-19 в регионах России, были выделены

следующие ключевые параметры: накопленные количественные значения заражений, выздоровлений, смертей, а также их ежедневные приросты.

Рассмотрим такую характеристику пандемии, как количество подтвержденных заболеваний. Во-первых, данная характеристика не в полной мере отражает реальное количество больных, так как около 80% случаев заражения COVID‑19 характеризуются легким или бессимптомным течением [7], а тестирование на наличие вируса, в подавляющем большинстве, проводят при проявлении явных симптомов. Во-вторых, фиксирование заболевания происходит после получения результатов анализов на COVID-19, результаты которых становятся известными с определенной задержкой. Этот фактор, особенно на ранних этапах пандемии сильно сказывается на неравномерности роста количества подтвержденных случаев заболевания, так могут возникать логистические задержки доставки биоматериала, перегрузки лабораторий и прочие факторы, замедляющие получение результатов анализов. Данные факторы еще сильнее влияют на параметр суточного прироста заражений.

Первоначально для анализа использовались данные регионального Роспотребнадзора по ХМАО – Югре, который аккумулирует сведения от больниц, работающих с больными COVID-19. Однако на начальном этапе предоставленные данные содержали неточности и пропуски, которые нуждались в дополнительной валидации и обработке.

В ходе исследования было решено перейти к данным Роспотребнадзора России, которые публикуются сервисом Yandex DataLens [8]. Эти сведения использовались при построении прогнозов по региону в целом, а данные регионального Роспотребнадзора использовались, в первую очередь для уточнения анализа развития ситуации в регионе в разрезе муниципалитетов, а после накопления достаточного количества данных и для построения прогнозов. Однако, ввиду наличия уже упомянутых задержек в тестировании, все исходные данные содержали явные значительные выбросы в данных. Поэтому для повышения качества прогнозов выполнялось сглаживание исходных данных методом скользящего среднего с размером окна равным 7 суток.

Ключевыми показателями для прогнозирования были выбраны:

1) накопленное количество заражений;

2) суточный прирост количества заражений;
3) накопленное количество выздоровлений;
4) накопленное количество смертей.

Суточное изменение количества смертей от COVID-19 не является статистически значимой характеристикой, так как в большинстве отдельно взятых регионов России (в том числе в ХМАО – Югре), количество смертей в день варьируется от 1 до 10 в день.

Сведения полученные от сервиса Yandex DataLens не содержат классификации по степени тяжести заболевания. Прогнозы по степени тяжести больных не строились, а оценка количества больных средней и тяжелой степени производилась по прогнозам накопленного количества заражений с учетом того, какой процент от общего количества заражений они составляют (данный процент определялся медиками по результатам их наблюдений).

**Регрессионный анализ**

Одним из наиболее распространенных количественных методов прогнозирования является регрессия. Прогнозирование на основе экстраполяции по регрессионной модели успешно применяется для построения краткосрочных прогнозов. Для предсказания накопленного количества заражений, выздоровлений, смертей, а также их прироста в ХМАО – Югре, использовались линейная, полиномиальная и экспоненциальная регрессия. Однако по мере увеличения объема анализируемых данных было решено отказаться от линейной модели как имеющей низкую точность.

В качестве инструмента для анализа и прогнозов использовался язык программирования Python с математическими пакетами NumPy и SciPy. Исходные данные представлены 5 наборами данных:

1) Накопленное количество заражений, выздоровлений, смертей, а также их приросты по регионам России (Источник – сервис Yandex DataLens);
2) Накопленное количество заражений, выздоровлений, смертей, а также их приросты по региону (Источник – Роспотребнадзор по ХМАО – Югре);

3) Накопленное количество заражений, выздоровлений, смертей, а также их приросты по странам мира (Источник – сервис Yandex DataLens);

4) Индекс самоизоляции по городам России (Источник – сервис Yandex DataLens);

5) Статистические данные по регионам России: численность населения, соотношение городского и сельского населения, плотность населения, средний возраст населения, численность населения в возрасте 65 лет и более, среднемесячная заработная плата (Источник – Росстат РФ).

При построении краткосрочных прогнозов указывается регион, источник данных (Yandex DataLens или другой, на примере ХМАО – Югры это Роспотребнадзор по ХМАО – Югре), дата начала выборки данных, количество дней прогноза, количество исключаемых с конца временного ряда дней, размер окна сглаживания, точность выгружаемых прогнозных значений. Показателями для анализа и прогнозирования являются накопленное количество заражений, выздоровлений, смертей, а также их суточные изменения. Из исходных данных по регионам России выполнялась выборка данных с 15.04.2020 (дата достижения 100 подтвержденных случаев заражений для ХМАО – Югры) и строились прогнозы по накопленному количеству заражений, выздоровлений, смертей и их приростам. Исходные данные сглаживались методом скользящего среднего с размером окна равным 7 суток. При необходимости модель позволяет исключить из выборки заданное количество последних дней, что позволяет в дальнейшем оценить точность прогнозов путем сравнения фактических и прогнозных значений показателей. Прогнозы строились на 7 суток.

Для построения прогноза методом экспоненциальной регрессии использовалась функция вида

$$f(x) = a \cdot e^{-b \cdot x + c} + d \qquad (1)$$

где a, b, c, d – коэффициенты, вычисляемые по исходным данным.

По итогам прогнозов строятся графики по каждому из показателей. Все численные данные по прогнозам сохраняются в файл формата Microsoft Excel. На рисунке 1 показан график прироста заражений с прогнозными кривыми, а на рисунке 2 пример отчета по показателю количества заражений.

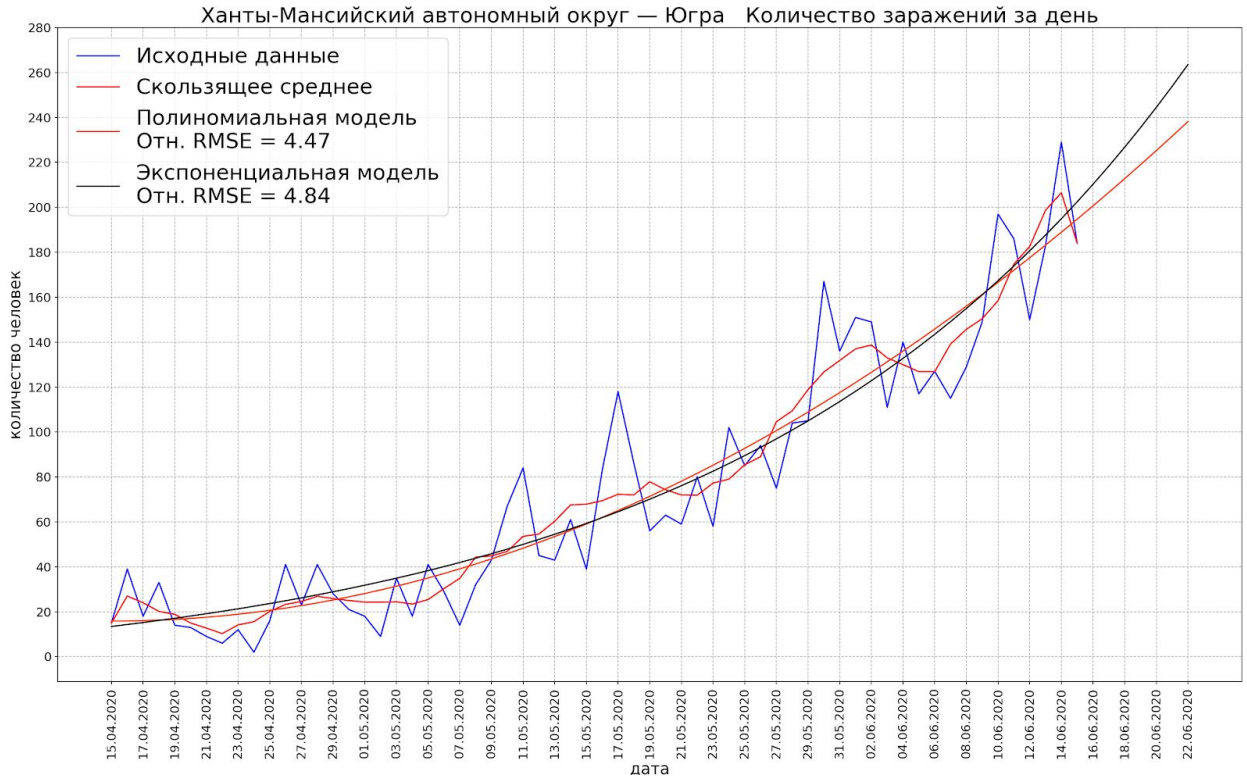


Рисунок 1. График прироста заражений с прогнозными кривыми на примере ХМАО – Югры

| Регион | Показатель | Дата | Исходные данные | Скользящее среднее | Полиномиальная модель | Экспоненциальная модель |
|---|---|---|---|---|---|---|
| ХМАО – Югра | Заражений за день | 15.04.2020 | 15 | 15 | 16 | 13 |
| ХМАО – Югра | Заражений за день | 16.04.2020 | 39 | 27 | 16 | 14 |
| ХМАО – Югра | Заражений за день | 17.04.2020 | 18 | 24 | 16 | 15 |
| ХМАО – Югра | Заражений за день | 18.04.2020 | 33 | 20 | 16 | 16 |
| ... | | | | | | |
| ХМАО – Югра | Заражений за день | 13.06.2020 | 183 | 199 | 183 | 188 |
| ХМАО – Югра | Заражений за день | 14.06.2020 | 229 | 206 | 189 | 195 |
| ХМАО – Югра | Заражений за день | 15.06.2020 | 184 | 184 | 195 | 202 |
| ХМАО – Югра | Заражений за день | 16.06.2020 | нет | нет | 201 | 210 |
| ХМАО – Югра | Заражений за день | 17.06.2020 | нет | нет | 207 | 218 |
| ХМАО – Югра | Заражений за день | 18.06.2020 | нет | нет | 213 | 227 |
| ХМАО – Югра | Заражений за день | 19.06.2020 | нет | нет | 219 | 235 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ХМАО – Югра | Заражений за день | 20.06.2020 | нет | нет | 225 | 245 |
| ХМАО – Югра | Заражений за день | 21.06.2020 | нет | нет | 232 | 254 |
| ХМАО – Югра | Заражений за день | 22.06.2020 | нет | нет | 238 | 264 |

Таблица 1. Прирост количества заражений и прогнозные значения примере ХМАО – Югры

Оценка точности прогнозов производилась по значениям MSE, RMSE, MAE, MAPE, которые вычислялись для каждого прогнозного показателя. Погрешность прогнозов на горизонте 7-10 дней по накопленному количеству заражений, приросту заражений и накопленной смертности не превышала 10%. При этом точность отдельных видов экстраполяции менялась с течением эпидемии. Если на начальном этапе точность экспоненциальной модели была выше, то по мере приближения к пику заболеваемости полиномиальная модель дает более достоверные прогнозы. Полученные на основе регрессионного анализа краткосрочные прогнозы использовались при подготовке аналитических справок для Оперативного штаба по коронавирусу Ханты-Мансийского автономного округа - Югры.

**SIR-модель и ее модификации**

Одним из существенных недостатков прогнозных моделей на основе регрессионного анализа является то, что они не учитывают закономерности развития заболеваний в популяции, и как следствие, не могут предсказать пик заболеваний. Поэтому другим важным направлением прогнозирования пика пандемии является использование компартментальных моделей, которые рассматривают популяцию как совокупность нескольких групп. Классическая SIR-модель включает восприимчивых (англ. susceptible), больных (англ. infectious) и выздоровевших (англ. recovered). Применение данной модели позволяет достаточно точно моделировать эпидемии гриппа и других заболеваний в больших городах, вводить новые параметры и анализировать разные сценарии. Но использование модели для анализа распространения эпидемий гриппа и других подобных заболеваний не гарантирует адекватность модели в случае с COVID-19. Существуют модификации, например учитывающие тех, у кого болезнь находится в инкубационном периоде (англ. exposed), и умерших (англ. dead)[9]. Однако SIR-модель и ее модификации  слабо интерпретируемы, а подбор их

параметров неявно зависит от множества факторов: предпринимаемые государством меры ограничения, общий уровень соблюдения изоляции, доля бессимптомных больных и вероятность заражения ими других членов популяции, уровень выявления зараженных путем тестирования, задержки в отчетах об испытаниях и ряда других. Кроме того, модели данного класса в большинстве случаев исключают из рассмотрения фактор мобильности населения. Очевидно также, что каждая популяция имеет неоднородный характер по возрастному составу, роду деятельности и прочим параметрам, который в моделях не учитывается. Мобильность групп внутри популяции влияет на скорость передачи инфекции как внутри группы, так и между ними.

Также при моделировании необходимо учитывать и характер самого заболевания. Согласно данным ВОЗ COVID-19 и грипп схожи по клиническим проявлениям болезни, однако COVID-19 имеет более длительный инкубационный период (время от момента заражения до возникновения симптомов) и большее время генерации (время между заражением одного человека и заражением другого)[7]. Многие закономерности распространения эпидемии гриппа в настоящий момент применимы с ограничениями или не применимы вовсе, ввиду малого количества данных и, как следствие, научных исследований о характере распространения коронавирусной инфекции.

**Описательные математические модели**

В настоящее время также находят применение методы прогнозирования, основанные на описательных математических моделях. Анализ динамики наблюдаемых данных позволяет разделить весь временной ряд данных на несколько характерных интервалов, характеризующихся отдельным трендом изменения данных. Характерные тренды наблюдаются даже несмотря на неоднородности и неопределенности во временных рядах данных и различия в мерах социального ограничения. Например, в работе [10] отмечается наличие характерного перехода от длительного периода экспоненциального роста количества активных случаев (количество больных членов популяции) к короткому периоду с полиномиальным ростом и последующим достижением пика эпидемии. Количество активных случаев определяется как разность между накопленным количеством заражений и суммой накопленного количества выздоровлений и смертей. После

прохождения пика эпидемии количество активных случаев постепенно уменьшается. Кривая спада определяется следующей формулой:

$$I(t) = A/T_G \cdot (t/T_G)^{\alpha} \cdot e^{-t/T_G} \quad (2)$$

где A, $T_G$, α – коэффициенты, вычисляемые по исходным данным.

Несмотря на отсутствие строго доказанной явной функциональной зависимости между количеством больных и динамикой распространения заболевания, данная модель может использоваться для прогнозирования пика нагрузки на систему здравоохранения и оценки количества необходимых коек. На рисунке 2 показан построенный по результатам моделирования прогноз изменения количества больных на примере ХМАО – Югры.

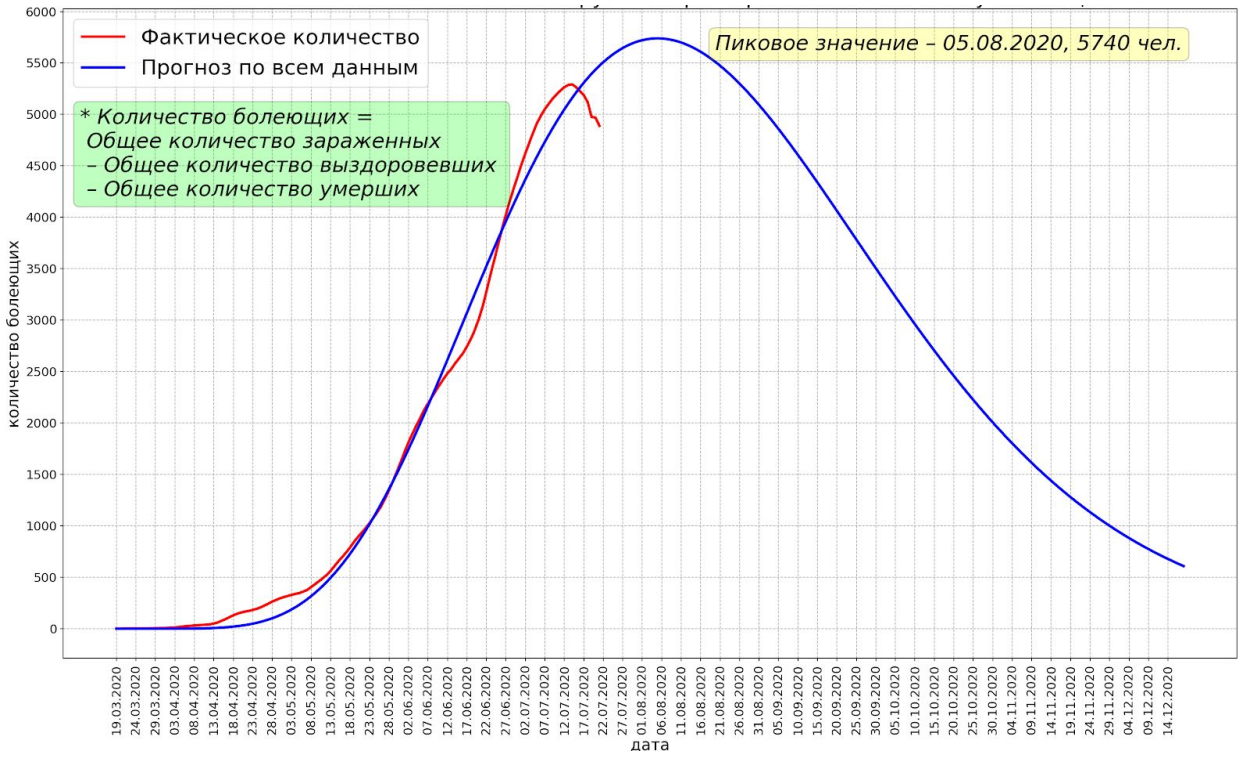


Рисунок 2. Прогноз количества больных на примере ХМАО – Югры

Отметим, что описанная выше модель может использоваться для прогнозирования только после прохождения пика заболеваемости, когда количество больных в популяции начинает постепенно убывать. При этом необходимо получить достаточное количество данных, свидетельствующих о

тренде снижения количества больных в популяции. На рисунке 3 показаны варианты прогнозов в зависимости от исходных данных. Рост количества больных приводит к смещению даты пика на более поздний срок, а спад – на более ранний. До достижения пика лучшие результаты показывают, как и в случае с регрессионным анализом, вначале экспоненциальная, а затем полиномиальная модели роста.

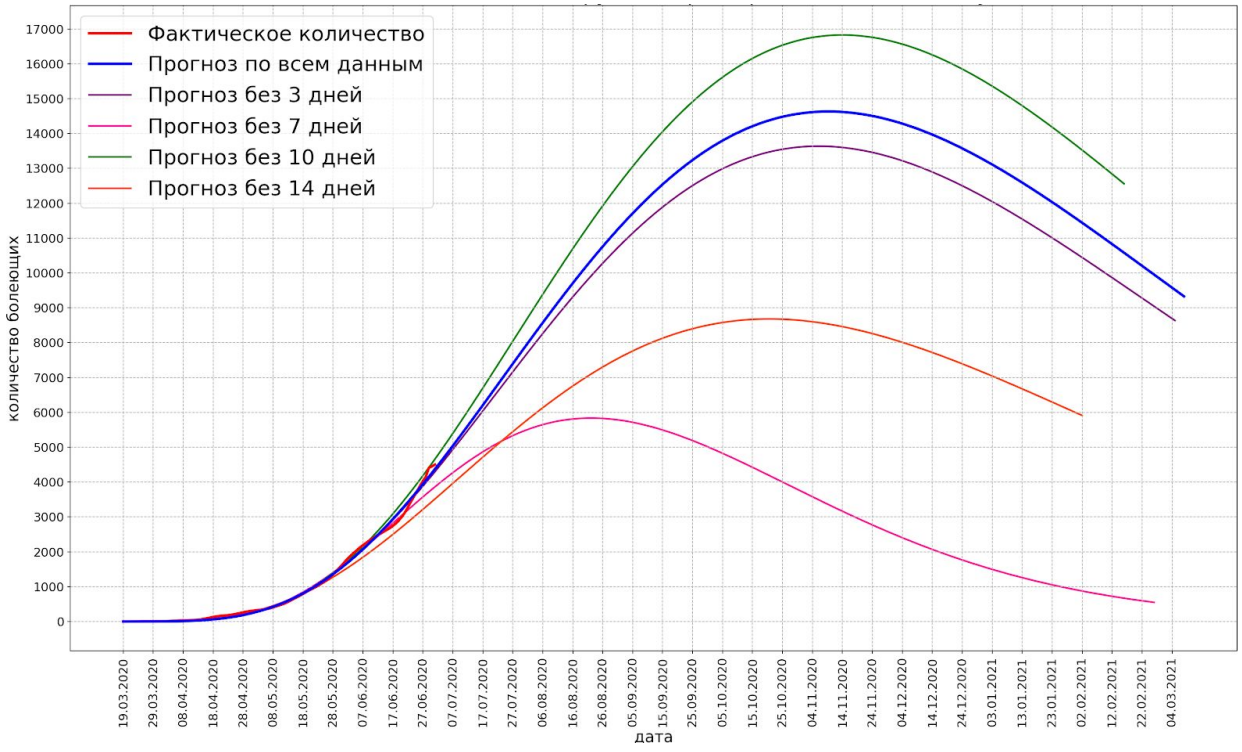


Рисунок 3. Изменение прогнозов количества больных в зависимости от исходных данных на примере ХМАО – Югры

**Модель прогнозирования на основе аналогий**

Помимо рассмотренных моделей прогнозирования в ходе работы применялся также и метод прогнозирования по аналогии. В данном случае, идет речь об исторической аналогии, как методе, основанном на установлении и использовании аналогии объекта прогнозирования с одинаковым по природе объектом, опережающим первый в своем развитии. Так, для прогнозирования развития эпидемии в ХМАО – Югре использовались сведения о динамике распространения COVID-19 в странах и регионах, которые уже прошли пик прироста заражений и при этом имеют

близкие к объекту прогнозирования параметры (половозрастной состав, особенности климата, состояние экономики).

Для сравнительного анализа развития пандемии в ХМАО - Югре были выбраны такие страны как Канада, Норвегия, Финляндия, Швеция, США, Германия, Россия, а также город Санкт-Петербург. Для каждого из выбранных объектов на графике отображался относительный прирост заражений в день на 1 млн. человек. График от точки достижения первых 100 накопленных заражений приведен на рисунке 4.

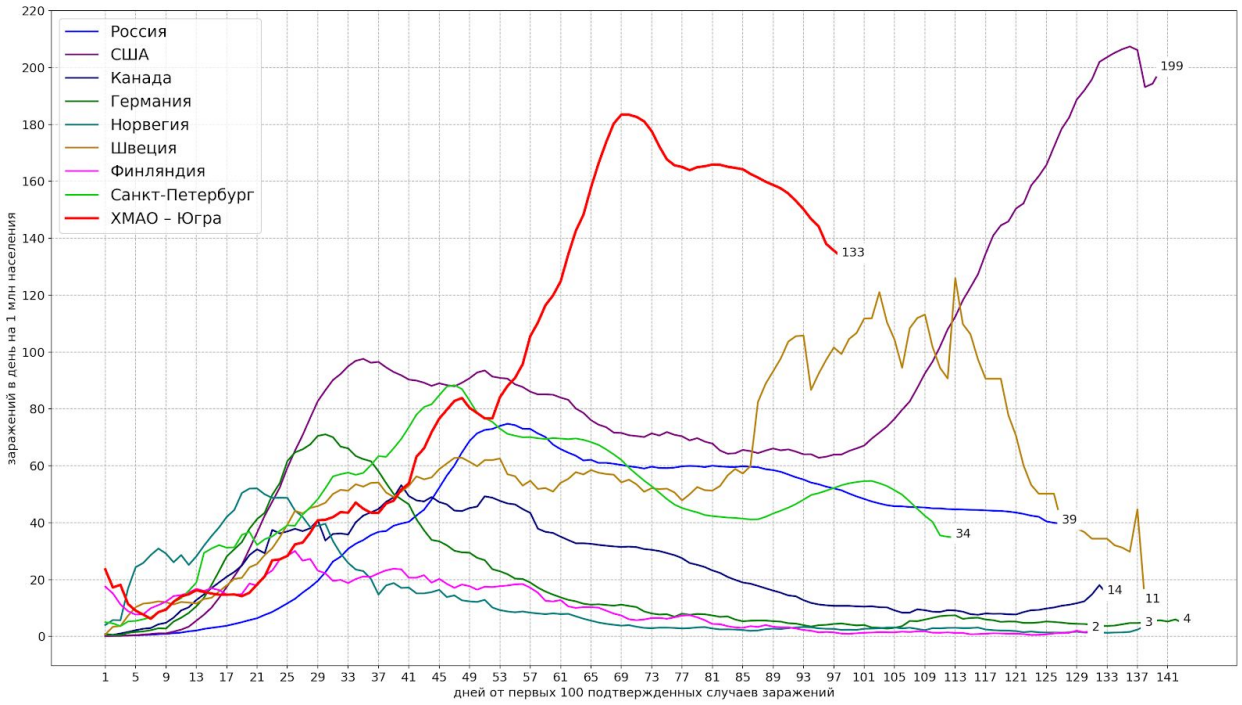


Рисунок 4. Прирост заражений для различных стран и регионов

Описанный метод использовался для предсказания приблизительной даты пика заболеваемости в регионе. Полученные при этом прогнозы оказались достаточно точными: дата пика была определена с точностью 4-5 дней. Конечно данный метод не позволяет получить достоверных количественных значений прогнозных показателей. В нашем случае прогноз накопленного количества подтвержденных случаев заболевания с начала пандемии в момент пика заболеваемости оказался существенно ниже фактических значений. По нашей оценке это связано с высокой мобильностью населения и завозными случаями заражения вахтовых рабочих в Ханты-Мансийском автономном округе - Югре.

## Выводы

Проведенный анализ существующих подходов к прогнозированию распространения коронавируса и практическая апробация существующих математических моделей на конкретных данных отдельного региона Российской федерации позволяет сделать следующий выводы. Ни одна из существующих моделей не позволяет с достаточной точностью прогнозировать основные параметры распространения пандемии на достаточно большой период времени (4-6) месяцев. На коротких интервалах (7-10 дней) хорошо работают традиционные регрессионные модели на основе полиномиальной или экспоненциальной зависимости. SEIR-модели дают прогнозы существенно превышающие реальные значения накопленных уровней заражений COVID-19, а погрешность прогнозирования пика пандемии составляет 1-2 месяца. Для качественного анализа развития пандемии хорошие результаты показывает метод аналогий, но он не дает точных количественных оценок.

**Литература**


1. Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115(772), 700-721.
2. Hethcote HW. The mathematics of infectious diseases. Siam Review 2000; 42(4): 599-653.
3. Dye C, Gay N. Modeling the SARS epidemic. Science 2003; 300(5627): 1884-5.
4. Riley S, Fraser C, Donnelly CA, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. Science 2003; 300(5627): 1961-6.
5. Wang WD, Ruan SG. Simulating the SARS outbreak in Beijing with limited data. Journal of Theoretical Biology 2004; 227(3): 369-79.
6. Wang L, Wu JT. Characterizing the dynamics underlying global spread of epidemics. Nature Communications 2018; 9.
7. Вопросы и ответы: сходства и различия возбудителей COVID‑19 и гриппа. / Всемирная организация здравоохранения. URL: https://www.who.int/ru/news-room/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza (дата обращения: 20.07.20).
8. Yandex DataLens [Электронный ресурс] : Пресет Коронавирус: Дашборд и данные. URL: https://datalens.yandex.ru/datasets (дата обращения: 20.08.2007).
9. Nicholas B Noll, Ivan Aksamentov, Valentin Druelle, Abrie Badenhorst, Bruno Ronzani, Gavin Jefferies, Jan Albert, Richard Neher. COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2, 2020
10. K. Bod’ov´a, R. Koll´ar Emerging Polynomial Growth Trends in COVID-19 Pandemic Data and Their Reconciliation with Compartment Based Models, 2020