Научная визуализация, 2020, том 12, номер 4, страницы 85 - 97, DOI: 10.26583/sv.12.4.08

Метод визуального анализа динамики изменения структурированных данных на основе цветовых маркеров

А.А. Трубакова[1], А.О. Трубаков[2]

Брянский государственный технический университет

[1] ORCID: 0000-0003-0280-1760, trubakovaaa@gmail.com
[2] ORCID: 0000-0003-0058-1215, trubakovao@gmail.com

Аннотация

Одной из важных задач для систем, помогающих в принятии решений, является задача наглядной и интуитивно понятной визуализации исходных данных. При правильном и удобном отображении данный инструмент может стать очень серьезным помощником для лица, принимающего решение. От подобного инструмента зависит не только трудоемкость процесса принятия решений, но и корректность и объективность принятых решений. Именно поэтому данному вопросу уделяется столько внимания в различных исследованиях. Однако на сегодняшний день вопрос отображения динамики изменения структурированных данных (данных, в которых наблюдаемая величина имеет внутреннюю структуру и состоит из большого числа компонентов) недостаточно проработан. В данной работе основное внимание уделено именно этим вопросам, приведено формальное описание структурированных данных, предлагается подход к отображению динамики их изменения. Предложенный подход к визуализации позволяет увидеть, как общий процесс изменения наблюдаемой величины, так и характер изменения её внутренней структуры. Отдельные диаграммы позволяют рассмотреть вклад каждой компоненты структурированной величины и оценить динамику этого вклада во времени. Так же в статье предложена область потенциального применения данного подхода, выделены его особенности и основные возможности. Анализ данных, предложенный в работе, был получен на основе разработанного программного комплекса моделирования ситуации с распространением COVID-19 в Брянской области, в котором использовался предложенный вариант визуализации. Рассмотрены ситуации, в которых предложенный подход к визуализации помогает получить новую информацию, не видимую при использовании других подходов к отображению структурированных величин.

Ключевые слова: визуализация данных, структурированные данные, принятие решений, SEIRD-модели.

1. Введение

Очевидным фактом является то, что объем данных, который необходимо анализировать при принятии управленческих решений, растет с каждым годом большим темпом. При этом анализ и конечное решение продолжает оставаться прерогативой человека и зависит от лица, принимающего решение (ЛПР). Современные системы поддержки принятия решений (СППР), использующие в своей основе математические методы, позволяют частично заменить человеческие ресурсы на этапе обработки исходных данных, но не на этапе окончательного принятия решения. Отсутствие четкой и структурированной информации в ходе процесса принятия решений затрудняется, в виду обработки данных различного характера и природы, частоты обновления инфор-

мации в течение короткого периода времени, при этом поддержка, принятие решения всегда остается за ЛПР. В связи с такой потребностью в оперативных исследованиях, возникает необходимость в качественной и наглядной визуализации постоянно растущего и развивающегося объема данных, которые быстро появляются в результате глобальных научных исследований во всем мире [1]. Круг исследований, которые активно привлекают визуализацию в качестве средства взаимодействия с данными произвольного типа, постоянно расширяется. В свою очередь, это закономерно приводит к значительному увеличению объемов данных, усложнению их структуры, использованию в анализе данных новых типов в качестве вспомогательных или, даже, основных, источников информации [2].

Особенно важно развивать методы визуализации, если речь идет об анализе не обычных данных, а динамики изменения во времени. В этом случае перед методами визуализации стоит очень важная и не простая задача наглядного и интуитивно понятного отображения развертки изменения некоторой величины по времени, отображения характера изменения, анализ и кооперация с принятыми решениями, эффект от этих решений. Для этого используют такие варианты визуализации, как графики, столбчатые диаграммы, гистограммы.

Но ситуация с визуализацией может сильно усложниться, если рассматриваемая величина является многокритериальной [3] или состоит из нескольких компонентов. В определённых случаях ЛПР бывает полезно и очень важно анализировать не изменение самой наблюдаемой величины, а влияние принятых мер на структуру составных частей, входящих в неё, динамику изменения структуры и вклада составляющих. Именно визуализации таких данных посвящена данная статья.

2. Визуальный анализ динамики изменения структурированных величин

2.1. Понятие структурированной величины

Для начала введем формализованное понятие структурированной величины. В рамках данной статьи будут рассматриваться вопросы визуализации динамики изменения именно таких данных.

Пусть имеется некоторая величина P, наблюдение за которой представляет интерес в рамках исследования. При этом пусть в состав этой величины входят ряд компонент k_i, таких что:

$$P = \sum_i k_i \tag{1}$$

Таким образом получается, что наблюдаемая величина P раскладывается по ряду компонент. При этом с точки зрения визуализации таких величин можно выделить два случая:

- величина P раскладывается на сравнительно небольшое количество компонент (менее 5);
- величина P состоит из большого числа составляющих.

Оба этих случая идентичны с точки зрения математического описания, но имеют значительные отличия при их визуализации. Если небольшое количество составляющих компонент человек способен наглядно представить и увидеть их структуру, то в случае большого числа – наглядное представление затрудняется и визуальный анализ становится более сложным.

Описанную выше величину назовем структурированной. При этом динамикой изменения структурированной величины будем называть такую зависимость, при которой в каждый конкретный момент t_j величина $P(t_j)$ будет определяться как сумма составляющих компонентов $k_i(t_j)$:

$$P(t_j) = \sum_i k_i(t_j) \quad (2)$$

Принципам визуализации такой динамики и таких величин посвящена данная статья.

2.2. Практическая необходимость анализа динамики изменения структурированных величин

Рассмотрим практические ситуации, в которых визуальная аналитика динамики структурированных данных может быть очень полезной.

В период карантинных мер первой половины 2020 года в Брянском государственном техническом университете группа заинтересованных студентов и преподавателей занималась моделированием распространения COVID-19 в среде AnyLogic. При этом ряд моделируемых и наблюдаемых величин как раз являлись составными и имели характер, описанный выше.

Одним из самых важных и критических показателей распространения COVID-19 является количество заразившихся людей – *infected* (обозначим её как *INF*). Именно от этого показателя в первую очередь зависят карантинные меры, сложность борьбы с эпидемией в том или ином регионе, различные управленческие решения [4,5]. Для отображения и визуальной аналитики динамики изменения количества заболевших во времени можно использовать классический метод – отображения кривой в виде графика функции.

Однако, как показала практика, цифра суммарной величины заболевших в том или ином регионе (или даже стране) не всегда адекватно описывают ситуацию. Даже если эту величину нормировать на численность региона (сделать относительной), этого явно будет недостаточно для анализа. Дело все в том, что сама по себе величина заразившихся людей *INF* по своей природе является составной и входящие в её состав части имеют разную степень критичности и важности. Например, величину *INF* можно разложить на такие составляющие, как количество тяжелых больных (требующих подключения к аппарату искусственной вентиляции легких), количество больных средней степени тяжести, количество больных, не требующих госпитализации. Поэтому одно и то же значение *INF* в разных регионах может абсолютно по-разному показывать общую ситуацию. Для правильного понимания ситуации важно видеть не только сам показатель, но и структуру входящих в него компонентов, оценивать вклад каждой составляющей.

Если количество компонентов не очень большое, для отображения и визуального анализа можно использовать традиционные схемы, такие как несколько графиков (по каждому критерию по отдельности), столбчатые диаграммы, гистограммы. Кроме того, в последние годы исследовательское сообщество накопило неопровержимые доказательства в пользу сложных и разнородных схем визуализации данных, основанных на визуализации Temporal Networks [6]. Однако ситуация усложняется, если количество составных частей увеличивается. В этом случае наглядность от используемых подходов сильно ухудшается и в данном направлении ведутся попытки разработки новых методов и подходов [7].

Рассмотрим еще один пример. По данным Всемирной организации здравоохранения (ВОЗ) заболеваемость и степень тяжести протекания болезни при COVID-19 находится во взаимосвязи с возрастом пациента. Поэтому полезным может быть анализ количества заболевших (введенный ранее показатель *INF*) в разрезе возраста заболевших. Т.е. в качестве составных частей этого показателя рассматривать число заболевших какого возраста отразились на величине общего числа заболевших. Данное отображение может быть полезно для оценки принятых мер и решений. Например, некоторые ограничительные меры могут не привести к уменьшению общей численности заболевших, но

могут изменить структуру этой величины, уменьшив число заболевших пожилого возраста, что так же является очень важным фактором в борьбе с эпидемией [8].

При этом визуальный анализ по возрастам является очень трудоемким. Сложно оценить данный показатель ввиду большого числа вариантов возрастов. Поэтому в большинстве случаев прибегают к сокращению вариантов (составных частей) путем округления возраста по диапазонам [9]. Пример такого округления, взятый из официальной статистики, показывает зависимость вероятности заражения и смертности в зависимости от возраста пациента показан в Таблице 1. Данные основаны на официальном документе от 28 февраля 2020 года из доклада ВОЗ WHO-China Joint Mission [10].

Табл. 1. Данные по зависимости заболевания от возраста ВОЗ WHO-China Joint Mission

Возраст	**Количество зараженных (человек)**	**Количество умерших (человек)**	**Вероятность смерти от COVID-19**
0-9 лет	416	-	-
10-19 лет	549	7	0,2%
20-29 лет	3619	1	0,2%
30-39 лет	7600	18	0,2%
40-49 лет	8571	38	0,4%
50-59 лет	10008	130	1,3%
60-69 лет	8583	309	3,6%
70-79 лет	3918	312	8,0%
80+ лет	1408	208	14,8%

В данном примере возраст пациентов разбит группами по 10 лет. Однако такое округление сопряжено с определёнными проблемами. Во-первых, неудачно выбранные диапазоны возрастов могут привести к некорректному анализу и сильному искажению реальной картины происходящего. Во-вторых, производя округление по диапазонам возрастов, мы теряем часть информации для анализа, которая может быть очень существенной. В-третьих, выбранные диапазоны могут сильно разниться для разных регионов и стран, а делать подбор в каждой конкретный ситуации является задачей весьма трудоемкой [11]. Поэтому важным моментом является не отказ от полной линейки возрастов, а разработка новых принципов визуализации подобных данных, при которых можно будет учитывать все данные в первоначальном их виде, не зависимо от того, сколько в них градаций и составных частей.

2.3. Существующие подходы к отображению динамики изменения данных

Существует достаточно много способов отображения динамики изменения данных. Рассмотрим наиболее часто используемые варианты на примере данных динамики развития эпидемии COVID-19 на территории Брянской области. На графиках представлена динамика численности заражений, выздоровлений, смертей от коронавирусной инфекции COVID-19, динамика численности новых заражений, по сравнению с предыдущим днем, выявленных на территории Брянской области, по дням начиная с первого выявленного больного.

Самым часто используемым вариантом для отображения таких данных является использование традиционных графиков. При этом время откладывается по оси абсцисс, а на оси ординат – рассматриваемая величина (см. рис. 1).

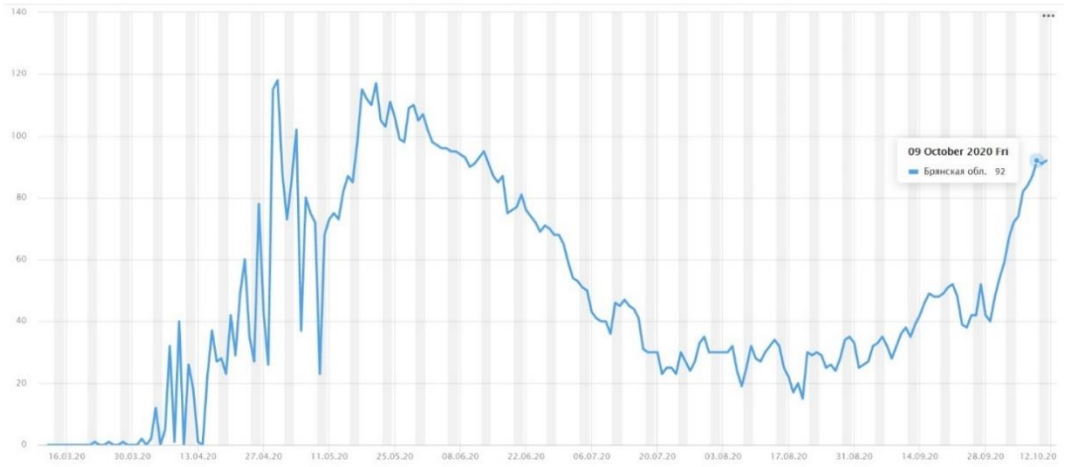

Рис. 1. График изменения количества заболевших COVID-19 в день

Если данные являются структурированными, то строится несколько самостоятельных графиков или диаграмм на одних координатных осях по каждой из составляющих. Пример таких графиков показан на рис. 2.

Рис. 2. Количество заражений, выздоравливании, смертей от COVID-19

Однако, такие графики хорошо анализировать в том случае, если наблюдаемых величин достаточно мало (две-три). Если же таких величин становится много, график становится трудно читаем и делать какие-то выводы по нему становится совсем сложно. Еще одной проблемой является сложность оценки вклада каждой составляющей в общую величину. На основе таких способов отображения зачастую становится невозможно в совокупности проанализировать ситуацию роста эпидемии с точки зрения возраста пациента, пола, сопутствующих заболеваний, его активности и т.п. [12].

Другим вариантом, который также часто встречается в открытых источниках статистики, является столбчатая диаграмма. В этом случае каждая составляющая отображается в виде отдельного столбика своего цвета, а динамика – в виде набора таких столбиков. По данным Яндекса, Apple и Otonomo приведен пример столбчатой гистограммы с группировкой изменения уровня активности населения России в период с февраля по июнь (см. рис.3).

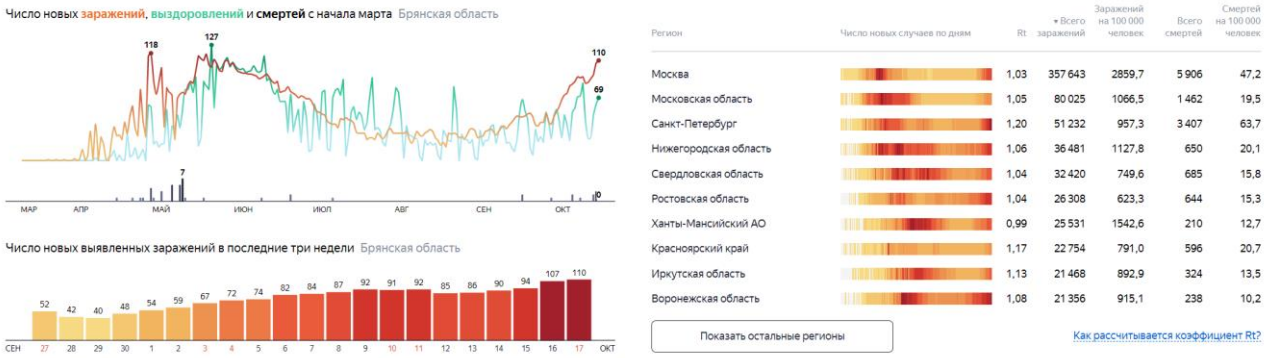

Рис. 3. Графики отображения динамики COVID-19 по Брянской области

Еще одним достаточно хорошим вариантом отображения является лепестковая диаграмма и круговая диаграмма [13]. В отличии от предыдущих вариантов эти способы очень хорошо визуализируют структуру данных, наглядно показывают вклад каждой составляющей. Однако их очень сложно применить для отображения динамики изменения этих данных во времени.

3. Система моделирования распространения COVID-19 и проблемы визуализации динамики структурированных величин в ней

Вспышки эпидемии в различных областях региона представляют собой особую ситуацию с высоким уровнем неопределенности. В некоторых регионах была отмечена заметная повторяемость ситуация по одному и тому же сценарию. В других – ситуация значительно отличалась. В Брянском государственном техническом университете был разработан программный комплекс в имитационной среде моделирования AnyLogic [14]. Авторами было предложено использовать несколько направлений моделирования. В основе смешанной имитационной модели, использовалось сочетание подходов: дискретно-событийного моделирования, агентного моделирования и раздела системной динамики. В качестве основного метода было предложено использование классических подходов моделирования эпидемии, основанной на SEIRD-модели [15]. Модель распространения заболевания SEIRD, относится к классу так называемых компартментальных моделей, суть которых состоит в том, чтобы разделить численность популяции на несколько групп (англ. compartments), т.е.: S (англ. susceptible) – число людей восприимчивых к заболеванию, E (англ. exposed) – группа инфицированных людей, находящихся в стадии инкубационного периода, I (англ. infectious) – инфицированные, R (англ. recovered) – число людей, выздоровевших, после перенесенной инфекции, D (англ. dead) – умершие. Каждый из заданных параметров образует переменные, входящие в состав системы дифференциальных уравнений, решая которую, можно спрогнозировать динамику развития эпидемии.

В результате анализа серии экспериментов, проведенных с моделью развития заболевания COVID-19, возможно наглядно представить скорость распространения инфекции, при различном поведении людей с различными параметрами, условиями и ограничениями. Скриншот системы показан на рис. 4. Система была представлена на конкурсе на лучшую научную работу «Современные научные достижения. Брянск – 2020».

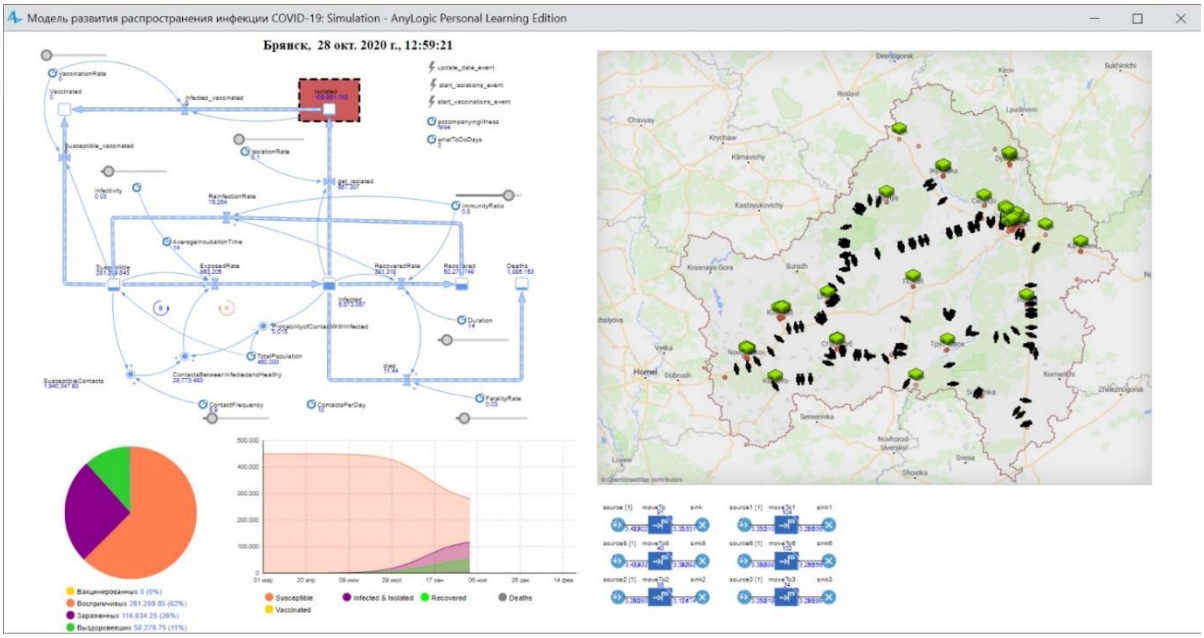

Рис. 4. Система моделирования COVID-19 в Брянской области

Взяв за основу известные критерии, возможно наглядно показать развитие скорости распространения заражения людей, показателей уровня сложности протекания заболевания [16]. В работе использовались модель, в которую были заложены основные факторы, известные на тот момент о распространении инфекции среди людей. Конкретные показатели и коэффициенты были рассчитаны по аналогичным данным по Московскому региону [17] и заложены в комплекс для моделирования ситуации в Брянской области. В итоге была построена модель (см. рис. 5), учитывающая следующие факторы эпидемии COVID-19:

1. Иммунитет. Чем выше у человека иммунитет, тем вероятность заражения COVID-19 ниже (параметр *ImmunityRatio* типа *double*, принимающий значение от 0 до 1). Каждому объекту *Person* (модель человека) в конструкторе присваивается случайное значение в указанном выше диапазоне.
2. Социальная ответственность (параметр *IsolationRate* типа *double* имеющий значение по умолчанию 0.45) [18]. По данным оперативного штаба РФ степень социальной ответственности составляет примерно 45% (в точности соблюдение предписания и рекомендации врача, а также карантинные или самоизоляционные меры). Значение данного параметра динамически изменяется на основании данных, полученных от Роспотребнадзора и Университета Джонса Хопкинса.
3. Параметр *ContactsPerDay*, определяет среднее количество близких контактов объекта *Person* за один день.
4. Вероятность контактов между людьми с различной степенью протекания инфекции: вероятность контактов между инфицированными людьми и здоровыми (параметр *ProbabilityofContactWithInfected*), частота контакта и соблюдение социальной дистанции (параметр *ContactFrequency*).
5. Сопутствующие коронавирусу заболевания (фактор обращения человека в больницу в случае заражения, параметр *accompanyingIllness* типа *boolean*).
6. Время на принятие решения после появления симптомов инфекции – фактор, на основании которого можно сделать вывод о степени протекания заболевания (параметр *whatToDoDays* типа *int* и задается случайным образом из вышеуказанного диапазона для каждого объекта *Person*). По данным статистики в регионе [19], человек обращается в больницу или отправляется на самоизоляцию в течение 1-5 дней. В данном временном диапазоне человек начинает замечать активное проявление симптомов и ухудшение своего состояния.

7. Принудительная самоизоляция людей (событие *start_isolations_event*, при условии превышения порога заболевших людей, а также на основании данных в периоде самоизоляции в Брянской области и распоряжений властей *intervention*).

8. Вероятность распространения вируса – параметр *InfectionProbability*. Значение вероятности зависит от нескольких факторов, например, от таких факторов, как использование средств индивидуальной защиты, ношение масок, мытьё рук или прикосновение к органам дыхания и слизистой человека.

Исходя из поставленных задач и факторов, влияющих на развитие распространения инфекции Covid-19, построение модели осуществляется через систему уравнений, описываемую ниже:

$$ExposedRate = Susceptible \cdot ContactsBetweenInfectedandHealthy \cdot Infectivity \quad (3)$$

$$InfectionRate = \frac{Exposed}{AverageIncubationTime} \quad (4)$$

$$RecoveredRate = \frac{Infected}{Duration \cdot ImmunityRatio} \quad (5)$$

$$Died = RecoveredRate \cdot FatalityRate \quad (6)$$

$$ReInfectionRate = RecoveredRate \cdot (1 - ImmunityRatio) \quad (7)$$

где *ExposedRate* – число заболевших людей за единицу модельного времени, *Susceptible* – число людей восприимчивых к заражению Covid-19 (учет факторов, таких как, ослабленный иммунитет, нахождение в общественных местах массового скопления людей и т.п.), *ContactsBetweenInfectedandHealthy* – среднее число контактов между инфицированными и здоровыми людьми, *Infectivity* – вероятность заражения при распространении эпидемии, *InfectionRate* – скорость протекания заболевания, *AverageIncubationTime* – усредненное значение инкубационного периода, *RecoveredRate* - число людей, определяющих восстановление после болезни (отсутствие в ходе тестирования наличия вируса в организме) за единицу модельного времени, *Infected* – инфицированные, *Duration* – среднее время протекания болезни, *FatalityRate* – количество умерших людей в единицу модельного времени, *Died* – процент смертности (вероятность летального исхода на основе статистики пациента).

Структура имитационной модели протекания заболевания согласно единице модельного времени, равной одному дню представлена на рис. 5.

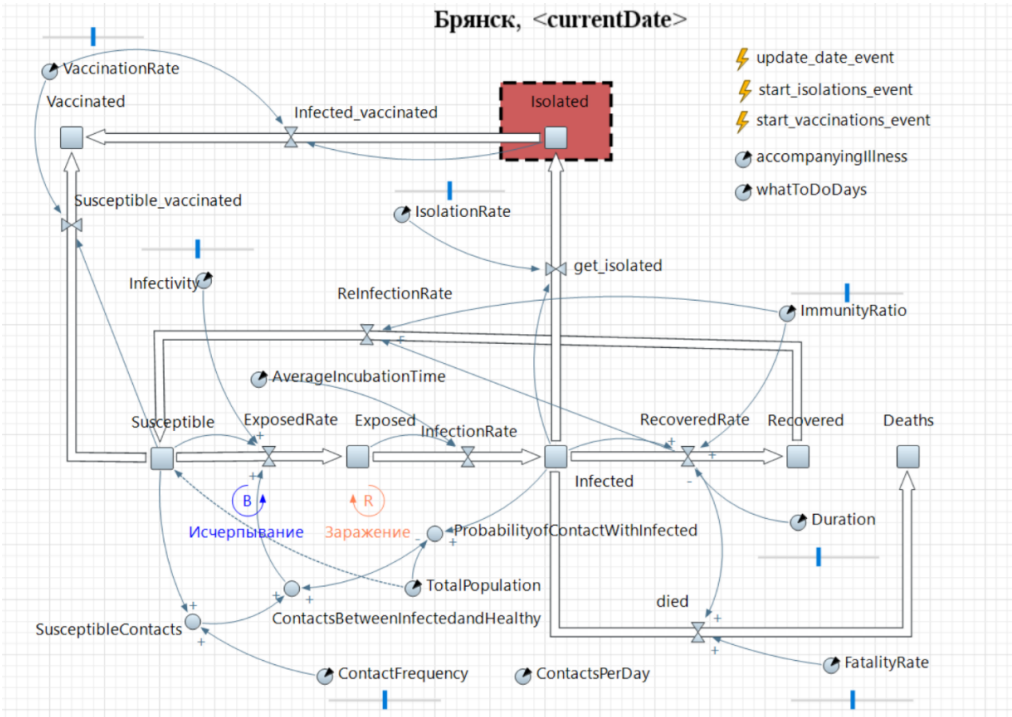

Рис. 5. Модель системной динамики распространения Covid-19

Для отображения и визуализации данных о динамике изменения структурированных величин использовался описанный выше подход на основе цветовых маркеров. Это позволило более наглядно представить полученные результаты.

Полученные при моделировании с помощью системы данные по распространению инфекции в Брянской области на основе открытой и общедоступной информации показывают жизнеспособность модели. Расхождение на текущей момент смоделированных и реальных кривых не превосходит 8%.

4. Предлагаемое решение для визуализации динамики изменения структурированных величин на основе цветовых маркеров

4.1. Использование цветовых маркеров для отображения структуры составной величины

Как было описано выше, в рамках исследований и моделирования ситуации с распространением COVID-19, проводимых на базе Брянского государственного технического университета, для отображения динамики изменения структурированных величин мы использовали подход, условно названный методом цветовых полос или графиком цветовых маркеров. Основная идея данного подхода заключается в следующем. Пусть у нас есть некоторая величина $P(t)$, значение которой меняется со временем (исследуемая величина зависит от параметра t). При этом данная величина состоит из разных компонент. Т.е. в момент времени t значение величины $P(t)$ можно разложить по некоторым составляющим:

$$P(t) = \sum_i k_i(t). \tag{3}$$

Для наглядного отображения подобных данных (как графика изменения самой величины, так и её составляющих), присвоим каждой составной части k_j некоторый цвет (так называемый цветовой маркер):

$$\{k_1 \leftrightarrow \text{цвет}_1;\ k_2 \leftrightarrow \text{цвет}_2;\ \dots;\ k_n \leftrightarrow \text{цвет}_n\}. \tag{4}$$

Затем отобразим кривую изменения величины $P(t)$ в виде стандартного графика, отложив по оси абсцисс время, а по оси ординат – саму наблюдаемую величину. При этом составляющие её части $k_j(t)$ закрасим соответствующим цветом-маркером. В этом случае график изменения $P(t)$ будет выглядеть в виде цветовых полос разной ширины (в зависимости от значения составляющих $k_j(t)$).

Рассмотрим пример, описанный в предыдущей главе (анализ изменения количества заболевших в разрезе тяжести протекания заболевания). Пусть в качестве анализируемой величины $P(t)$ выступает общее количество заразившихся COVID-19 в определенный момент времени (введенная ранее величина *INF*). При этом, как было описано выше, зачастую интерес представляет не график изменения количества заболевших, а анализ того, сколько среди этих заболевших тяжелых больных, сколько больных средней тяжести, сколько человек не требуют госпитализации и сколько бессимптомных. Это и будут составные части нашей величины. В этом случае данная величина в момент времени t раскладывается следующим образом:

$$INF(t) = \sum_{i=1}^{4} k_i(t). \tag{5}$$

где: $k_1(t)$ – количество тяжелых больных в момент времени t; $k_2(t)$ – количество госпитализированных больных средней тяжести; $k_3(t)$ – количество заболевших, не требующих госпитализации; $k_4(t)$ – количество бессимптомных больных.

Присвоим каждой величине некоторый цветовой маркер. Тогда общий график изменения эпидемиологической ситуации примет вид, показанный на рис. 6.

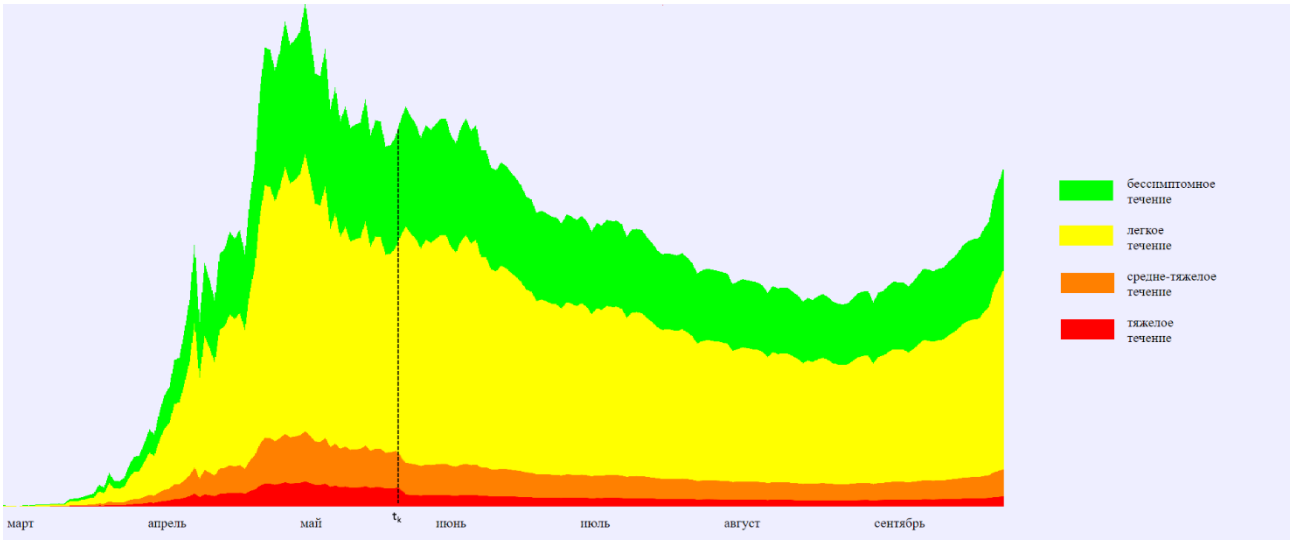

Рис. 6. График отображения структурированной величины в виде цветовых маркеров в разрезе тяжести протекания болезни

Этот график является более наглядным способом отображения, чем табличные данные или другие варианты, рассмотренные ранее. Данное отображение дает больше информации для анализа и дополнительных выводов, позволяет увидеть закономерности или эффект, не видимый при других вариантах отображения.

Отличительной особенностью отображения в виде цветовых маркеров является то, что данный метод позволяют ЛПР производить визуальный анализ, недоступный ранее при отображении в виде обычного графика. Например, на рис. 6 по представленному графику можно увидеть одну особенность. В момент времени t_k, отмеченном на графике пунктиром, были приняты некоторые меры, которые не позволили существенно уменьшить прирост общего количество заболевших и существенно повлиять на ситуацию с распространением COVID-19 (график величины *INF* сохранил свою природу). Однако в данный момент времени принятые решения позволили существенно изменить структуру величины *INF* в сторону уменьшения числа тяжелых больных (k_1) и больных средней тяжести (k_2), что является даже более важным моментом, чем общие тенденции целевого показателя. График на рис. 6 наглядно показывает эти изменения и дает возможность визуального анализа и поиска подобных периодов. Например, экспоненциальная форма кривой на графике развития эпидемии в обычных условиях. Такая форма определяет быстрый рост заболеваемости, что в свою очередь сильно увеличивает нагрузку на органы здравоохранения, тем самым увеличивая смертность, в связи с отсутствием должного охвата медицинской помощи и соответствующего оборудования. При уменьшении социальной активности человека в пять раз (самоизоляция) наблюдается плавное сглаживание кривой на графике. Такой вариант развития увеличивает общую продолжительность эпидемии, но уменьшает нагрузку на здравоохранение, тем самым уменьшая количество умерших пациентов.

В ряде ситуаций поиск и визуальный анализ таких структурных изменений для ЛПР является более важной задачей, чем даже анализ самой суммарной величины. Однако, в выше рассмотренном примере (рис. 6) есть ряд сложностей, которые затрудняют визуальный анализ структуры. В качестве таковых можно отметить следующие:

- в областях с малым значением целевого показателя *INF* полосы цветовых маркеров становятся достаточно узкими и какие-либо изменения в них становятся не видны;
- в областях с резким изменением целевой величины *INF* скачок целевого показателя затрудняет визуальный анализ входящих в него компонент k_i.

Для того что бы преодолеть указанные выше недостатки мы предлагаем перейти от абсолютных чисел к относительным, т.е. строить графики цветовых полос в нормированном варианте:

$$k_i'(t) = \frac{k_i(t)}{INF(t)}, \tag{6}$$

$$\sum_{i=1}^{n} k_i'(t) = 1. \tag{7}$$

При таком отображении колебания целевого показателя либо малая его величина не будут затруднять восприятие структуры. Анализ изменения структуры становится еще более наглядным и любые изменения или внешние влияния, которые смещают внутренний состав величин, проявляются нагляднее (см. рис. 7).

Рис. 7. График структурных изменений

4.2. Обобщение метода для непрерывных данных

Рассмотрим еще один пример динамики изменения структурированных данных. Допустим наблюдаемая величина $P(t)$ состоит из достаточно большого количества составляющих компонент k_i. Примером таких данных может служить рассмотренный выше анализ заболеваемости в разрезе возраста пациента. Как показала мировая практика с распространением COVID-19, данный параметр также очень важен для анализа ситуации и принятия управленческих решений.

Формально данная задача выразится следующим образом. Как и раньше в качестве наблюдаемой целевой величины будем использовать значение заболевших в момент времени t – $INF(t)$. Однако будем рассматривать её в разрезе составных частей $k_{age}(t)$ – количество заболевших в момент времени t в возрасте age, где age находится в диапазоне:

$$0 \le age \le MaxAge, \tag{8}$$

где $MaxAge$ – максимальный возраст заболевших за весь период наблюдения.

Наглядное отображение такой величины общепринятыми способами (графиками, диаграммами, таблицами) является достаточно проблематичной задачей в виду большого количества компонент (в предельном случае age может быть непрерывной величиной) [20,21]. Очень часто для анализа и интерпретации таких данных прибегают к округлению параметра и переход к некоторым диапазонам. Например, во всероссийском оперативном штабе параметр age для публикации предложили заменить на диапазоны:

- дети, подростки, молодые люди (до 30 лет);
- люди среднего возраста (30-49 лет);
- люди в возрасте 50-59 лет;
- пожилые люди старше 60 лет.

Однако такое деление весьма условно. Это подтверждается многочисленными публикациями, в которых диапазоны возрастов могут быть абсолютно другими. Это связано с тем, что выделить наиболее объективные диапазоны для анализа достаточно сложно. Прежде всего трудности вызывают следующие моменты:

- необходимость предварительного анализа данных и ввод дополнительных критериев, по которым будут сформированы диапазоны;
- потенциальная возможность изменения диапазонов во времени, которая в случае введения жестких границ приводит к недостоверности данных;
- потеря информативности и части информации за счет округления данных и приведение их к заранее заданным диапазонам;
- увеличение времени, необходимого для анализа данных за счет необходимости дополнительного шага (подбор границ диапазона);
- сложности в организации зависимых от внешних влияний диапазонов границ (например, границы диапазонов могут иметь географические или иные зависимости).

К тому же стоит заметить, что выделение заранее заданных диапазонов и округление наблюдаемой величины к ним автоматически приводит задачу к анализу в сугубо субъективных величинах и объективность результатов отображения будет сильно зависеть от того, на сколько удачно или неудачно были выбраны границы диапазонов.

Для анализа подобных данных мы предлагаем несколько другой подход. В связи со сложностями выбора диапазонов, мы предлагаем отказаться от них и визуализировать данные не в виде цветовых полос, а в виде непрерывного градиента. Для этого возрасту сопоставляется плавное изменение цвета и график целевой величины закрашивается внутри согласно полученным значениям составных компонент. Полученный график показан на рис. 4.

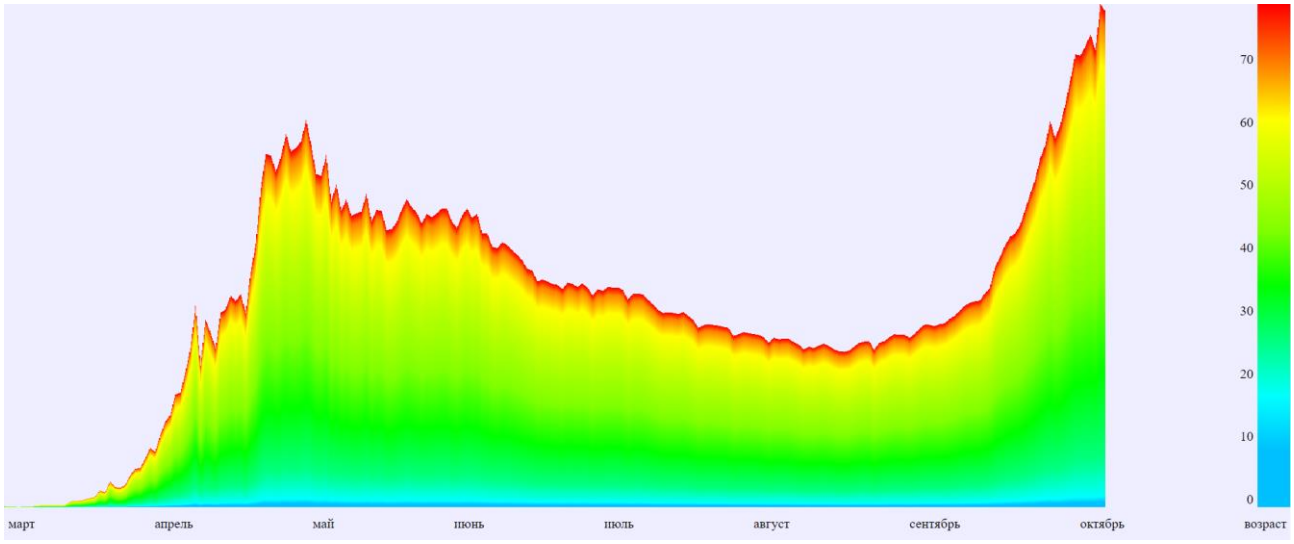

Рис. 8. График отображения структурированной величины в виде цветовых маркеров в разрезе возраста заболевших

Такое отображение не требует предварительного анализа и выбора границ округления. Данные можно визуализировать сразу же в необработанном виде. При этом закономерности по составу будут автоматически подстраиваться и адекватно отображаться вне зависимости ни от каких внешних или внутренних влияний.

5. Заключение

В данной статье был предложен метод, сочетающий анализ и визуализацию структурированных данных. Большинство примеров в статье приводилось на примере ситу-

ации с развитием COVID-19. Это связано с тем, что данный подход визуализации мы опробовали и протестировали на базе соответствующего комплекса, разработанного нами. Однако предложенный метод отображения динамики изменения структуры данных в виде цветовых маркеров и градиентов можно применять абсолютно в любых областях, где есть необходимость отображать динамику составных величин, видеть и анализировать структуру этих величин.

Список литературы

1. Podvesovskii A.G., Isaev R.A.: Constructing Optimal Visualization Metaphor of Fuzzy Cognitive Maps on the Basis of Formalized Cognitive Clarity Criteria // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 115-129. – DOI: 10.26583/sv.11.4.10
2. Zakharova A., Shklyar A. Basic principles of data visual models construction, by the example of interactive systems for 3D visualization // Scientific Visualization, 2014, Vol. 6, Num. 2, P. 62-73.
3. Bondarev A.E., Galaktionov V.A.: Generalized Computational Experiment and Visual Analysis of Multidimensional Data // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 102-114. – DOI: 10.26583/sv.11.4.09
4. Russell Timothy W., Hellewell J., Jarvis Christopher I., van Zandvoort K., Abbott S., Ratnayake R., Flasche S., Eggo R.M., Edmunds W.J., Kucharski A.J. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, 2020, https://doi.org/10.2807/1560-7917.ES.2020.25.12.2000256, last accessed 2020/07/10.
5. Yang L.Y., Yan L.M., Wan L., Xiang T.-X., Le A., Liu J.M., Peiris M., Poon L.L.M., Zhang W. Viral dynamics in mild and severe cases of Covid 19. Lancet Infect Dis, 2020.
6. Kapoor A., Ben X., Liu L., et al. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks, 2020.
7. Zakharova A.A., Korostelyov D.A., Fedonin O.N.: Visualization Algorithms for Multicriteria Alternatives Filtering // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 66-80. – DOI: 10.26583/sv.11.4.06
8. Murray C.J.L. Forecasting COVID 19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. IHME COVID 19 health service utilization forecasting team, https://doi.org/10.1101/2020.03.27.20043752, last accessed 2020/07/10.
9. Khrapov P.V., Loginova A.A. Mathematical modelling of the dynamics of the coronavirus COVID-19 epidemic development in China // International Journal of Open Information Technologies, 2020, Vol. 8(4), P. 13-16, http://www.injoit.org/index.php/j1/article/view/908/874, last accessed 2020/07/10.
10. Матвеев А.В. Математическое моделирование оценки эффективности мер против распространения эпидемии COVID-19 // Национальная безопасность и стратегическое планирование, 2020, №1(29), С. 23-39.
11. Kosara R. Presentation-oriented visualization techniques // IEEE Comput Grap Appl, 2016, Vol. 36, P. 80-85.
12. Кольцова Э.М., Куркина Е.С., Васецкий А.М. Математическое моделирование распространения эпидемии коронавируса COVID-19 в ряде европейских, азиатских стран, Израиле и России // Проблемы экономики и юридической практики, 2020, №2.
13. Wang D.Q., Guo D.H., Zhang H. Spatial temporal data visualization in emergency management: a view from data-driven decision // Proceedings of the 3rd ACM SIGSPATIAL Workshop on Emergency Management, 2017, P. 1-7.
14. AnyLogic, https://www.anylogic.ru/, last accessed 2020/10/15.

15. Jüni P., Rothenbühler M., Bobos P., Thorpe K.E., da Costa B., Fisman D., Slutsky A.S., Gesink D. Impact of climate and public health interventions on the COVID-19 pandemic: A prospective cohort study. CMAJ May 08, 2020.
16. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med, 2020, Vol. 8(5), P. 475-481.
17. Борисов В.В., Круглов В.В., Федулов А.С. Нечеткие модели и сети. – М.: Горячая линия – Телеком, 2012. – 284 с.
18. Родкин М.В., Шихова Н.М. Математическое моделирование развития эпидемии COVID-19, попытка прогноза // Уральский геологический журнал, 2020, №3.
19. World Health Organization. Coronavirus disease 2019 (COVID-19): Situation Report – 38 from 27 February 2020, http://www.who.int/docs/default-source/coronaviruse/ situation-reports/20200227-sitrep-38-covid-19. pdf, last accessed 2020/07/10.
20. Dimara E., Perin C. What is interaction for data visualization? // IEEE Trans Visual Comput Graph, 2020, Vol. 26, P. 119-129.
21. Robertson G, Fernandez R, Fisher D, et al. Effectiveness of animation in trend visualization // IEEE Trans Visual Comput Graph, 2008, Vol. 14, P. 1325-1332.

Scientific Visualization, 2020, volume 12, number 4, pages 85 - 97, DOI: 10.26583/sv.12.4.08

Visual Analysis of Dynamic Changes in Structured Data on the Basis of Colour Markers

A.A. Trubakova[1], A.O. Trubakov[2]

Bryansk State Technical University, Bryansk, Russia

[1] ORCID: 0000-0003-0280-1760, trubakovaaa@gmail.com
[2] ORCID: 0000-0003-0058-1215, trubakovao@gmail.com

Abstract

One of the most important tasks for decision-making systems is obvious and intuitive visualization of input data. If they are displayed correctly and conveniently, this tool can become a very serious assistant for a decision-maker. Such a tool influences not only on the complexity of the decision-making process, but also the correctness and objectivity of the decisions made. Because of this so much attention is paid to this issue in various studies. However, nowadays the issue of displaying the dynamics of changes in structured data (data in which the observed value has an internal structure and consists of a large number of components) is not sufficiently developed. This paper focuses on these issues, provides a formal description of structured data, and suggests an approach to displaying the dynamics of their changes. The proposed approach to visualization allows to see both the general process of changing the observed value and the nature of changes in its internal structure. Separate charts give the opportunity to see the contribution of each component of a structured value and evaluate the dynamics of this contribution over time. The paper also shows the area of potential application of this approach, highlights its features and main prospects. Most of examples in the paper are based on the developed software package for modeling the situation with the spread of COVID-19 in Bryansk region, which used the proposed visualization option. There are also examples in which the proposed approach to visualization helps to get new information that is not visible when using other approaches to displaying structured data.

Keywords: Data Visualization, Structured Data, Making Decisions, SEIRD-model.

References

1. Podvesovskii A.G., Isaev R.A.: Constructing Optimal Visualization Metaphor of Fuzzy Cognitive Maps on the Basis of Formalized Cognitive Clarity Criteria // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 115-129. – DOI: 10.26583/sv.11.4.10
2. Zakharova A., Shklyar A. Basic principles of data visual models construction, by the example of interactive systems for 3D visualization // Scientific Visualization, 2014, Vol. 6, Num. 2, P. 62-73.
3. Bondarev A.E., Galaktionov V.A.: Generalized Computational Experiment and Visual Analysis of Multidimensional Data // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 102-114. – DOI: 10.26583/sv.11.4.09
4. Russell Timothy W., Hellewell J., Jarvis Christopher I., van Zandvoort K., Abbott S., Ratnayake R., Flasche S., Eggo R.M., Edmunds W.J., Kucharski A.J. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, 2020, https://doi.org/10.2807/1560-7917.ES.2020.25.12.2000256, last accessed 2020/07/10.
5. Yang L.Y., Yan L.M., Wan L., Xiang T.-X., Le A., Liu J.M., Peiris M., Poon L.L.M., Zhang W. Viral dynamics in mild and severe cases of Covid 19. Lancet Infect Dis, 2020.

6. Kapoor A., Ben X., Liu L., et al. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks, 2020.
7. Zakharova A.A., Korostelyov D.A., Fedonin O.N.: Visualization Algorithms for Multi-criteria Alternatives Filtering // Scientific Visualization, 2019, Vol. 11, Num. 4, P. 66-80. – DOI: 10.26583/sv.11.4.06
8. Murray C.J.L. Forecasting COVID 19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. IHME COVID 19 health service utilization forecasting team, https://doi.org/10.1101/2020.03.27.20043752, last accessed 2020/07/10.
9. Khrapov P.V., Loginova A.A. Mathematical modelling of the dynamics of the coronavirus COVID-19 epidemic development in China // International Journal of Open Information Technologies, 2020, Vol. 8(4), P. 13-16, http://www.injoit.org/index.php/j1/article/view/908/874, last accessed 2020/07/10.
10. Matveev A.V. Mathematical modeling of evaluating the effectiveness of measures against the spread of COVID-19. // Natsionalnaya Bezopasnost I Strategicheskoe Planirovanie (National Security and Strategic Planning), 2020, Vol. 1(29), P. 23-39.
11. Kosara R. Presentation-oriented visualization techniques // IEEE Comput Grap Appl, 2016, Vol. 36, P. 80-85.
12. Koltsova E.M., Kurkina E.S., Vasetsky A.M. Mathematical modeling of the spread of COVID-19 coronovirus epidemic in a number of European, Asian countries, Israel and Russia // Economic problems and legal practice, 2020, Vol. 2
13. Wang D.Q., Guo D.H., Zhang H. Spatial temporal data visualization in emergency management: a view from data-driven decision // Proceedings of the 3rd ACM SIGSPATIAL Workshop on Emergency Management, 2017, P. 1-7.
14. AnyLogic, https://www.anylogic.ru/, last accessed 2020/10/15.
15. Jüni P., Rothenbühler M., Bobos P., Thorpe K.E., da Costa B., Fisman D., Slutsky A.S., Gesink D. Impact of climate and public health interventions on the COVID-19 pandemic: A prospective cohort study. CMAJ May 08, 2020.
16. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med, 2020, Vol. 8(5), P. 475-481.
17. Borisov V.V., Kruglov V.V., Fedulov A.S. Fuzzy models and networks. – M.: Hot line – Telecom, 2012, P. 284.
18. Rodkin M.V., Shikhova N.M. Mathematical modeling of COVID-19 epidemic, an attempt to forecast // Uralian Geological Journal, 2020, Vol. 3.
19. World Health Organization. Coronavirus disease 2019 (COVID-19): Situation Report – 38 from 27 February 2020, http://www.who.int/docs/default-source/coronaviruse/ situation-reports/20200227-sitrep-38-covid-19. pdf, last accessed 2020/07/10.
20. Dimara E., Perin C. What is interaction for data visualization? // IEEE Trans Visual Comput Graph, 2020, Vol. 26, P. 119-129.
21. Robertson G, Fernandez R, Fisher D, et al. Effectiveness of animation in trend visualization // IEEE Trans Visual Comput Graph, 2008, Vol. 14, P. 1325-1332.