

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ЭКОНОМИКИ И ОРГАНИЗАЦИИ
ПРОМЫШЛЕННОГО ПРОИЗВОДСТВА
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

А. В. Костин

**БАЗА ЗНАНИЙ КАК ИНСТРУМЕНТАРИЙ
РЕШЕНИЯ ЗАДАЧ В ЭКОНОМИЧЕСКОЙ,
СОЦИАЛЬНОЙ И ПРОИЗВОДСТВЕННОЙ СФЕРАХ**

Ответственный редактор
М.А. Ягольницер

Новосибирск
2023

УДК 338.984

ББК 65.050.9(2)2

К 723

DOI 10.36264/978-5-89665-384-4-2023-020-160

Рецензенты:

д.э.н. Коломак Е.А., д.э.н. Ибрагимов Н.М., д.э.н. Дубина И.Н.

К 723

Костин А.В. База знаний как инструментарий решения задач в экономической, социальной и производственной сферах / под ред. М.А. Ягольничера. – Новосибирск: Изд-во ИЭОПП СО РАН, 2023. – 160 с.

ISBN 978-5-89665-384-4

В монографии представлены детальные результаты работ ИЭОПП СО РАН по базовому проекту плана НИР ИЭОПП СО РАН: № 121040100283-2. Одновременно работа рассматривается как второе издание и развитие другой «Новый импульс Азиатской России», изданной в 2022 г. при поддержке крупного научного проекта по приоритетным направлениям научно-технологического развития: «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий». Содержание данной монографии представляет интерес для широкого круга исследователей в области экономики и архитектуры Баз Знаний, магистрантов и аспирантов, работников органов власти и управления, чья деятельность связана с принятием решений в области политики развития федерального и регионального уровней.

УДК 338.984

ББК 65.050.9(2)2

ISBN 978-5-89665-384-4

© ИЭОПП СО РАН, 2023.

© Костин А.В., 2023.

INSTITUTE OF ECONOMICS AND INDUSTRIAL ENGINEERING
SIBERIAN BRANCH OF RUSSIAN ACADEMY OF SCIENCES

A.V. Kostin

**KNOWLEDGE BASE AS A TOOL
FOR SOLVING PROBLEMS IN THE ECONOMIC,
SOCIAL AND INDUSTRIAL SPHERES**

Executive editor
M.A. Yagolnitser

Novosibirsk
2023

Kostin A.V. Knowledge base as a tool for solving problems in the economic, social and industrial spheres / ed. M.A. Yagolnitser – Novosibirsk: IEIE SB RAS. – 2023. – 160 c.

The presented monograph provides a detailed description of the concept of a knowledge base aimed at scientific support of the processes of making strategic decisions to create conditions for the progressive and sustainable socio-economic development of the territories of Asian Russia. The work describes an information system that represents an interconnected set of tools, methods and models used for storing, processing and issuing information in the interests of achieving the goal - developing scenarios for the development of socio-economic systems of Asian Russia based on the interaction of natural resource, technological and human potentials, ensuring institutional, investment and infrastructural harmony

ОГЛАВЛЕНИЕ

Словарь терминов	7
Введение	10
Глава 1. ОБЩАЯ СТРУКТУРА БАЗЫ ЗНАНИЙ	15
1.1. DSGE-модели для краткосрочного прогнозирования экономики России	18
1.1.1. Задачи базы данных	18
1.1.2. Техническое задание функциональных возмож- ностей базы данных	19
1.2. Общая схема работы базы знаний	23
1.3. Принципы работы с данными в рамках базы дан- ных	27
Глава 2. ОСНОВНЫЕ КОМПОНЕНТЫ БАЗЫ ДАННЫХ	30
2.1. Блок информационного обеспечения (компании, сек- торы, комплексы, регионы, кластеры)	31
2.1.1. Социально-экономические статистические дан- ные	31
2.1.2. Данные о Компаниях	33
2.1.3. Анализ достоверности данных Компаний	37
2.1.4. Таможенная статистика	40
2.2. Блок инвестиционных, программных, институцио- нальных проектов	41
2.3. Блок по углеводородному сырью, отдельным видам твердых полезных ископаемых и подземным водам	42
2.4. Блок технологий	47
2.5. Оптимизация и ускорение работы с данными	50
2.6. Автоматизация сбора данных	51
Глава 3. ФОРМИРОВАНИЕ МОДЕЛЬНОГО АППАРАТА БАЗЫ ЗНАНИЙ	53
3.1. Модели КАМИН и ОМВЕАР	53
3.1.1. Общее техническое описание модели КАМИН	54
3.1.2. Общее техническое описание модели ОМВЕАР	57
3.2. Когнитивные модели	59
3.3. Система анализа и прогнозирования межрегиональных и межотраслевых взаимодействий на основе анализа больших данных	60

Глава 4. ИНТЕРФЕЙС ПОЛЬЗОВАТЕЛЯ БАЗЫ ЗНАНИЙ	71
4.1. Техническое задание к интерфейсу пользователя базы знаний ИЭОПП СО РАН	71
4.1.1. Функциональные части системы	71
4.1.2. Каталог данных	71
4.1.3. Поиск и фильтрация	72
4.1.4. Операция с данными	73
4.1.5. Экспорт	73
4.1.6. Визуализация	73
4.1.7. Аналитические запросы	73
4.1.8. Моделирование	74
4.2. Веб-интерфейс базы знаний	75
4.2.1. ГИС-моделирование	78
4.2.2. Модели	82
4.2.3. Работа с данными	100
4.3. ГИС-интерфейс базы знаний	111
Глава 5. Примеры использования платформенных решений	122
5.1. Оценка инвестиционного импульса с использованием когнитивной модели базы знаний	122
5.2. Влияние цифровизации городской среды на высокотехнологичные компании	133
5.3. Применение базы знаний для комплексной оценки возможностей реализации цепочек инвестиционных проектов на примере развития черной металлургии в Азиатской России	134
5.4. Оценки инвестиционного импульса с использованием прогностического комплекса ДММ-КАМИН	138
5.5. Предварительный анализ крупных инфраструктурных инвестиционных проектов	141
Заключение	151
Список использованных источников	154
Благодарности	159

Словарь терминов

Apache Airflow – открыто распространяемый инструмент, который позволяет разрабатывать, планировать и осуществлять мониторинг сложных рабочих процессов.

BNN (Bayesian Neural Network) – Байесовская нейронная сеть. Графовая вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей по Байесу.

ETL (Extract, Transform, Load) – извлечение, преобразование, загрузка. Один из основных процессов в управлении хранилищами данных.

LSTM (Long short-term memory) – рекуррентная нейронная сеть с долгой краткосрочной памятью. особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долгосрочным зависимостям.

MLP (Multilayered perceptron) – многослойный линейный перцептрон. Это класс искусственных нейронных сетей прямого распространения, состоящих как минимум из трех слоёв: входного, скрытого и выходного.

PCA (Principal Component Analysis) – метод главных компонент.

RBF (radial basis function) – сеть радиальных базисных функций. Программная система, способная классифицировать данные и делать предсказания.

REST/SOAP API – это способ взаимодействия сайтов и веб-приложений с сервером. Его также называют RESTful.

Термин состоит из двух аббревиатур, которые расшифровываются следующим образом.

API (Application Programming Interface) – это код, который позволяет двум приложениям обмениваться данными с сервера. На русском языке его принято называть программным интерфейсом приложения.

REST (Representational State Transfer) – это способ создания API с помощью протокола HTTP. На русском его называют «передачей состояния представления».

HTTP (HyperText Transfer Protocol) – «протокол передачи гипертекста») – протокол прикладного уровня передачи данных, изначально – в виде гипертекстовых документов в формате HTML, в настоящее время используется для передачи произвольных данных.

Shiny – программная среда, предназначенная для удобства создания на основе языка программирования R интерактивных веб-приложений с графическим интерфейсом пользователя.

SQL (Structured Query Language) – декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционной базе данных, управляемой соответствующей системой управления базами данных.

UI (User Interface) – дизайн пользовательского интерфейса.

URL (Uniform Resource Locator) – адрес ресурса в сети Интернет.

БД – база данных.

БЗ – База знаний.

Веб-интерфейс – веб-страница или совокупность веб-страниц, предоставляющая пользовательский интерфейс для взаимодействия с сервисом или устройством посредством протокола HTTP и веб-браузера.

ГИС – геоинформационная система.

ГМК – горно-металлургическая компания (горно-металлургический комбинат).

ГОК – горно-обогатительный комбинат.

Градиентный бустинг (Gradient boosting) – техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений.

ГРЭС – районная электростанция государственного образца.

ГЭС – гидроэлектростанция.

ЕМ-алгоритм (Expectation-maximization (EM) algorithm) – алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных.

ЗИФ – золотоизвлекательная фабрика.

КАМИН – Комплексный анализ межотраслевой информации.

КОМБИ – Комплекс обработки матричных блоков информации.

Лог – это текстовый файл, куда автоматически записывается важная информация о работе системы или программы.

ОКВЭД – Общероссийский классификатор видов экономической деятельности. Коды ОКВЭД в обязательном порядке необходимо указывать при регистрации юридического лица (ООО, АО, НКО и др.), а также ИП (индивидуальных предпринимателей).

ОМВЕАР – оптимизационная модель взаимодействия Европейской и Азиатской России.

ОФ – Обоганительная фабрика.

Скрипт – это набор команд, прописанных в коде, которые необходимы для выполнения задачи.

СУБД – Система управления базами данных.

ТПИ – твердые полезные ископаемые.

ТЭЦ – теплоэлектроцентраль.

УВС – углеводородное сырье.

ФО – федеральный округ.

Введение

Постоянное развитие современных технологий существенно увеличивает возможности исследователя. Прогресс цифровых технологий и методов анализа больших данных позволяют повышать точность и скорость математических вычислений. Благодаря развитию машинного обучения и искусственного интеллекта стало возможным автоматизировать и оптимизировать множество задач, которые ранее требовали больших затрат времени и ресурсов. В онлайн пространстве появляются разнообразные сервисы, предоставляющие большие объемы социально-экономических данных и возможности их обработки на удаленных серверах.

За последние десятилетия меняется концепция исследовательской работы. В соответствии с требованиями и возможностями современного научно-технологического развития применение цифровых технологий становится одним из важных условий обеспечения конкурентоспособности в научной сфере.

Подобные изменения повышают требования и к самому исследователю. Растёт необходимость в практических навыках эконометрии и математического моделирования. Постоянное снижение барьеров изучения языков программирования, их упрощение, а также конкуренция в научной среде, стимулирует ученого к погружению в современный инструментарий, особенно когда теоретическая модель для подтверждения выдвинутых гипотез предполагает использование больших данных. Все чаще исследователи используют методы ГИС-моделирования для пространственного анализа. Ускорение накопления данных и знаний требует изучения инструментов их сбора, хранения и структурирования.

Ссылки на эмпирические исследования OECD, Д. Форей и Б. Лундвалл показали, что современная экономика все больше зависит от накопленных знаний [Foray, Lundvall, 1996]. Этот процесс в последнее время только усиливается, в том числе и в научной сфере. Знания могут быть важным стратегическим активом и необходимым элементом конкурентоспособности и успеха в высокодинамичных средах, поскольку они расширяют возможности, необходимые для решения проблем.

Всё вышеперечисленное стало причиной развития новых подходов к анализу и структурированию данных. В монографии описываются идеи и подходы к анализу данных, сформированные в рамках разработки Базы знаний Института экономики и организации промышленного производства Сибирского отделения Российской академии наук (ИЭОПП СО РАН). Создание Базы знаний происходило в рамках реализации проекта «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий»¹.

Цель построения базы знаний состоит в таком преобразовании исходных данных о ресурсах, объектах, технологиях, состоянии и развитии макрорегиона Азиатская Россия, которое на качественно новом уровне обеспечит понимание сложных процессов синергии в экономике на основе сетевых комбинаций цепочек объектов, взаимосвязанных по движению природных, финансовых, трудовых, информационных ресурсов, продукции, технологий, поддерживаемых институтами развития.

Анализ литературы показал, что среди достаточно большого потока статей и книг по базам знаний лишь в малой доле стоит вопрос её определения. Официальный международный словарь терминов ISO по ИТ-технологиям дает следующее определение: «База знаний – это база данных, содержащая правила вывода и информацию о человеческом опыте и знаниях в определенной области» [ISO/IEC/IEEE 24765:2017]. А.Ф. Тузовский, С.В. Чириков и В.З. Ямпольский рассматривали базу знаний как объединение неявных и явных знаний [Тузовский и др., 2005]. В.Л. Арлазаров и Н.Е. Емельянов в своей работе рассматривали базу знаний как систему работы с декларативными (факты) и процедурными (алгоритмы) знаниями через формы ввода/вывода и интерфейсы [Арлазаров, Емельянов, 2006]. Последнее определение наиболее

¹ Крупный научный проект «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий», соглашение № 075-15-2020-804 (№ 13.1902.21.0016) от 2 октября 2020 г.

полно описывает концепцию объединения баз данных и программных продуктов и взято за основу в разработке Базы знаний ИЭОПП СО РАН.

В основе всех определений Базы знаний лежит определение слова «знание». Определение Ф. Бэкона, ставшее крылатым – «Знание – сила!», впоследствии было уточнено Т. Гоббсом: «Знание есть только путь к силе». В философии «Знание есть такой результат познавательной деятельности, который обладает непреходящей истинностью, может быть логически или фактически обоснован и допускает эмпирическую или практическую проверку» [Ивин, 2004]. Одним из часто используемых определений знания является следующее «Знания – это комбинация данных и информации, к которой добавлены мнения, навыки и опыт экспертов, что дает в результате ценный актив, который может быть использован при принятии решений» [Тузовский, 2007]. О.О. Варламов в своей работе пишет, что наиболее близкое слово к «знаниям» в концепции Базы Знаний является «правило» – термин, который в целом более хорошо формализован [Варламов 2007].

Д. Форея, Б. Лундвалл и Б. Джонсон [Foray, Lundvall, 1996; Lundvall, Jonson 1994], предложили разделение знаний на 4 вида:

- Know-what,
- Know-why,
- Know-how,
- Know-who.

Know-what описывается авторами как знания о «фактах». Например, численность населения в Новосибирской области, динамика показателя ВРП и др. Это информация, которая просто структурируется и хранится на цифровых носителях. Она описывается в виде баз данных и входят в основу Базы знаний.

Know-why описываются как научные знания о принципах и фундаментальных законах. Это знания, которые являются базовыми, и такие знания чаще всего формируются у специалистов во время обучения в ВУЗе, ускоряют и упрощают получение новых знаний. Эти знания не сформировать в формате исследовательской Базы знаний, но они ложатся в основу всех её алгоритмов и методов.

Know-how больше относится именно к навыкам и знаниям, связанными с технологиями. В научной среде это связано с развитием исследовательского инструментария и компетенций. В Базе знаний они накапливаются в виде развивающегося инструментария, онтологий и базы данных о технологиях.

Know-who относится к социальным навыкам. Знание того, кто является специалистом в той или иной сфере, упрощает процедуры привлечения экспертов. В результате усложнения экономических процессов такой вид знаний становится все более важным. В Базе знаний он может быть представлен в виде сетевых графов и систем взаимосвязей.

В монографии А.Ф. Тузовского, С.В. Чирикова и В.З. Ямпольского знания разделяются на явные и неявные, оцифрованные знания называются «явными» [Тузовский и др., 2005]. Явные знания можно сохранить и передать, а неявные знания относятся к знаниям, которые не могут быть легко переданы, потому что не были сформулированы в явной форме. Развитие технологий обработки знаний приводит к изменению границ между явными и неявными знаниями, что не только не уменьшает важность неявных знаний, а наоборот делает навыки и компетенции, связанные с отбором и эффективным использованием информации еще более важными.

Л.А. Трофимова и В.В. Трофимов [Трофимова, Трофимов, 2012], вслед за И. Нонака и Х. Такеучи [Нонака, Такеучи, 2011], показывают систему перехода неявных знаний в явные и наоборот (рис.).



Рис. Спираль знаний Нонака и Такеучи

Источник: [Трофимова, Трофимов, 2012].

Социализация предполагает перенос неявных знаний от одного человека к другому при взаимодействии и обсуждении.

Экстернализация – это перевод неявных индивидуальных знаний в явные знания путем их формулирования и фиксации.

Комбинирование (формирование системного знания) означает перенос знаний путем накопления, структурирования и распространения существующих знаний.

Интернализация предполагает трансформацию человеком явных знаний в неявные, формирование нового прикладного знания путем обучения на практике. Это также процесс непрерывного индивидуального и коллективного размышления, а также способность видеть связи и распознавать закономерности, а также способность находить смысл между областями, идеями и концепциями.

В рамках развития концепции увеличения эффективности накопления научных знаний, предполагается, что развитие Базы знаний может ускорить все описанные процессы, кроме процесса Социализации. Он формируется в рамках научного сообщества. Развитие спирали знаний приводит к постоянному накоплению явных знаний и трансформации неявных знаний. В результате, концептуально База знаний не может являться фиксированным элементом и предполагает постоянное изменение вслед за изменениями структуры знаний организации. Только в этом случае будет достигаться высокий уровень эффективности её внедрения.

Подводя итог, можно описать Базу знаний как динамично развивающуюся систему работы со знаниями через формы ввода/вывода данных и программные интерфейсы, способствующую ускорению накопления знаний в организации.

Создание базы знаний предполагает проектирование её общей структуры, наполнения, накопления инструментария и разработку интерфейса. Эти этапы и примеры применения базы знаний будут описаны в последующих главах.

Глава 1

ОБЩАЯ СТРУКТУРА БАЗЫ ЗНАНИЙ

«Именно от структуры системы главным образом и зависят суждения о ее единстве и основательности»

Э. Кант. Критика чистого разума

В работе Шинган Хоу (с соавторами) говорится о том, что большинство исследователей в целом согласны с тем, что практически невозможно построить большую систему, основанную на знаниях, без структурированного подхода [Нои и др., 2023]. Полный жизненный цикл построения систем такого рода, как База знаний, можно представить в виде трехэтапного процесса: создание, эксплуатация и обслуживание [Felfernig, Wotawa, 2013].

Существует несколько этапов создания базы знаний, которые начинаются с определения требований пользователя и заканчиваются итоговой реализацией. В целом можно выделить следующие этапы разработки базы знаний:

1. Определение целей и задач базы знаний и построение концептуальной схемы.
2. Сбор и анализ информации, определенной на этапе формулирования целей и задач.
4. Разработка структуры базы знаний и базы данных с учетом анализа информации и концептуальной схемы.
4. Разработка пользовательского интерфейса базы знаний.
5. Тестирование и отладка моделей, методов и данных.
6. Внедрение, поддержка, масштабирование и развитие.

Следуя за концепцией спирали знаний предполагается, что после начальной реализации База знаний будет дорабатываться, расширяться и постоянно включать новые элементы, расширяя блоки явных и неявных знаний.

На первом этапе моделирования данных пользовательские требования синтезируются в концептуальную схему, которая позже преобразуется в конкретные модели хранения знаний и данных, например, реляционную модель [Antony и др., 2005]. Примеров концептуальных схем много, но в общем виде она предложена в работе Родионовой (рис. 1.1) [Родионова, 2017].

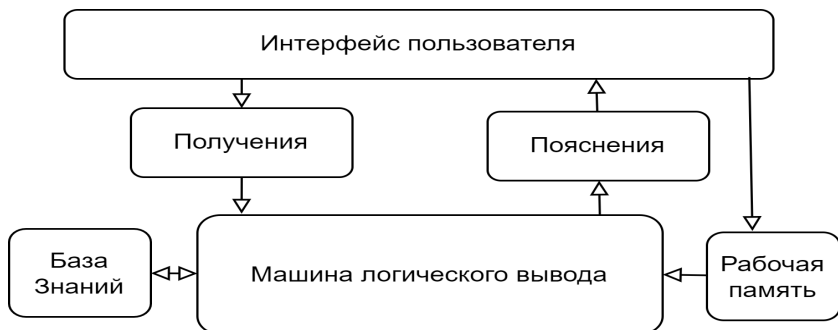


Рис. 1.1. Общая структура экспертной системы, основанной на правилах

Источник: [Родионова, 2017]

Предложенный подход помогает раскрыть примерную концепцию построения БЗ, но для её реализации необходима более детализированная схема с включенным экспертным участием. Примером работы, где предложена детальная схема, является статья Черненко В.В. и Пискорской С.Ю., где более детально описана экспертная составляющая системы (рис. 1.2) [Черненко, Пискорская, 2012].

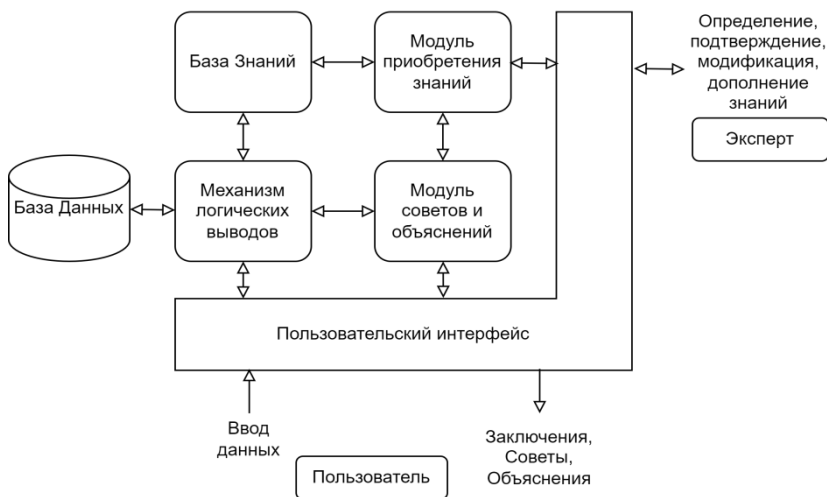


Рис. 1.2. Структура экспертной системы

Источник: [Черненко, Пискорская, 2012].

Анализ предложенных концептуальных схем позволил разработать общую схему, описывающую взаимосвязи и структуру базы знаний для научно-исследовательских задач (рис. 1.3).

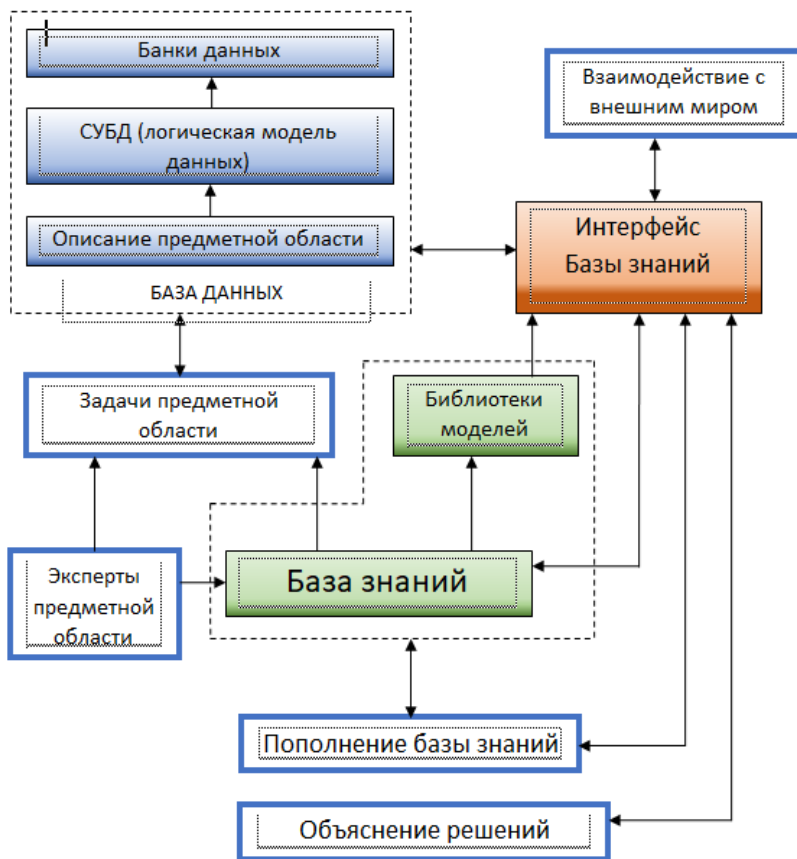


Рис. 1.3. Структура базы знаний, адаптированной для исследовательского проекта «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий»

Дальнейшие шаги будут приводить к выстраиванию базы знаний на основе предложенной структуры.

Видно, что основой разрабатываемой базы знаний является база данных. Она представляет костяк, наращивающейся и формирующейся системы. Задачи, которые должна решать База знаний, формируют требования к структуре базы данных, а также дополнительные требования к системе, которые были зафиксированы перед выбором основной платформы БД.

1.1. Основные требования к платформе базы данных

Экспертное сообщество ИЭОПП СО РАН, после анализа поставленных задач сформировало ряд требований к системе, которым она должна соответствовать в процессе формирования, развития и эксплуатации. Предлагаемые общие требования к базе знаний включают требования к базе данных, инструментально-аналитическое наполнение и требование к интерфейсу БЗ.

К базе данных, как структурообразующей единице, необходимо особо тщательное формулирование как общих целей, так и технического задания к её основным функциональным возможностям.

1.1.1. Задачи базы данных

Создаваемая база данных должна обеспечивать:

- сбор и хранение актуальных данных, описывающих развитие отдельных экономических объектов на территории Азиатской России и за её пределами;
- формирование удобного интерфейса, предоставляющего доступ к информации сотрудникам и заинтересованным организациям;
- предоставление доступа к информации с использованием программных средств и языков программирования;
- построение на основе базы данных прогнозно-аналитического комплекса и базы знаний, включающих программный интерфейс для визуализации и обработки данных.

1.1.2. Техническое задание функциональных возможностей базы данных

Работа с данными

Для удобства должна быть создана иерархическая структура хранения данных. Данные хранятся блоками, которые в свою очередь группируются в более крупные структуры и т.д.

Основные классы данных, с которыми должна работать система:

- табличные данные без привязки ко времени,
- временные ряды,
- пространственные (геопривязанные),
- текстовые.

Возможно появление и других классов данных.

Сбор данных

Необходимо обеспечить сбор и хранение данных из различных источников, имеющих разнообразную структуру и природу. Сбор данных должен быть автоматизирован, запускаться по расписанию либо по требованию. В случае повторной загрузки должна быть предусмотрена система фильтрации/обновления совпадающих данных.

Права доступа

Поскольку предполагается совместное пользование базой знаний сотрудниками разных категорий, необходимо разделение классов пользователей на администратора, эксперта, пользователя и др. Разные пользователи должны иметь разный уровень доступа к данным. Должны быть предусмотрены следующие уровни работы с данными (и их комбинации):

- создание,
- изменение,
- аудит/проверка,
- только чтение;
 - всех данных,
 - некоторой части данных.

Предоставление доступа на чтение должно гибко настраиваться. К разным блокам данных могут применяться различные уровни

доступа. Нужно иметь возможность ограничить доступ пользователю вплоть до минимально возможного в системе блока данных.

Доступ к системе возможен только авторизованным пользователям.

Валидация

ГОСТ Р ИСО 9000–2008 определяет термины «верификация» и «валидация» следующим образом: «Верификация – подтверждение на основе представления объективных свидетельств того, что установленные требования были выполнены. Валидация – подтверждение на основе представления объективных свидетельств того, что требования, предназначенные для конкретного использования или применения, выполнены» [Макарычев и др., 2016].

Система должна позволять осуществлять проверку части данных экспертом. Пользователь с соответствующими правами просматривает блок данных, проверяет его и при необходимости вносит изменения. После этого ставится пометка о валидации.

Кроме эксперта необходимо предусмотреть автоматизированную проверку корректности данных.

Валидированные и не валидированные данные хранятся в системе одинаково. Валидация данных не влияет на их доступность пользователям, однако каждый блок данных должен содержать информацию о его проверке.

Версионирование

Система должна отслеживать возможные изменения в сохранённых данных, предоставлять возможность просмотра предыдущих версий данных и отката изменений.

Каталог данных

Каталог – это один из способов структурированного доступа к данным в системе. Каталог содержит список всех данных, доступных конкретному пользователю. Список не фиксирован, меняется в следующих ситуациях:

- поступление новых данных,
- удаление устаревших данных,
- разрешение/запрет пользователю доступа к конкретному блоку данных.

Поиск и фильтрация

Система должна предоставлять гибкий поиск и фильтрацию содержимого.

- Полнотекстовый поиск по:
 - названию,
 - описанию,
 - внутреннему содержанию.
- Фильтрация по характеристикам:
 - категории каталога,
 - экономическим (отрасль, производственный комплекс, предприятие и т.д.),
 - временным (год, месяц, квартал и т.д.),
 - пространственным:
 - регион (субъект РФ, федеральный округ, экономический район и т.д.),
 - код (ОКВЭД и т.д.),
 - произвольная область на карте,
 - прочим характеристикам, специфичным для конкретного вида данных.
- Прочее:
 - теги,
 - ключевые слова,
 - другое.

Операции с данными

Система должна позволять следующие операции с данными:

- экспорт для локальной работы,
- визуализация,
- моделирование:
 - экономические модели,
 - ГИС-модели,
- аналитические запросы,
- копирование в пользовательское пространство.

Экспорт

У пользователя должна быть возможность выгрузить данные из системы. Нужно предусмотреть выгрузку в форматы: CSV,

XLSX, PDF, PNG как наиболее часто используемые в научных исследованиях форматы. Возможно появление других форматов. Выбор конкретного формата зависит от типа данных.

Также следует предусмотреть пользовательский интерфейс написания запросов к данным.

Моделирование

Система должна позволять хранить, извлекать и визуализировать данные для экономического и геопространственного моделирования. Впоследствии – строить и рассчитывать эти модели на сервере.

Моделирование включает:

- построение специализированных экономико-математических моделей (система OMVEAR, КАМИН и др.),
- ГИС-моделирование (кластеризация, суммирование по области и др.),
- моделирование взаимосвязей данных (корреляционных, регрессионных и прочих взаимосвязей),
- построение прогнозных моделей (подбор оптимальной прогнозной модели на основе регрессионных моделей, машинного обучения, нейронных сетей).

Аналитические запросы

В системе должна быть возможность формирования аналитических запросов к данным, построения отчетов по шаблонам (предопределенным или пользовательским). Следует предусмотреть также пользовательский интерфейс написания аналитических запросов и построения шаблонов отчетов.

Пользовательское пространство

Система должна иметь раздел, в котором пользователь индивидуально работает со своей копией данных отдельно от основного хранилища. Он называется пользовательским пространством или "песочницей".

Рассмотренные требования станут основными критериями дальнейшего выбора системы управления базами данных.

1.2. Общая схема работы базы знаний

Анализ технического задания к базе данных и потенциального потока данных, необходимый для формирования пространственных цепочек создания социальной ценности, показал необходимость построения в рамках БЗ двух видов БД: структурированной и неструктурированной.

Структурированная БД предусматривает жесткий формат данных и их взаимосвязей. Это позволяет выполнять необходимые расчеты и в последующем применять различные инструменты по анализу данных, которые работают именно со структурированными данными. Для реализации подобной БД было принято решение использовать структуры реляционной базы данных PostgreSQL. Она удовлетворяет всем требованиям к платформе БД, описанным в разделе 1.1, и имеет ряд важных свойств – таких как:

- возможность хранения структурированных данных и создание SQL запросов к их извлечению;
- интеграция с ключевыми в проекте языками программирования: Python и R;
- высокая скорость выполнения запросов;
- интеграция с Веб и ГИС интерфейсами, используемыми при реализации базы знаний;
- совместимость с инструментами постоянного обновления БД;
- наличие интерфейса первичного пользования для анализа и пополнения БД.

Неструктурированная БД включает в себя наборы неструктурированных данных, таких как тексты, отчеты и неструктурированные табличные данные. На данном этапе реализации проекта предполагается их хранение в общем файловом хранилище с сегментацией и тщательным описанием метаданных. Алгоритмы запросов к данным предполагается создать в рамках реализации интерфейса пользователя.

Оба вида БД формируют основу БЗ и встраиваются в общую архитектуру системы, дополняемые системой загрузки данных, внутренними сервисами в виде модельного аппарата, а также интерфейсом пользователя. Общая схема работы базы данных представлена на рис.1.4.

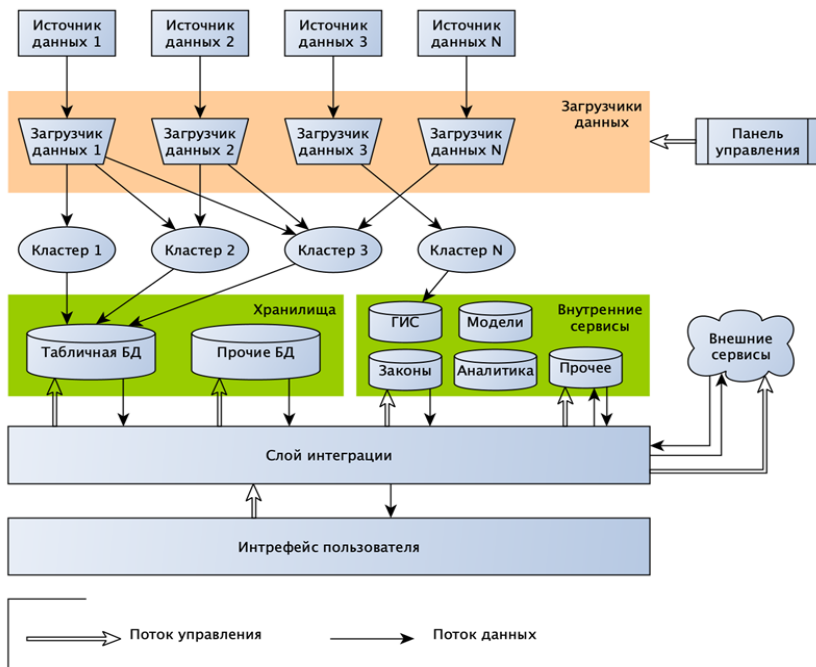


Рис. 1.4. Общая схема базы знаний

В рамках реализации БЗ для увеличения эффективности и скорости обработки и анализа данных было принято решение создать вычислительный комплекс прогнозирования пространственно-распределенных взаимодействий регионов Азиатской России на основе сервис-ориентированной геоинформационной системы анализа и оценки различных сценарных условий их комплексного социально-экономического развития во взаимосвязи с экономикой России.

Уникальность данного комплекса обеспечена и реализована в ходе решения двуединой задачи в отношении используемого оборудования, конструктивного построения взаимоувязанного комплекса и, самое главное, оснащения его специализированным программным обеспечением, как приобретенного у профессиональных ИТ компаний, так и собственно разрабатываемого авторского программного обеспечения в виде ин-

терфейса пользователя базы знаний и пакетов для прямого подключения к базе знаний через языки программирования SQL, R и Python.

Комплекс представляет собой специализированную платформу, которая позволит собирать и хранить большие объемы структурированных и плохо структурированных данных (природных ресурсов территорий Азиатской России, используемых в мире технологий добычи и переработки полезных ископаемых, транспортных путей и т.д.), выявлять взаимосвязи различных процессов при помощи разработанных алгоритмов, и в том числе, обосновывать видение развития территории Азиатской России и генерации сценариев развития для принятия, в том числе, управленческих решений.

Комплекс создается на основе высокопроизводительного сервера, настроенного на потоковый сбор, обработку и хранение данных. Принципиальная Схема комплекса соответствует общей схеме базы знаний (см. рис. 1.4) и расширена дополнительной настройкой кластеров рабочих станций (рис. 1.5).

В составе Комплекса организованы специализированные рабочие станции исследователей, каждая из которых оснащена специальным ПО для решения задач в соответствии техническим заданием и достижением поставленных целей. Создание сети рабочих станций также позволяет проводить многопоточные вычисления, сбор и анализ данных.

Связь систем специализированных рабочих станций с сервером базы знаний по локальной сети предполагается с использованием Веб-интерфейса или авторских разрабатываемых пакетов базовых аналитических языков программирования: Python, R и др.

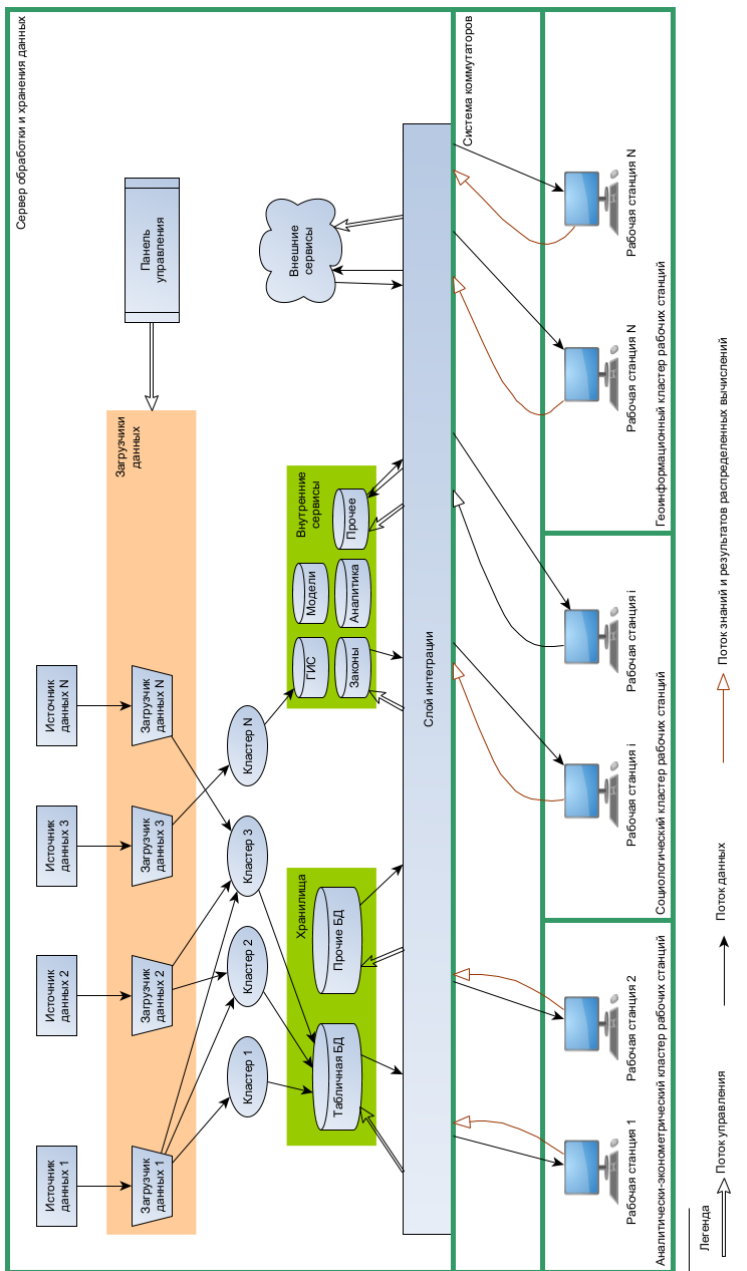


Рис. 1.5. Принципиальная Схема Комплекса Базы Знаний

1.3. Принципы работы с данными в рамках базы данных

Работа с большим объемом информации создает необходимость проработки общих требований и подходов к организации и пополнению БД. В этом разделе рассматриваются принципы работы с данными в рамках создания БД.

Всё множество данных, которое нужно собирать и хранить, делится на *кластеры* по принципу «похожести» и «близости» их смысла и/или внутренней структуры. Таким образом кластер представляет собой набор примерно однородных данных. Кластер может содержать данные из нескольких источников, если они дополняют друг друга. Например, данные об инвестиционных проектах собираются из разных источников, но преобразуются в единую структуру.

Сбор данных

Для каждого источника данных разрабатывается специальный, специфичный для него загрузчик, который знает:

- структуру данных внешнего источника,
- формат внутреннего хранилища (например, схему таблиц в БД),
- правила, по которым входные данные преобразуются,
- правила, по которым преобразованные данные складываются в хранилище.

При запуске загрузчик:

- получает данные внешнего источника (загружает их по URL, читает из файловой системы и т.д.),
- анализирует данные,
- извлекает из них нужную информацию,
- преобразует/обогащает их при необходимости,
- помещает в хранилище.

Процесс работы всех загрузчиков осуществляется:

- централизованно,
- повторяемо,
- автоматизированно,
- по расписанию / по требованию (на выбор),
- стабильно (с отчётом об ошибках).

Конфигурирование работы всей системы загрузки и каждого загрузчика в отдельности осуществляется через панель управления.

Таким образом, система загрузки реализует ETL-процесс (включает в себя извлечение, преобразование и загрузку данных). Так же на текущий момент внедряются ELT процессы, которые в отличие от ETL процессов предполагают загрузку данных без преобразований в озеро данных с последующей обработкой.

Хранение данных

Хранение данных системы осуществляется в реляционной БД.

Основные классы таблиц:

- таблицы с данными кластеров,
- таблицы с общими данными,
- таблицы со служебными данными.

Данные кластеров

Хранение данных каждого кластера осуществляется в виде набора таблиц, специфичных для конкретного кластера. Схема этих таблиц известна заранее, но может меняться со временем при добавлении источников данных или изменении структуры данных в них.

Общие данные

Формируется набор таблиц с данными, общими для двух или более кластеров. Схема этих таблиц также известна заранее, но меняется реже схемы данных кластеров.

Данные в этих таблицах служат:

- частичным описанием данных кластеров,
- опосредованной связью между кластерами.

Примером таких данных являются ведомственные классификаторы, коды и названия регионов, единиц и т.д.

Служебные данные

Служебные данные необходимы для корректного функционирования системы в целом. Они включают в себя:

- метаданные о кластерах (типы, источники, описания, прочие характеристики содержащихся в каждом кластере данных),
- данные о пользователях, правах доступа, настройки и т.д.

Слой интеграции

Слой интеграции представляет собой серверное приложение, основными задачами которого являются:

- предоставление пользователям доступа к данным,
- интеграция данных из хранилищ / внутренних и внешних сервисов / прочих подсистем для удобного представления и использования в UI,
- сокрытие внутренней реализации хранилищ,
- разделение прав доступа к системе и её частям (авторизованные / не авторизованные, доступы к сервисам и т.д.),
- администрирование системы,
- ограничение нагрузки,
- валидация и фильтрация данных от пользователей.

Слой интеграции реализует основную бизнес-логику работы системы. С одной стороны, он взаимодействует с интерфейсом пользователя через RESTful Web API. С другой стороны – с подсистемами через их родные протоколы работы (драйверы реляционных БД, REST/SOAP API веб-сервисов и т.д.) Таким образом, он позволяет совместно функционировать разнородным частям системы и пользовательскому интерфейсу.

Мониторинг

При эксплуатации системы важно:

- знать, что происходит в системе,
- вовремя/заранее реагировать на нештатные ситуации.

Для этого каждая её часть снабжается системой диагностического мониторинга событий, а также сбора логов. Собранная информация о статусе системы отображается в специализированном графическом интерфейсе.

Глава 2

ОСНОВНЫЕ КОМПОНЕНТЫ БАЗЫ ДАННЫХ

Для решения основных задач БЗ были выбраны необходимые блоки данных, которые планируется постоянно накапливать и обновлять:

- Блок информационного обеспечения (компании, секторы, комплексы, регионы, кластеры),
- Блок проектов – инвестиционных, программных, институциональных,
- Блок по углеводородному сырью (ресурсные кластеры, компании, участки недр, состояние геологоразведочных и добычных работ),
- Блок по отдельным видам твёрдых полезных ископаемых (ресурсные кластеры, компании, участки недр),
- Блок по подземным водам (месторождения),
- Блок технологий.

Данные, указанные выше, имеют разнородный характер (типы, форматы, логику), так как принадлежат разным организациям и ведомствам, и с учетом этого требовалось разработать оригинальную структуру базы данных, что подтверждается свидетельством о государственной регистрации программ для ЭВМ¹.

Для этого к блокам основных данных были добавлены блоки дополнительных, технических данных, включающих в себя блоки «common» и «history». Блок «common» имеет таблицы для хранения данных о пользователях, их сессиях и разграничении прав доступа к данным. Блок «history» состоит из двух таблиц, которые хранят информацию об изменении записей в таблицах о компаниях и статистических данных.

Реализация функционала для работы записи истории изменений была осуществлена с помощью триггеров, по результатам ра-

¹ Жирнов А.А., Харлампенков И.Е., Юрченко А.В. Программа для создания таблиц, связей и полей с разными типами данных для хранения структурированной и слабоструктурированной информации: свидетельство № 2021666300; патентообладатель – Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ); заявл. 04.10.2021; опубл. 12.10.2021, заявка № 2021665475.

боты было получено свидетельство о государственной регистрации программ для ЭВМ¹.

2.1. Блок информационного обеспечения (компании, секторы, комплексы, регионы, кластеры)

Блок информационного обеспечения содержит в себе информацию макро-, мезо- и микроуровня, имеющую сопроводительный характер. Для решения задач БЗ были сформированы группы кластеров данных, позволяющих производить анализ территории и прогнозирование траекторий роста. Группы кластеров включают в себя:

- социально-экономические статистические данные,
- данные о компаниях,
- таможенную статистику.

В дальнейшем планируется дополнение базы данных информацией, описывающей общую экономическую конъюнктуру, такие как курсы валют, стоимость основных товарных групп на фондовой бирже и т.д.

2.1.1. Социально-экономические статистические данные

Статистические данные представляют собой информацию, собранную из разных статистических источников, но имеющую единую структуру формата панельных данных (временные ряды). Структура панельных данных имеет формат:

- Год,
- Регион (также может быть страна или федеральный округ),
- Параметр,
- Значение,
- Заметка (в случае отсутствия данных указывается причина),
- Дата создания,
- Автор, внесший данные.

¹ Жирнов А.А., Харлампенков И.Е., Юрченко А.В. Программа сохранения изменений в ячейках одной таблицы базы данных в другую таблицу с записью версии: свидетельство № 2021666299; патентообладатель Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ); заявл. 04.10.2021; опубл. 12.10.2021, заявка № 2021665476.

Общая схема блока статистических данных представлена на рис. 2.1. Пример основной таблицы представлен на рис. 2.2.

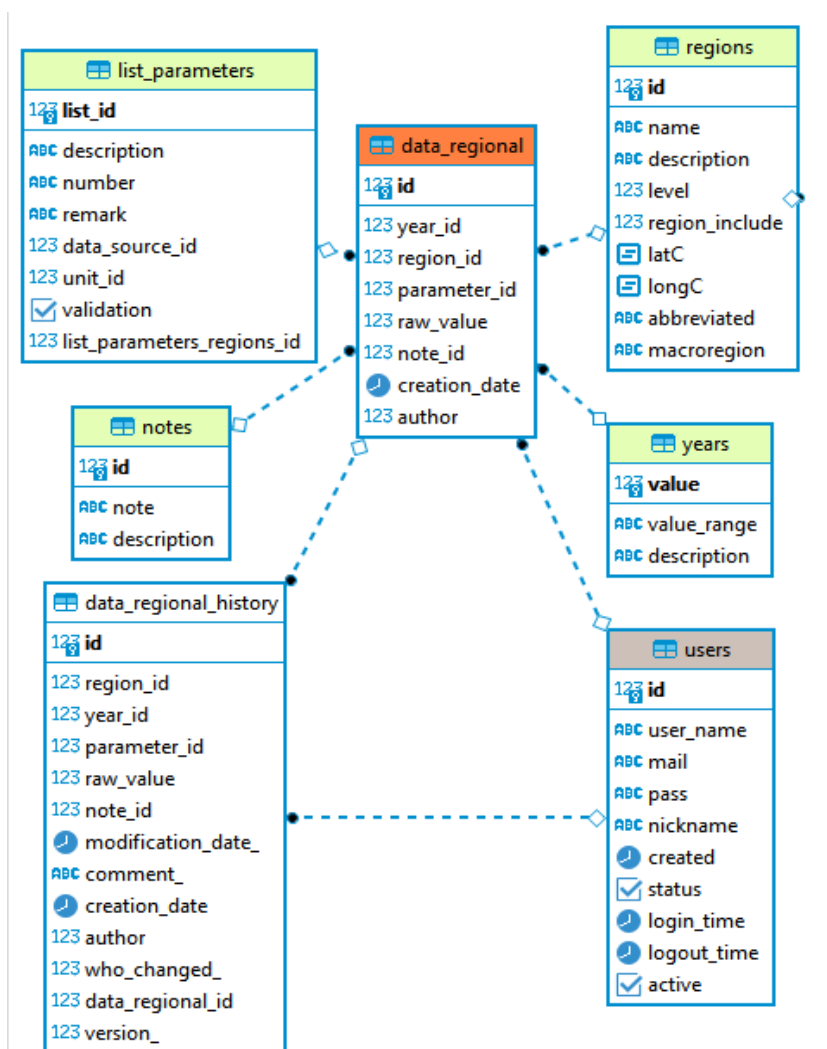


Рис. 2.1. Схема блока статистических данных

data_regional Введите SQL выражение чтобы отфильтровать результаты									
Таблица	123 year_id	123 region_id	123 parameter_id	123 raw_value	123 note_id	creation_date	123 author	id	
	1	1	3 078	0,25999999	4	2021-11-17 22:01:41.760 +0700	3	1 110 289	
	2	2	1	38 175,1	4	2021-08-31 10:44:21.590 +0700	3		2
	3	4	1	1 407,9	4	2021-08-31 10:44:21.590 +0700	3		4
	4	6	1	2 422,4	4	2021-08-31 10:44:21.590 +0700	3		6
	5	7	1	1 179	4	2021-08-31 10:44:21.590 +0700	3		7
	6	8	1	1 060,3	4	2021-08-31 10:44:21.590 +0700	3		8
	7	9	1	754	4	2021-08-31 10:44:21.590 +0700	3		9
	8	10	1	1 266,4	4	2021-08-31 10:44:21.590 +0700	3		10
	9	11	1	1 228,1	4	2021-08-31 10:44:21.590 +0700	3		11
	10	12	1	6 613,5	4	2021-08-31 10:44:21.590 +0700	3		12
	11	13	1	876,7	4	2021-08-31 10:44:21.590 +0700	3		13
	12	14	1	1 258,8	4	2021-08-31 10:44:21.590 +0700	3		14
	13	15	1	1 082,2	4	2021-08-31 10:44:21.590 +0700	3		15

Рис. 2.2. Фрагмент основной таблицы статистических данных

Источниками данных являются статистические сборники, такие как «Регионы России», «Российский статистический ежегодник», «База данных ЕМИСС» и др.

Ключевой точкой создания такой базы данных является их верификация. Проведение верификации осуществляется через эксперта, для которого по каждому показателю формируются карточки, включающие показатели:

- Название;
- Источник данных;
- Регион;
- Наличие всплесков темпов прироста;
- График показателя;
- Заметки по причинам пропусков данных.

Визуализация и первичный анализ ускоряют работу эксперта и позволяют быстро выявить проблемные точки, которые возникают в процессе наполнения базы данных.

2.1.2. Данные о Компаниях

Информационный блок о компаниях представляет собой связку данных из разных источников, таких как СПАРК-Интерфакс, list-org, РБК Компании и др. Ключевым источником является база данных СПАРК (<https://spark-interfax.ru/>). Пример данных основной информационной таблицы о компаниях представлен на рис. 2.3. Структура данных о компании представлена в табл. 2.1.

Таблица 2.1

Структура данных о компании

Внутренний идентификатор компании	Сайт в сети Интернет	Коды всех видов деятельности
Краткое наименование	Форма собственности	Код ОКОПФ Внешний ключ
Адрес (место нахождения)	Налоговый орган	Руководитель – ИНН
Руководитель – ФИО	Сумма исполнительных производств	Код ОКATO
Дата регистрации	Размер компании	Код ОКТМО
Дата ликвидации	Уставный капитал, RUB	Ранее использованные ИНН
Регион регистрации	Налоговый режим	Ранее использованные ОГРН
Предмет поставки	Координаты объекта (ГИС)	Руководитель – должность
Наименование на английском	Код налогоплательщика	Источник данных
Наименование	Вид деятельности/отрасль	Наименование ОКATO

С этой таблицей (см. рис. 2.3) взаимосвязаны еще две ключевые таблицы: Бухгалтерская информация и Коды видов деятельности.

Бухгалтерская статистика о компаниях ведет начало с 2005 г. Пример данных таблицы бухгалтерской отчетности представлен на рис. 2.4. Структура показателей бухгалтерской отчетности компаний выглядит следующим образом:

- Внутренний идентификатор компании,
- Год,
- Идентификатор показателя (связан с таблицей перечня показателей),
- Значение показателя.

Не все компании имеют данные о бухгалтерской отчетности, и не все данные указаны корректно, что стимулирует поиск новых источников данных для расширения и корректировки потока информации. В рамках расширения и обновления блока данных о компаниях используются данные из открытых источников, та-

ких как <https://www.list-org.com/> и <https://companies.rbc.ru/>. В случае отклонения показателей проводится дополнительная процедура связывания данных из разных источников.

company_financial_statements Введите SQL выражение чтобы отфильтровать результаты

	123 company_id	123 year_id	123 value	123 spark_code_id	123 parameter_id	123 id
1	[NULL]	2 020	-7 000	363 452	63	3 841
2	[NULL]	2 020	1 204 589 000	102 907	63	3 842
3	[NULL]	2 020	32 000	135 369	63	3 843
4	[NULL]	2 020	5 708 000	135 091	63	3 844
5	[NULL]	2 020	3 800 000	1 319 059	63	3 845
6	[NULL]	2 020	31 000	84 581	63	3 846
7	[NULL]	2 020	164 000	189 208	63	3 847
8	[NULL]	2 020	295 000	101 931	63	3 848
9	[NULL]	2 020	8 678 000	226 008	63	3 849
10	[NULL]	2 020	99 000	933 749	63	3 850
11	[NULL]	2 020	20 000	1 971 001	63	3 851
12	[NULL]	2 020	-324 000	115 670	63	3 852
13	[NULL]	2 020	5 028 000	1 066 337	63	3 853
14	[NULL]	2 020	3 000	8 558 900	63	3 854
15	[NULL]	2 020	-14 000	54 152	63	3 855

Рис. 2.4. Фрагмент основной таблицы данных
о бухгалтерской отчетности компаний

Проблемой существующего набора данных является наличие только юридического адреса, а не места осуществления экономической деятельности, что не дает полной картины насыщенности территории. Для решения этой проблемы было принято выделить 1000 крупных компаний, осуществляющих деятельность в разных регионах Российской Федерации, и создать ручную базу данных «дочерних компаний», которые входят в тот или иной холдинг. На данный момент, этот раздел находится в разработке.

С использованием показателя Адрес (юридический адрес компании) была реализована привязка компаний к ГИС-пространству. Привязка происходит с использованием Кластера рабочих станций экономико-математического моделирования, на котором был реализован потоковый запуск скрипта, написанного на языке программирования R и использующего открытый API – nominatim.openstreetmap.org. Но точность инструмента оказалась низкой, в результате отсутствия части адресов и направленности инструмента на европейский рынок. В дальнейшем была проведена процедура уточнения координат через API Яндекс карт и оценки качества результатов через проверки соответствия координат компании координатам полигона её региона. Это позволяет

проверить данные и сделать перенос данных о компаниях в ГИС-интерфейс и применять для анализа пространственное и ГИС-моделирование.

Связка таблиц компаний, ОКВЭД, бухгалтерской отчетности и привязки к ГИС позволяет осуществлять пространственный анализ территории и дает богатую основу для моделирования. Но ключевым вопросом становится надежность данных.

2.1.3. Анализ достоверности данных компаний

В отличие от статистических данных, данные о компаниях имеют более высокую погрешность, поскольку они поступают в базу данных без каких-либо дополнительных проверок. В связи с этим для данных компаний в качестве этапа предварительной обработки выступает анализ достоверности предоставленной информации.

В ходе интервью с экспертами ИЭОПП СО РАН было выявлено, что основной ошибкой при работе с данными из используемых нами источников был резкий прирост выручки компаний – более чем в 100 или даже в 1000 раз за год. Предположительно, данные ошибки появились вследствие некорректно внесённой информации в первоисточник.

Также в ходе предварительного анализа данных был выявлен ряд ошибок в других показателях отчётности компании. Например, расхождение активов и пассивов, расчета прибыли и т.д. Таким образом, для нахождения компаний, показатели которых содержат ошибки, была поставлена задача реализовать алгоритм, который будет:

- находить компании, в которых темп прироста выручки превышает 100 и 1000 раз и присваивать им соответствующий код ошибки;
- сравнивать строки отчётности с агрегированными показателями: и в случае, если агрегированный показатель оказывается меньше показателя, который является составной частью данного агрегированного показателя – компании присваивается соответствующий код ошибки.

Используемая в работе база данных компаний состоит из трех таблиц:

- Общее описание компаний – содержит данные весом 3 ГБ о 12 миллионах компаний,

- Финансовая отчетность компаний весом более 60 ГБ,
- ОКВЭД – 3 ГБ.

Так как вес основной рабочей таблицы с данными слишком большой, этот элемент базы данных был разбит на отдельные файлы по каждому из регионов.

Реализация программного модуля для реализации метода проверки темпов прироста компании осуществляется по следующему алгоритму:

1. Загрузка необходимых библиотек,
2. Загрузка данных в программу,
3. Получение уникального списка компаний по региону,
4. Реализация функции для определения значения параметра,
5. Реализация функции для расчета суммы параметров компании,
6. Реализация функции для расчета темпов прироста,
7. Реализация функции для нахождения ошибок в других показателях отчетности компании,
8. Реализация цикла для расчета темпов прироста и проверки прочих показателей отчетности.

Для работы данного программного модуля были использованы библиотеки на языке Python.

Для получения уникального списка компаний по каждому из файлов с данными, чтобы не усложнять систему, код для расчета параметра был реализован в виде отдельной функции. В качестве входных параметров функция принимает название блока данных, идентификатор параметра и идентификатор компании, а в качестве результата работы возвращает уникальное значение параметра. Далее реализовывалась функция для расчета суммы параметров компании, которая принимает на вход название набора данных, список параметров для суммирования и идентификатор компании.

Реализация алгоритма расчета темпов роста выручки компании осуществлялась следующим образом. Его основная функция – прием на вход идентификатора компании и выдача кода ошибки. Внутри самой функции по каждому идентификатору компании производится выборка показателя выручки по всем годам (с 2005 по 2020 год). Соответственно, в случае отсутствия данных за какой-либо год показатель прироста не рассчитывался и на код ошибки не влиял.

Далее производился расчёт темпов прироста и в случае:

- 1) если был найден темп прироста от 100 до 1000, алгоритм присваивал данной компании код ошибки «1»;
- 2) если же темп прироста оказывался больше 1000, алгоритм присваивал данной компании код ошибки «2»;
- 3) если же не было никакой ошибки, то алгоритм присваивал данной компании код ошибки «0».

Тем не менее данные коды ошибок ещё не являются конечными, так как после проверки темпов роста производится переход к следующему этапу проверки, на котором осуществляется проверка всех остальных показателей компании. Для реализации данного этапа была отдельно составлена таблица соответствия строчковых и агрегированных показателей отчётности.

В случае если агрегированный показатель оказывается меньше какого-либо из входящих в него показателей отчетности, компании присваивался код ошибки «3». Для корректной работы алгоритма в ходе проверки было оставлено допустимым отклонение в 5%, которое может возникнуть вследствие округления или каких-либо иных причин.

Входным параметром для функции нахождения ошибок в показателях отчетности компании является идентификатор компании, а в качестве выходного значения возвращается код ошибки. Таким образом, на данном этапе проверяются все связанные между собой показатели на явные ошибки, и при нахождении ошибок в показателях компании присваивается соответствующий код.

Цикл расчетов темпов прироста и проверки прочих показателей отчетности не отличается от процедуры для показателя выручки. В результате работы программы получено однозначное соответствие компании и параметра, который определяет степень достоверности данных:

- 0 – ошибок нет,
- 1 – темп прироста от 100 до 1000,
- 2 – темп прироста больше 1000,
- 3 – ошибка в показателях отчетности,
- 13 – ошибка в темпах прироста и показателях отчетности,
- 23 – ошибка в темпах прироста и показателях отчетности.

Хотелось бы отметить, что данные значения параметров не являются конечным выводом, а являются лишь маркером для экспертов, которые будут осуществлять дальнейшую проверку показателей компании.

Финальным этапом реализации рассмотренного алгоритма проверки данных на ошибки является его добавление в базу данных компаний.

Подобные этапы проверки позволяют в последующем исключать показатели с наличием ошибок при проведении пространственной аналитики, потому что наличие всплесков в данных приводит к смещению агрегированных показателей. Далее планируется этап корректировки ошибок за счёт сверки показателей компаний с ошибками с данными альтернативных источников и формировании карточки компании с описанием проблем для ручной проверки экспертами.

2.1.4. Таможенная статистика

Блок данных таможенной статистики имеет в основе открытый источник Федеральной таможенной службы (URL: <http://stat.customs.gov.ru/unload>) и полностью повторяет его структуру, с учетом привязки всех данных к общей таблице регионов. Пример данных таблицы таможенной статистики представлен на рис. 2.5.

customs_data Введите SQL выражение чтобы отфильтровать результаты

	id	123 region_id	abc napr	abc invned_id	123 unit_id	123 stoim	123 netto	123 kol	123 year_id	abc month_id	abc list_countries_id
1	779440	20	ИМ	6110209100	109	169,44	5,949	15	2018	Апрель	JP
2	902958	33	ИМ	6109902000	109	678,59	1,4	7	2018	Апрель	JP
3	2682860	12	ИМ	6203431900	109	4477,16	10,76	19	2018	Март	JP
4	3109045	33	ЭК	4407129100	104	247100,97	492132	1105,2	2018	Ноябрь	JP
5	3421117	86	ЭК	4407119300	104	58386,94	124500	225,24	2018	Октябрь	JP
6	2061	20	ИМ	2106909003	[NULL]	15185,15	392	0	2018	Август	JP
7	2190	20	ИМ	1902301000	[NULL]	616,54	15	0	2018	Август	JP
8	2367	90	ЭК	2301200000	[NULL]	203503,5	198540	0	2018	Август	JP
9	2467	90	ИМ	2103100000	[NULL]	1851,19	843,54	0	2018	Август	JP
10	2610	20	ИМ	2918300000	[NULL]	51,82	0,08	0	2018	Август	JP
11	2715	94	ИМ	2103909009	[NULL]	2138,81	526	0	2018	Август	JP
12	2831	29	ЭК	2818101900	[NULL]	6360	8000	0	2018	Август	JP
13	3047	78	ИМ	4016910000	[NULL]	217,6	135	0	2018	Август	JP
14	3069	91	ИМ	4016995709	[NULL]	1575,18	1,27	0	2018	Август	JP
15	3183	60	ИМ	3926909709	[NULL]	97844,91	2152,266	0	2018	Август	JP

Рис. 2.5. Фрагмент основной таблицы таможенной статистики

Привязка к ГИС-формату производится так же, как в случае статистических показателей за счет привязки к полигонам соответствующих регионов. В целом выбранный формат данных позволяет производить любые аналитические срезы данных в необходимом исследователю формате.

2.2. Блок инвестиционных, программных, институциональных проектов

Блок проектов содержит информацию об инвестиционных проектах различной стадии реализации и размера, сконцентрированных как на территории Азиатской России, так и на территории Европейской России. Основное наполнение Блока проектов на начальном этапе происходило на базе информации об инвестиционных проектах, представленных на цифровой платформе «Инвестиционные проекты»¹, находящихся на разных стадиях реализации (планирование, предпроектные проработки, проектирование, подготовка к строительству, строительство, модернизация) в различных регионах России, но уже была разработана система поиска и сбора данных из сети источников. Отраслевой и пространственный анализ проектов платформы позволят оценить роль использования ресурсов в синергии инвестиций, выявить положительные и отрицательные их эффекты, сформировать наиболее эффективные цепочки взаимодействия проектов и направления инвестиционных импульсов. При этом под ресурсами здесь понимаются: материальные активы; результаты интеллектуальной деятельности; человеческие ресурсы; рынки сбыта товаров или услуг, возможности продвижения и т.д. Тем самым сформированы

¹ Право использования Программного обеспечения «Инвестиционные проекты» предоставлено ООО «ПКР Аналитика» по Лицензионному договору №119-11/21, заключенному при финансовой поддержке Минобрнауки РФ в рамках выполнения работ по крупному научному проекту «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий» (регистрационный номер заявки 2020-1902-01-377).

научно-обоснованные, ориентированные на достижение оптимального системного эффекта проекты, и на их основе сформулирована концепция стратегии комплексного освоения и развития территорий Азиатской России. Кроме того, Блок проектов позволяет разработать концепцию территориальной инвестиционной политики, призванной обеспечить устойчивый экономический рост и повышение благосостояния населения территорий Азиатской России.

Текущая реализация Блока проектов предполагает два вида хранения данных: структурированный и неструктурированный.

Неструктурированный вид реализован в виде хранилища карточек инвестиционных проектов с сортировкой по отраслям и регионам, а также включает структурированное описание метаданных. В рамках формирования Веб интерфейса был реализован блок работы с карточками инвестиционных проектов с текстовым поиском.

Другими источниками данных является платформа «B2b Global» и платформа «Росинфра». Данные с этих платформ преобразовались в базовый формат и использовались для расширения имеющейся базы. Наличие нескольких источников информации может приводить к дублированию информации и её несогласованности. Поэтому в базу добавлен идентификатор источника данных, и пользователь получает возможность сортировки данных по нему. В дальнейшем предполагается введение процедуры экспертного согласования проектов.

2.3. Блок по углеводородному сырью, отдельным видам твердых полезных ископаемых и подземным водам

Спецификой созданной базы данных Проекта «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий» является включение в нее данных, характеризующих ресурсный потенциал Азиатской России по углеводородному сырью (УВС), твердым полезным ископаемым (ТПИ) и подземным водам.

Блок базы данных по углеводородному сырью представлен следующим списком приложений (рис. 2.6):

- Анадырско-Охотоморский перспективный кластер;
- Ангаро-Ленский действующий кластер;
- Байкитско-Катангский действующий кластер;
- Вилуйский перспективный кластер;
- Демьянский перспективный кластер;
- Енисей-Хатангский перспективный кластер;
- Западно-Сибирский действующий кластер;
- Карабашский перспективный кластер;
- Курейский перспективный кластер;
- Непско-Ботубинский действующий кластер;
- Полуйско-Ляпинский перспективный кластер;
- Пур-Тазовский перспективный кластер;
- Сахалинский действующий кластер;
- Хатангско-Ленский перспективный кластер;
- Юго-Восточный (Томский) перспективный кластер;
- Южно-Карский перспективный кластер.

В составе электронных приложений, соответствующих перспективным кластерам УВС, заполнялись следующие шаблоны:

- характеристика месторождений перспективного кластера;
- характеристика подготовленных объектов перспективного кластера;
- характеристика выявленных объектов в пределах перспективного кластера;
- характеристика перспективных площадей в пределах перспективного кластера.

В составе электронных приложений, соответствующих действующим кластерам УВС, заполнялись следующие шаблоны:

- характеристика запасов действующего кластера;
- характеристика ресурсов действующего кластера.

По инфраструктуре, имеющей отношение к УВС, заполнялись следующие шаблоны:

- нефтепроводы;
- газопроводы;
- конденсатопроводы;
- железные дороги;

- автомобильные дороги;
- автозимники;
- порты;
- аэропорты.

Блок базы данных по направлению работ на углеводородное сырье через ИНН компаний связан с Блоком данных о Компаниях и также имеет ГИС привязку.

Блок базы данных по твердым полезным ископаемым представлен следующим списком приложений (рис. 2.7):

- Уральский кластер;
- Таймырский кластер;
- Норильский кластер;
- Ангара-Енисейский кластер;
- Кузбасско-Алтайский кластер;
- Тувинско-Хакаский кластер;
- Ангара-Анабарский кластер;
- Бурятский кластер;
- Забайкальско-Амурский кластер;
- Южно-Якутский кластер;
- Приморско-Сахалинский кластер;
- Верхоянский кластер;
- Хабаровский кластер;
- Верхнеиндигиро-Колымско-Чукотский кластер;
- Камчатский кластер.

В составе электронных приложений, соответствующих кластерам ТПИ, заполнялись следующие шаблоны:

- изученность территорий;
- месторождения и оценённые проявления ресурсов минералогической зоны;
- территории с ресурсами P_1 , P_2 , P_3 (прогнозные ресурсы, возможное нахождение полезных ископаемых в геологически слабо изученных участках земной коры и гидросферы).

По инфраструктуре, имеющей отношение к ТПИ, заполнялись следующие шаблоны:

- ГОК / ГМК / ЗИФ / ОФ;
- ТЭЦ / ГРЭС / ГЭС;
- железные дороги;
- порты;
- аэропорты.

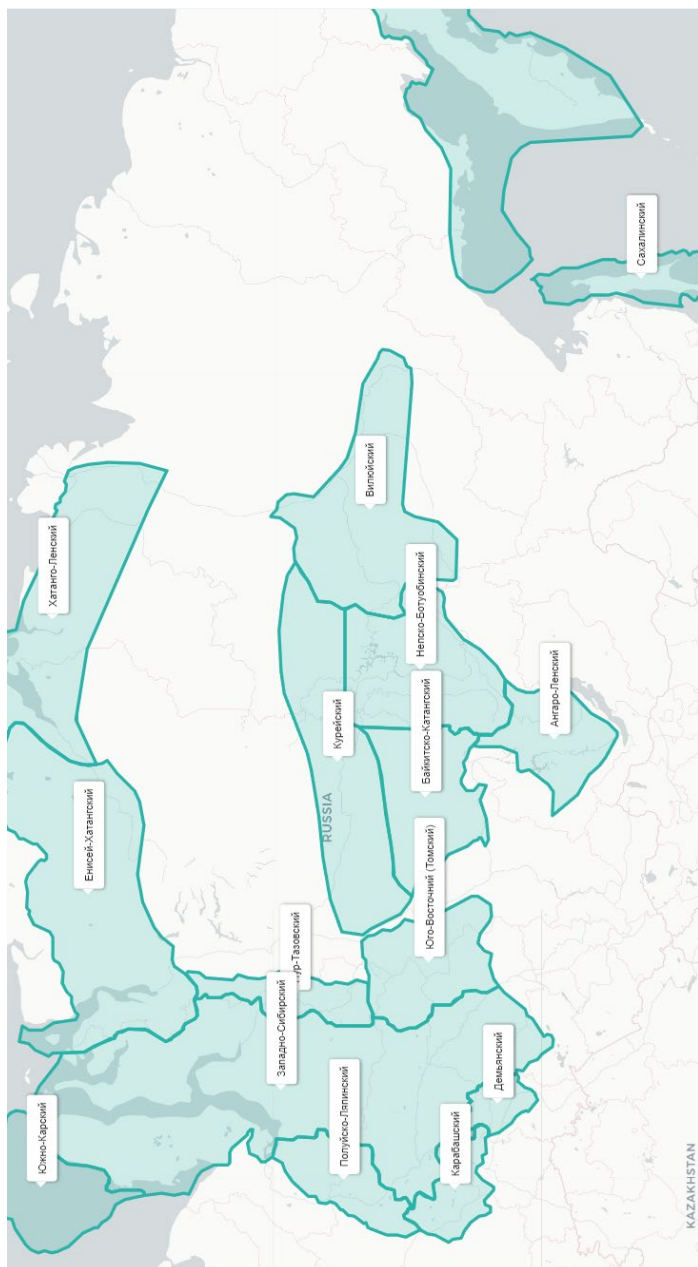


Рис. 2.6. Кластеры углеводородного сырья

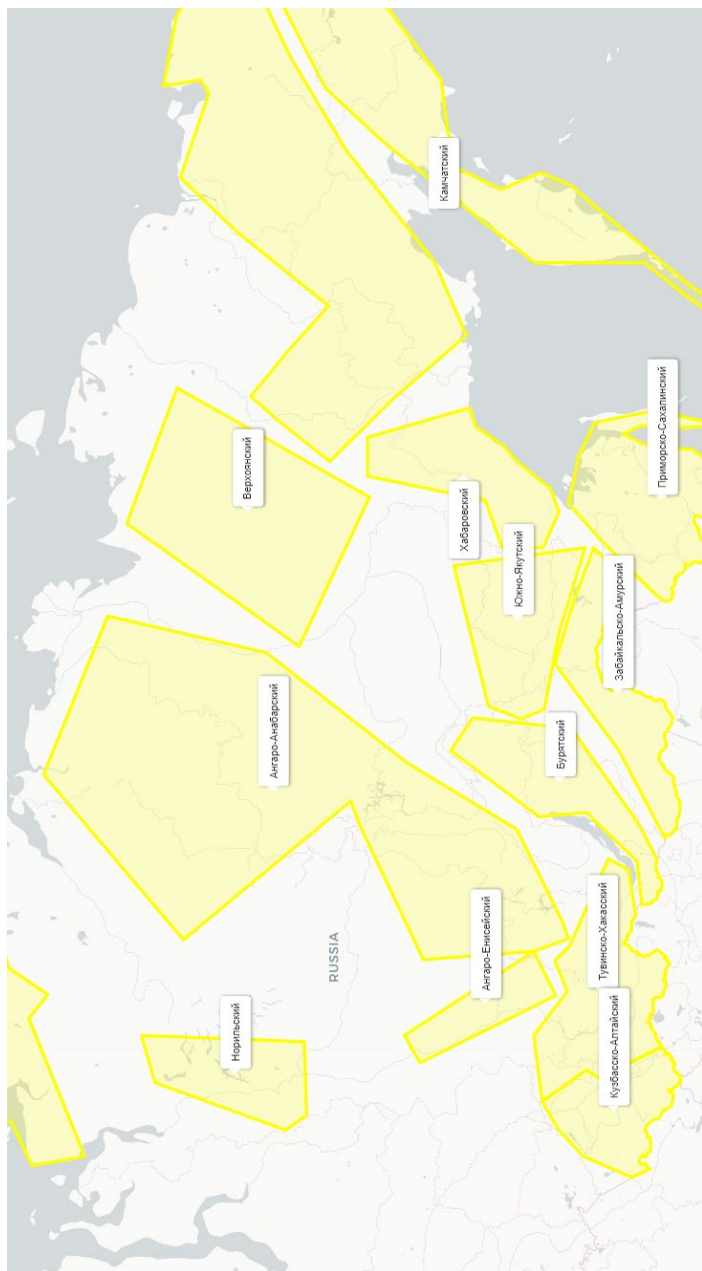


Рис. 2.7. Кластеры твердых полезных ископаемых

Блок базы данных по подземным водам был представлен в формате следующих шаблонов:

- общие сведения о ресурсах, запасах, добычи в ФО,
- источники водоснабжения,
- потребность в источниках водоснабжения,
- питьевые и технические воды,
- минеральные воды,
- термальные воды,
- пароводяная смесь,
- промышленные воды,
- месторождения питьевых и технических подземных вод,
- месторождения минеральных подземных вод,
- месторождения теплоэнергетических (термальных) подземных вод,
- месторождения теплоэнергетических (пароводяная смесь) подземных вод,
- месторождения промышленных подземных вод.

Все описанные сырьевые блоки строятся в формате взаимосвязи файлов ГИС-полигонов месторождений и описывающих их таблиц, хранящихся в общей базе данных. Подобный формат хранения ускоряет процессы анализа и визуализации данных.

2.4. Блок технологий

Блок технологий представляет собой информацию о характеристиках технологических процессов нефтепереработки и переработки твердых полезных ископаемых. Работа с информацией о технологических процессах предполагает два вида хранения данных: структурированные и неструктурированные данные.

Хранение неструктурированных данных о характеристиках технологических процессов предполагает файловое хранение отчетов, журналов и материалов с детальным описанием метаданных для удобного поиска информации и доступом через Веб интерфейс.

Хранение структурированных данных предполагает жесткие формы матричного представления, которые имеют следующие столбцы:

- Отрасль,
- Продукт,
- Промышленный способ получения,
- Применяемые технологии,
- Входные ресурсы,
- Дополнительные выходные ресурсы,
- Производитель в РФ (Азиатская часть России),
- Расположение информации,
- URL.

На текущее время реализованы два источника данных.

Первым источником являются отчеты Института Катализа¹.

Вторым источником данных является каталог Информационно-технических справочников по наилучшим доступным технологиям Росстандарта². Он включает в себя детальный анализ технологий в следующих областях:

- ИТС 2-2019 «Производство аммиака, минеральных удобрений и неорганических кислот»,
- ИТС 3-2019 «Производство меди»,
- ИТС 9-2020 «Утилизация и обезвреживание отходов термическими способами»,
- ИТС 10-2019 «Очистка сточных вод с использованием централизованных систем водоотведения поселений, городских округов»,
- ИТС 11-2019 «Производство алюминия»,
- ИТС 12-2019 «Производство никеля и кобальта»,
- ИТС 13-2020 «Производство свинца, цинка и кадмия»,
- ИТС 14-2020 «Производство драгоценных металлов»,
- ИТС 17-2021 «Размещение отходов производства и потребления»,
- ИТС 18-2019 «Производство основных органических химических веществ»,

¹ Федеральное государственное бюджетное учреждение науки «Федеральный исследовательский центр "Институт катализа им. Г.К. Борескова Сибирского отделения Российской академии наук"».

² <https://www.rst.gov.ru/portal/gost/home/activity/NDT>

- ИТС 19-2020 «Производство твердых и других неорганических химических веществ»,
- ИТС 21-2021 «Производство оксида магния, гидроксида магния, хлорида магния»,
- ИТС 22.1-2021 «Общие принципы производственного экологического контроля и его метрологического обеспечения»,
- ИТС 23-2017 «Добыча и обогащение руд цветных металлов»,
- ИТС 24-2020 «Производство редких и редкоземельных металлов»,
- ИТС 25-2021 «Добыча и обогащение железных руд»,
- ИТС 26-2021 «Производство чугуна, стали и ферросплавов»,
- ИТС 27-2021 «Производство изделий дальнейшего передела черных металлов»,
- ИТС 28-2021 «Добыча нефти»,
- ИТС 29-2017 «Добыча природного газа»,
- ИТС 30-2021 «Переработка нефти»,
- ИТС 31-2021 «Производство продукции тонкого органического синтеза»,
- ИТС 32-2017 «Производство полимеров, в том числе биоразлагаемых»,
- ИТС 33-2020 «Производство специальных неорганических химикатов»,
- ИТС 37-2020 «Добыча и обогащение угля»,
- ИТС 38-2017 «Сжигание топлива на крупных установках в целях производства энергии»,
- ИТС 40-2021 «Дубление, крашение, выделка шкур и кожи»,
- ИТС 41-2017 «Интенсивное разведение свиней»,
- ИТС 42-2017 «Интенсивное разведение сельскохозяйственной птицы»,
- ИТС 43-2017 «Убой животных на мясокомбинатах, мясохладобойнях, побочные продукты животноводства»,
- ИТС 44-2017 «Производство продуктов питания»,
- ИТС 45-2017 «Производство напитков, молока и молочной продукции»,

– ИТС 46-2019 «Сокращение выбросов загрязняющих веществ, сбросов загрязняющих веществ при хранении и складировании товаров (грузов)»,

– ИТС 47-2017 «Системы обработки (обращения) со сточными водами и отходящими газами в химической промышленности»,

– ИТС 48-2017 «Повышение энергетической эффективности при осуществлении хозяйственной и (или) иной деятельности»,

– ИТС 49-2017 «Добыча драгоценных металлов»,

– ИТС 50-2017 «Переработка природного и попутного газа».

При выходе новых справочников они будут обновляться с корректировкой структурированных данных, ссылающихся на них. В целом часть структурированных данных можно назвать сформированными и обработанными мета-данными (данными описывающими другие данные) неструктурированных источников.

2.5. Оптимизация и ускорение работы с данными

В разрабатываемой ИС более сотни таблиц, некоторые из них содержат большое число записей. Например, таблица таможенных данных – 18934617 строк; компаний – 12028209 строк; бухгалтерская отчетность – 590821434 строк.

С целью ускорения выполнения запросов к базе данных была проведена оптимизация информационной системы.

На первом этапе было решено провести индексирование таблиц. Обязательным является индексирование всех внешних ключей таблицы, что позволяет ускорить выборку в таблице компаний до приемлемого значения. Итоговое ускорение выборки составило примерно в 3–5 раз. Для остальных таблиц индексирование не позволило создать приемлемое время выборки.

На втором этапе оптимизации было решено изменить стандартные настройки СУБД и сервера. Данный прием позволил ускорить выборку незначительно, порядка от 3 до 10%.

В связи с этим было принято решение о применении партиционирования таблиц (разбиение таблиц, содержащих большое количество записей, на логические части по неким выбранным

администратором критериям) с таможенными данными и финансовой отчетностью. Данный метод вносит наименьшие изменения в структуру данных. Партиционирование таблиц было решено проводить по годам, исходя из анализа истории запросов. Эффект ускорения будет максимальный, если пользователь будет запрашивать данные только одного года. В этом случае метод партиционирования позволил ускорить выборку данных примерно в 30 раз. В случае других аналитических запросов, требующих весь временной период, эффект будет меньше.

2.6. Автоматизация сбора данных

Сбор данных является сложной и затратной по времени процедурой, но в случае периодического использования одного источника процесс сбора можно автоматизировать.

Ключевым элементом потокового сбора информации является компонент, реализующий процедуру ETL (Extract, Transform, Load – «извлечение, преобразование, загрузка») [Vassiliadis и др., 2005]. В процессе данные из нескольких систем-источников проходят несколько стадий преобразования и попадают в систему-приёмник (рис. 2.8). Одним из инструментов, реализующим данный подход, является программное обеспечение с открытым исходным кодом Airbyte и Airflow.

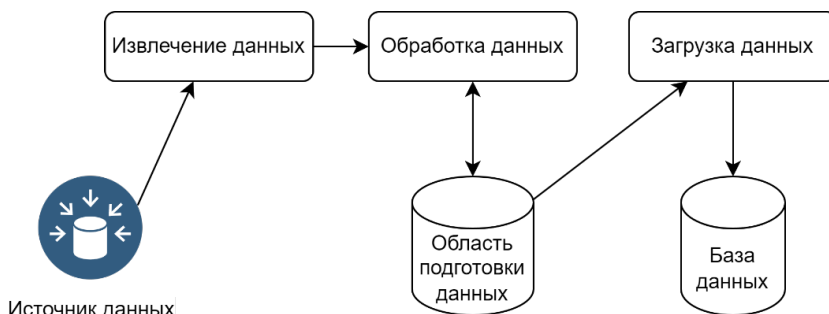


Рис. 2.8. Модель обработки данных ETL

Проведенный анализ Airbyte показал, что оно подходит для загрузки данных из широкого спектра источников. В то же время возможности по преобразованию данных после их получения несколько ограничены. Так, имеющееся средство трансформации (dbt) ориентировано на проведение операций для подготовки информации к последующему анализу, содержит средства предварительной оценки качества данных, их согласованности, удовлетворения определенных условий и т.д. В то же время dbt лишен конструкций для полноценного формирования таблиц и связей в реляционных СУБД, что затрудняет подготовку данных для их последующего хранения.

В связи с этим становится целесообразным использование Apache Airflow [Harenslak, Ruiter, 2021].

Глава 3

ФОРМИРОВАНИЕ МОДЕЛЬНОГО АППАРАТА БАЗЫ ЗНАНИЙ

Развитие БЗ происходит не только с накоплением новых данных, но также с развитием модельного и аналитического аппарата, позволяющего генерировать и фиксировать новые знания, которые за счет свойства кумулятивности могут являться основой создания последующих слоев знаний.

В этом разделе рассматриваются методы, модели и подходы, внедрение которых способствуют формированию новых знаний.

3.1. Модели КАМИН и ОМВЕАР

В рамках создания БЗ происходит инкорпорирование модельного комплекса в общую структуру БЗ, что также предполагает перенос уже существующего модельного аппарата, сложившегося и используемого в исследованиях Института экономики и организации промышленного производства Сибирского отделения Российской академии наук (ИЭОПП СО РАН). Итоговыми языками программирования выбраны Python и R, как базовые языки программирования БЗ. Перенос предполагает перепрограммирование, прямое связывание с БД и разработку подходов к фиксации новых знаний, возникающих в результате использования моделей. Это повысит их гибкость и простоту использования.

На первых этапах планировался перенос моделей КАМИН (комплексный анализ межотраслевой информации) и ОМВЕАР (оптимизационная модель взаимодействия Европейской и Азиатской России).

3.1.1. Общее техническое описание модели КАМИН

В системе КАМИН реализована трехфакторная оптимизационная межотраслевая динамическая модель экономики страны с распределенными строительными лагами и дискретным временем. Производственными факторами в модели являются: 1) основные производственные фонды, 2) сырье и материалы, 3) трудовые ресурсы.

Для ресурсов отрасли строится балансовое соотношение, из которого получаем основные балансовые соотношения для произведенных и использованных продуктов отраслей.

Предполагается, что структуризация экономики и нумерация отраслей в модели выбраны так, что отрасли можно разделить на два подразделения. Первое подразделение – производство средств производства, включая услуги, формирующие промежуточное потребление (причем внутри первого подразделения выделяются фондосоздающие отрасли, которые являются материальной основой инвестиционного комплекса страны, и сырьевые отрасли). Второе подразделение – производство предметов потребления, товаров длительного пользования, средств потребления и услуг населению.

Такое разделение отраслей в межотраслевой модели на группы связано с раздельным описанием воспроизводства каждого производственного фактора. Для каждой из групп отраслей балансовое соотношение модифицируется в соответствии со спецификой воспроизводственного процесса.

Балансовая модель используется для отладки базового варианта и исследования влияния численности занятых на траекторию.

Модель локальной оптимизации используется для исследования влияния инвестиционного комплекса на траекторию.

Гибридная модель используется для исследования совместного влияния инвестиционного комплекса и трудовых ресурсов на траекторию.

Различия балансовой модели и модели локальной оптимизации в системе КАМИН заключаются в методах использования следующих массивов при расчете траектории (табл.3.1): 1) численность занятых (трудоемкости), 2) инвестиции, 3) ввод основных фондов (ОФ) и 4) фондоемкости.

Таблица 3.1

**Различия балансовой модели и модели локальной оптимизации
в системе КАМИН**

Показатель	Источник данных	Характеристика данных
<i>Балансовая модель</i>		
1. Численность занятых в народном хозяйстве по годам	Берутся из БД	Экзогенные
2. Инвестиции в народное хозяйство по годам	Берутся из БД	Экзогенные
3. Вводы ОФ по годам	Берутся из БД	Экзогенные
4. Фондоёмкости продукта по отраслям и годам	Расчетные	Эндогенные
<i>Модель локальной оптимизации</i>		
1. Численность занятых в народном хозяйстве по годам	Расчетные	Эндогенные
2. Инвестиции в народное хозяйство по годам	Расчетные	Эндогенные
3. Вводы ОФ по годам	Берутся из БД	Экзогенные
4. Фондоёмкости продукта по отраслям и годам	Берутся из БД	Экзогенные

Для перехода от методики расчета траектории по балансовой модели к методике расчета траектории по модели локальной оптимизации в выбранной БД используется процедура пересчета матрицы фондоемкостей.

Алгоритм работы с программой

1. Пользователь указывает путь к файлу базы данных в разделе «Считывание БД из файла».
2. В разделе «Редактирование БД» есть возможность отредактировать загруженный набор данных.
3. Производится расчет показателей с помощью последовательного выполнения трех команд: «Расчет траектории по балансовой модели», «Согласование фондоемкостей», «Расчет траектории по модели локальной оптимизации».

4. Итоговые таблицы с показателями представлены в разделе «Табулятор».
5. Необходимо сохранить изменения в базе данных с помощью команды «Записать в файл». Также есть возможность выгрузки рассчитанных показателей в файл формата .xlsx (EXCEL).

Входные данные

Файл базы данных, который содержит следующую информацию:

1. Матрица материалоемкостей в базовом году периода расчетов.
2. Матрица материалоемкостей в последнем году периода расчетов.
3. Основные фонды на начало базового года.
4. Основные фонды на конец базового года.
5. Матрицы фондоёмкостей для каждого рассматриваемого года.
6. Вектор валового выпуска отраслей в базовом году.
7. Вектор трудоемкостей для каждого рассматриваемого года.
8. Сальдо (экспорт-импорт) по годам.
9. Прирост запасов продукции отраслей-производителей по годам.
10. Прочий продукт по годам.
11. Потери по годам.
12. Численность занятых в НХ по годам.
13. Структура распределенных лагов для каждого рассматриваемого года.
14. Сроки строительства для каждого рассматриваемого года.
15. Отраслевая структура вводов для каждого рассматриваемого года.
16. Вводы основных фондов в НХ по годам.
17. Схемы агрегирования отраслей.
18. Шкала второго подразделения.
19. Инвестиции в НХ по годам.
20. Структура инвестиций по видам в каждом году.

Выходные данные

Прогнозные значения по различным экономическим показателям, характеризующие свойства траекторий, на заданный период с выбранным уровнем агрегирования:

Отраслевые показатели:

- Валовой выпуск,
- Валовой внутренний продукт,
- Численность занятых.

Сырьевые показатели:

- Материальные затраты,
- Материалоемкость валового выпуска.

Инвестиционные показатели:

- Основные фонды,
- Фондоёмкость валового выпуска,
- Возмещение выбытия ОФ,
- Нормы возмещения выбытия ОФ,
- Приросты ОФ,
- Вводы ОФ.

Перераспределенные показатели:

- Основные фонды перераспределенные,
- Вводы ОФ перераспределенные,
- Инвестиции в ОК перераспределенные.

Описание Веб интерфейса модели дается в разделе 4.2, а содержательный пример использования – в разделе 5.4.

3.1.2. Общее техническое описание модели OMBEAP

На конец 2021 г. модельно-прогнозный комплекс OMBEAP 2035 реализован с помощью языка КОМБИ (Комплекс обработки матричных блоков информации). Этот комплекс создает единую программную среду, в которой размещаются различные компоненты модельно-методического комплекса и информационные массивы. Он представляет собой систему управления базами данных особого вида: на логическом уровне они служат банком информации, моделей, операций, методических схем, а также сообщений, директив и меню диалога пользователя с компьютерами.

Пользуясь его средствами, можно в диалоге реализовать любую оптимизационную модель или методическую схему комплекса, изменить старые и ввести новые элементы указанных банков. В частности, можно проводить все необходимые манипуляции с числовыми массивами, загружать модели новой информацией, строить новые модели, операции и модельные связки, генерировать различные диалоги, ориентированные на разных пользователей и разные типы задач.

Последовательность работы с Модельно-Программным Комплексом включает следующие пункты (рис. 3.1):

- 1) описание структуры ОМВЕАР (модели),
- 2) подготовка входных данных в виде числовых таблиц,
- 3) программное формирование входного файла (MPS) для оптимизатора,
- 4) программное решение задачи линейного программирования (оптимизация),
- 5) программное чтение выходного файла оптимизатора,
- 6) программный вывод результатов в виде числовых таблиц,
- 7) анализ решения.

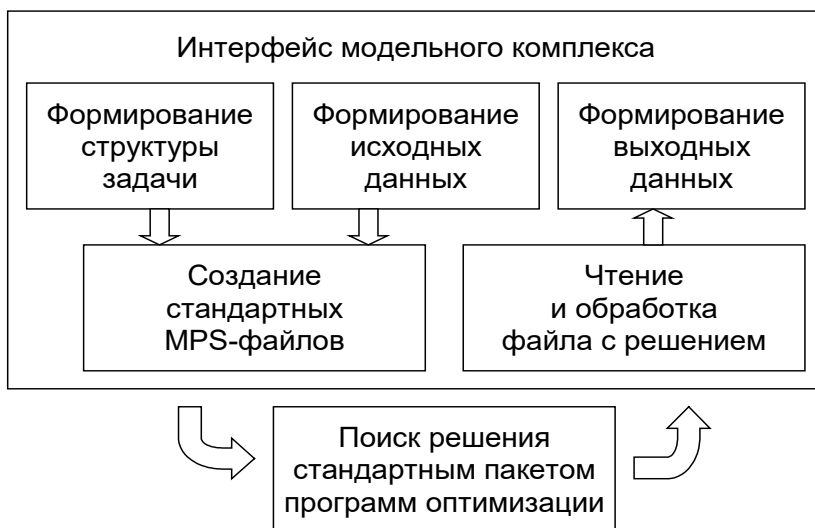


Рис. 3.1. Общая схема функционирования МПК ОМВЕАР

В последующих этапах развития базы знаний планируется поэтапное внедрение комплекса ОМВЕАР. Сейчас происходит внедрение на этапе автоматизации сбора первичных данных, далее планируется перенос кода модели на язык Python и R и прямое внедрение функционала.

3.2. Когнитивные модели

Вычислительная схема построения когнитивных моделей включает два этапа.

На первом этапе в результате анализа данных, выбранных из БД для решения поставленной задачи, происходит выделение информативной системы факторов и построение когнитивной схемы. Построение когнитивной схемы можно проводить с участием экспертов (1) либо с использованием априорной статистической информации (2). В первом случае по результатам экспертного анализа строится система взаимодействующих факторов, и оцениваются взаимосвязи между ними. То есть результатом этапа является построение ориентированного графа (знакового или взвешенного), называемого когнитивной схемой. Во втором случае – если имеется статистическая информация по взаимодействующим факторам – для построения когнитивной схемы можно использовать методы многомерного факторного анализа.

В этом случае анализ исходных данных проводится методом главных компонент (PCA, Principal Component Analysis). Данный метод представляет собой стандартный подход для анализа данных, в котором исходные данные проецируются на гиперплоскость меньшей размерности. Сильной стороной метода можно считать возможность выявления неочевидных, и скрытых поначалу, закономерностей и взаимосвязей между исходными переменными. Также он позволяет понижать размерность пространства данных и строить ортогональный базис факторов.

На втором этапе осуществляется импульсное воздействие на факторы путем задания значения в какой-либо вершине графа

с целью определения изменения в других его вершинах. Здесь, в отличие от задачи анализа, решается задача прогноза – определение, к чему приведет воздействие в сложной системе, заданное изменениями одного или нескольких управляющих факторов. Таким образом, проводится моделирование распространения управляющего воздействия по системе, заданной когнитивной схемой, в результате которого изменяется состояние системы (значения факторов).

Когнитивное моделирование применяется при формировании сценариев развития изучаемых хозяйственных систем, оценок влияния инвестиционных импульсов на экономический рост, а также решения других прогнозных задач.

Описание интерфейса модели и пример будут описаны в разделах 4.2.2 и 5.1.

3.3. Система анализа и прогнозирования межрегиональных и межотраслевых взаимодействий на основе анализа больших данных

Динамичное развитие цифровых технологий привело к появлению новых способов получения и распространения информации, которую необходимо хранить, обрабатывать, анализировать, а также строить прогнозы на основе проведенного анализа. Большие данные и способы работы с ними для XXI века стали важной задачей и вызовом.

Эти тенденции нашли отражение и в экономических процессах. Во многих направлениях происходит активное внедрение современных методов сбора, обработки и анализа данных, что повышает качество построенных математических моделей и увеличивает точность прогнозирования.

В рамках развития базы знаний началось формирование системы анализа и прогнозирования данных Азиатской части России в разрезе макро-, мезо- и микроуровней с использованием методов машинного обучения.

В качестве основы создания системы анализа и прогнозирования используются базы данных экономических показателей регионов Азиатской части России.

Программная часть выполнена на языке Python и R с использованием библиотек для анализа данных с помощью машинного обучения. Для получения информации из базы данных будут использованы SQL-запросы. Внедрение реализованной системы поиска взаимосвязей и прогнозирования в веб-интерфейсе – на основе Shiny.

В зависимости от объема и характеристических особенностей данных для решения вопросов прогнозирования и нахождения связей между факторами можно использовать различные подходы – как эконометрические, так и подходы, связанные с машинным обучением и нейронными сетями.

Методы машинного обучения

Использование методов машинного обучения значительно расширяют результативность экономических исследований. Однако следует отметить, что использование простых готовых алгоритмов машинного обучения не всегда корректно. Для эффективного решения задач в большинстве случаев требуется произвести тщательную настройку параметров алгоритмов машинного обучения.

Базовые понятия искусственного интеллекта и машинного обучения были продемонстрированы американским ученым А.Л. Самуэлем. В 1959 г. он изобрел первую самообучающуюся программу «Checkers-playing». В это же время он дал первое определение машинного обучения, как процесса, который позволяет компьютерам учиться выполнять что-либо без написания кода в явном виде [Samuel, 1959].

Современное машинное обучение можно определить как область знаний, исследующую алгоритмы, которые обучаются на данных с целью выявления закономерностей. Для реализации таких алгоритмов широко используются методы математической статистики, теории графов, теории вероятностей и другие [Флах, 2015].

Выделяют четыре основных (классификация, восстановление регрессии, ранжирование, кластеризация) и два вспомогательных (уменьшение размерности, выявление аномалий) типа задач машинного обучения.

Целью задачи *классификации* является обучение программы находить класс, к которому принадлежит объект. При этом количество классов неограниченно, начиная с 2, и, также, классы могут как не иметь общих объектов, так и пересекаться (в случае решения задачи обучения с учителем).

При прогнозировании на основе имеющихся данных решается задача *восстановления регрессии*. Методами решения данной задачи являются построение линейной регрессии, определение дерева решений и метод опорных векторов.

Задачи типа *«ранжирование»* решают проблему получения множества кластеров по заданному множеству объектов, согласно заданному отношению порядка. Основным примером использования метода является выдача поискового запроса. Решается данная задача методами поточечного, попарного или списочного подхода [Burges и др., 2005].

Способы *кластеризации* разбивают выборку на подвыборки так, чтобы объекты из каждой подвыборки были более схожи друг с другом, чем с объектами из другой подвыборки по какому-либо заранее обозначенному критерию. Для разбиения на кластеры возможно использование параметрической модели, методов группировки точек по какой-либо мере расстояния и методов теории графов [Ben-Hur и др., 2001; Hart и др., 2000].

В качестве методов кластеризации применяются: иерархическая кластеризация, эволюционная кластеризация и ЕМ-алгоритм.

Вспомогательные типы задач не несут в себе особенного смысла, а являются инструментами уточнения моделей и улучшения качества данных.

В зависимости от наличия «учителя» машинное обучение возможно разделить на 4 категории алгоритмов обучения:

- с учителем,
- с частичным привлечением учителя,
- без учителя,
- с подкреплением.

В случае обучения с учителем компьютер обучается на примерах с заранее известными выходными данными. Программа самостоятельно устанавливает взаимосвязь между входными и выходными данными, учится на основе наблюде-

ний, а также занимается прогнозированием. Далее «учитель» (человек – пользователь программы) корректирует прогнозы и снова начинается процесс обучения программы до тех пор, пока не будет достигнут необходимый уровень точности прогнозов. Наиболее распространёнными алгоритмами обучения с учителем являются:

- регрессии: линейная и логистическая,
- KNN (метод k-ближайших соседей),
- SVM (метод опорных векторов),
- CART (метод построения бинарных деревьев решений).

Данная категория алгоритмов используется на практике для решения задач регрессии и классификации [Полетаева, 2020].

Обучение без учителя охватывает широкий класс задач обработки данных, где есть лишь обучающая (исходная) выборка, и программе необходимо найти связь между объектами. В этом случае тренировочные данные находятся сразу в открытом доступе, но не существует заранее известного ответа для поставленной задачи.

Алгоритм обучения с частичным привлечением учителя на вход получает прецедент в виде пары «объект, ответ», но ответы известны только на малую часть прецедентов.

Частным случаем обучения с учителем является обучение с подкреплением. В этом методе правильность ответа оценивается не учителем, а средой, в которой программа обучается.

Для каждой рассматриваемой модели в литературе предлагаются огромное количество вариаций, и включить их все не представляется возможным. Базовые версии наиболее популярных моделей представлены в работе [Ahmed и др., 2010]. К ним относятся следующие.

А) Метод ближайших соседей

Метод ближайших соседей является одним из наиболее простых методов как для понимания, так и для реализации. Основная идея данного метода заключается в классификации объектов на основе их сходства.

Алгоритм метода ближайших соседей пропускает предварительную фазу настройки модели, сразу переходя к непосредственному обучению, которое представляет из себя классифика-

цию данных на основе большинства голосов ближайших соседей. Объекту назначается тип, который набрал наибольшее количество голосов среди соседей [Полетаева, 2020].

Б) Случайный лес

Случайный лес представляет из себя ансамбль решающих деревьев, которые представляют из себя графы решений (могут быть бинарными). В ходе работы алгоритма при решении задачи классификации осуществляется расчет прогнозных результатов для каждого из деревьев и выбирается класс, который набрал большее количество голосов.

В рамках решения задачи регрессии высчитывается среднее значение всех итоговых результатов ансамбля.

В) Метод опорных векторов

Алгоритм метода опорных векторов заключается в создании гиперплоскости в векторном пространстве для разбиения его на классы с максимальной граничной областью. То есть, данные отображаются из пространства данных в многомерное пространство признаков с помощью ядра Гаусса, которое представляет собой связь между двумя векторами. Для создания гиперплоскости на вход подается набор обучающих примеров, по которым строится модель. Классификация каждой новой точки данных происходит соответственно расположению точки относительно гиперплоскости [Bishop, 1995].

Нейронные сети

Искусственные нейронные сети являются альтернативным методом для работы с данными, основанным на построении и оценке математических моделей. Главное отличие нейронных сетей заключается в способности к самообучению за счет корректировки весовых параметров функций скрытого слоя. В процессе обучения нейронная сеть сравнивает выходные параметры с обучающими данными, и на основании разницы между ними производит корректировку параметров. В большинстве моделей нейронных сетей для коррекции параметров используется метод наименьших квадратов (МНК) в различных вариациях: классический, рекуррентный и взвешенный.

Далее будет произведен обзор наиболее популярных моделей нейронных сетей [Ahmed и др., 2010].

А) Многослойный линейный персептрон (MLP)

Многослойный персептрон (часто называемый просто нейронной сетью), пожалуй, является самой популярной сетевой архитектурой, используемой сегодня как для решения задач классификации, так и для оценки регрессии [Bishop, 1995]. Данная сеть содержит несколько скрытых слоев, причем нейроны одного слоя между собой никак не связаны, но при этом каждый нейрон каждого слоя связан с каждым нейроном соседнего слоя.

Б) Байесовская нейронная сеть (BNN)

Модель Байесовской нейронной сети образуется путём объединения нейронных сетей и вероятностного моделирования.

В качестве функции ошибки в данной сети используется функция правдоподобия, а функция активации строится на основе формулы Байеса. Огромным преимуществом данного метода является то, что регуляризация хорошо включается в BNN, что позволяет избежать проблемы переобучения и излишней сложности модели. Обработка параметров регуляризации может быть осуществлена последовательно в байесовской среде без дополнительного привлечения методов перекрестной валидации [Bishop, 1997].

BNN по своей сути является вероятностной моделью, к которой добавляется нейронная сеть как универсальный аппроксиматор функций.

В) Сеть радиальных базисных функций (RBF)

Сеть радиальных базисных функций похожа по архитектуре на многослойную нейронную сеть, основное отличие заключается в узлах сети – они имеют локализованную функцию активации. Чаще всего узловые функции выбираются как гауссовы функции, ширина которых определяет плавность приближения функции. Для получения конечного результата выходные данные узлов линейно объединяются [Moody, Darken, 1989; Powell, 1987].

Проделанный краткий анализ позволяет заключить, что использование алгоритмов машинного обучения, наиболее предпочтительно в решении задач прогнозирования в условиях сла-

бо формализуемых ситуаций. Также отдельно стоит отметить, что машинное обучение и нейронные сети помогают находить скрытые взаимосвязи в данных, которые не могут быть обнаружены в ходе статистического анализа. При этом для корректной работы алгоритмов машинного обучения и нейронных сетей ключевым вопросом является качество предоставленных данных, так как именно на них и строится модель. В связи с этим перед непосредственным использованием алгоритмов необходимо провести разведывательный анализ данных и предварительную их обработку.

Первичное преобразование данных

В рамках формирования системы анализа и прогнозирования межрегиональных и межотраслевых взаимодействий была создана модель первичных преобразований данных для их последующего использования.

Для формирования исходного набора данных были использованы данные базы знаний экономических показателей регионов Азиатской России, описанные в главе 2.

Большинство алгоритмов машинного обучения построены на основе статистики, нестационарность данных может привести к неправильной параметризации модели машинного обучения. В ходе обучения данные разделяются на две подвыборки: тестовую и тренировочную. В связи с этим, при наличии нестационарности в данных распределение в этих двух наборах данных может существенно различаться, что значительно образом снижает качество моделей и точность прогнозов.

Анализ на стационарность проводился с помощью расширенного теста Дики-Фуллера. Данный метод проверки используется в эконометрике и является тестом на наличие единичных корней. Тест был предложен и разработан американскими экономистами Дэвидом Дики и Уэйном Фуллером в 1979 г. Расширенный тест Дики-Фуллера получается путем включения в регрессию лагов первых разностей временного ряда.

В рамках анализа на стационарность было выявлено, что практически все переменные по всем регионам являются нестационарными, что требуется учесть в дальнейшем моделировании.

В качестве следующего этапа предварительного анализа данных был проведен *анализ параметров на гетероскедастичность*.

Большие данные открывают новые возможности для анализа и прогнозирования данных. При этом игнорирование непостоянства дисперсии в модели может привести к неточным доверительным интервалам и прогнозам [Antunes и др., 2015].

И хотя учет гетероскедастичности в данных ведет к значительному усложнению моделей и вычислений, он может значительно улучшить качество прогнозов. В машинном обучении при поиске взаимосвязей в случае гетероскедастичности возникает проблема того, что некоторые наблюдения будут иметь больший вес, чем другие, что приведет к неверным выводам.

Для анализа на наличие гетероскедастичности в данных использовались следующие тесты: Голдфелда-Квандта (ГК) и Брейша-Пагана (БП) на 95%-м и 90%-м уровнях значимости.

Тест Голдфелда-Квандта проверяет пропорциональность стандартного отклонения ошибок отдельной переменной [Goldfeld, Quandt, 1965]. Тест Брейша-Пагана используется для проверки зависимости дисперсии случайных ошибок от некоторого набора эндогенных переменных в модели и линейной зависимости дисперсии случайных ошибок от совокупности факторов [Breusch, Pagan, 1979].

Данные тесты были проведены отдельно для каждого из параметров каждого из регионов. В результате чего была рассчитана матрица результатов тестов на 90%-м и 95%-м уровнях значимости. Далее была составлена обобщающая переменная-вектор, которая представляет из себя среднее значение тестов ГК и БП на 90%-м уровне значимости для проведения экспертной оценки агрегированных результатов теста.

Помимо формального теста на гетероскедастичность, был дополнительно проведен визуальный анализ на наличие экспоненциального тренда. Это было необходимо, так как мы рассматривали ежегодные данные, соответственно объем данных не являлся достаточно большим – в таком случае вероятность ошибки тестов на гетероскедастичность достаточно велика. В связи с этим возникла необходимость использования дополнительного аналитически-экспертного подхода.

Для проведения аналитики для каждого из показателей был построен график и оценки экспоненциального тренда (рис. 3.2). Помимо этого для эксперта рядом с графиком дополнительно отображаются необходимые для анализа показатели: результаты тестов на гетероскедастичность, коэффициент детерминации R^2 , количество наблюдений и единицы измерения. Данный подход использовался также для экспертного анализа качества статистических данных. В результате была создана комплексная система оценивания, в ходе которой экспертом принималось решение о необходимости логарифмирования того или иного показателя.



Рис. 3.2. График и экспоненциальный тренд показателя валовой добавленной стоимости

В табл. 3.2 в первом столбце предоставлены сводные данные стационарности показателей. Тест на стационарность проводился отдельно по каждому из регионов, затем было рассчитано среднее значение показателя относительно всех регионов. Во втором столбце представлен процент данных, которые необходимо логарифмировать.

Таблица 3.2

Стационарные данные по группам показателей, %

№	Название раздела	Стационарные данные	Показатели, которые необходимо логарифмировать
1	Население	38	45
2	Труд	12,5	16
3	Уровень жизни населения	0	21
4	Образование	0	12
5	Здравоохранение	0	14
6	Культура, отдых и туризм	0	18
7	Земельные ресурсы и охрана окружающей природной среды	2,4	7
8	Валовой региональный продукт	0	22
9	Основные фонды	2	54
10	Инвестиции	0	65
11	Предприятия и организации	0	17
12	Промышленное производство	0	15
13	Сельское хозяйство	5,8	24
14	Строительство	2,8	35
15	Торговля и услуги населению	0	18
16	Транспорт	0	15
17	Информационные и коммуникационные технологии	0	23
18	Наука и инновации	0	14
19	Финансы	1,7	65
20	Цены и тарифы	0	89
21	Внешняя торговля	0	92
22	Государственное устройство РФ	0	14
23	Преступность	0	12
24	Система национальных счетов	0	29

Полученные результаты позволяют произвести первичные преобразования данных для последующего использования машинного обучения и нейронных сетей.

Следующим шагом является первичное *выявление долгосрочных взаимосвязей показателей*, вследствие чего возникла необходимость определения коинтегрированности параметров. Для этого был использован тест Педрони на коинтеграцию, а также матрица связанных показателей на основе теста Гренджера, в которой определена статистически значимая связь между соответствующими показателями, где один показатель является причиной, а другой – следствием. Нулевая гипотеза об отсутствии коинтеграции проверялась на уровне значимости 0.1. В результате были сформированы коинтеграционные и корреляционные матрицы взаимосвязей показателей.

Затем было проведено моделирование и прогнозирование показателей с использованием алгоритмов машинного обучения и нейронных сетей. В модельный пул были включены следующие модели машинного обучения и нейронных сетей:

- Рекуррентная нейронная сеть (RNN),
- Рекуррентная нейронная сеть с долгой краткосрочной памятью (LSTM),
- Градиентный бустинг (Gradient boosting).

Проведенный анализ качества прогноза показал преимущество LSTM моделей почти для всех кейсов региональных данных. В дальнейшем основной акцент будет направлен на улучшение работы модели, ускорение интерфейса пользователя и формирование новой сопроводительной базы метаданных.

В результате была разработана система автоматического поиска взаимосвязей и прогнозирования, которая была реализована в виде программного продукта. Она обладает высокой степенью автоматизации и позволяет более эффективно находить взаимосвязи между различными переменными и предсказывать будущие прогнозные значения на основе имеющихся данных.

Глава 4

ИНТЕРФЕЙС ПОЛЬЗОВАТЕЛЯ БАЗЫ ЗНАНИЙ

Формирование БЗ предполагает использование пользовательского интерфейса, через который будет происходить эксплуатация системы и её пополнение. БД PostgreSQL имеет встроенный пользовательский интерфейс в виде pgAdmin и DBeaver. Они подходят для администрирования БЗ, но требуют достаточно высокого уровня квалификации пользователя, что может представлять проблему для обычного научного сотрудника, не обладающего необходимыми квалификацией и навыками.

В целях преодоления этого затруднения было принято решение создать собственный пользовательский интерфейс.

4.1. Техническое задание к интерфейсу пользователя базы знаний ИЭОПП СО РАН

4.1.1. Функциональные части системы

Система состоит из двух основных разделов: служебного и основного. Служебный раздел предназначен для авторизации. Основной раздел – для работы, собственно, с базой знаний. Система предполагает доступ только авторизованных пользователей.

Служебный раздел предусматривает выполнение следующих операций:

- Регистрация, вход в систему, восстановление пароля,
- Внесение / изменение данных о пользователе.

Основной раздел обеспечивает доступ:

- к каталогу данных,
- страницам блоков данных,
- системе поиска,
- к действиям с данными.

4.1.2. Каталог данных

Каталог – один из способов структурированного доступа к данным в системе. Каталог содержит список всех данных, до-

ступных конкретному пользователю. Список не фиксирован, меняется в следующих ситуациях:

- поступление новых данных,
- удаление устаревших данных,
- разрешение / запрет пользователю доступа к конкретному блоку данных.

Необходимо обеспечить механизм регулирования доступа пользователей к различным блокам данных.

Информация по каждому блоку данных должна быть представлена на отдельной веб-странице. Страница содержит метainформацию и непосредственно данные из этого блока (либо способ их получения). Необходимо предусмотреть постраничный вывод для данных больших объемов.

В зависимости от типа, данные на веб-странице требуется представлять в виде таблиц, графиков, карт и т.д.

4.1.3. Поиск и фильтрация

Система должна предоставлять гибкий поиск и фильтрацию содержимого.

- Полнотекстовый поиск по:
 - названию,
 - описанию,
 - внутреннему содержанию.
- Фильтрацию по характеристикам:
 - категории каталога,
 - экономическим (отрасль, производственный комплекс, предприятие, др.),
 - временным (год, месяц, квартал и т.д.),
 - пространственным (регион, федеральный округ, экономический район и т.д.),
 - код (ФИАС, ОКТМО, ОКАТО и т.д.),
 - произвольная область на карте,
 - прочим характеристикам, специфичным для конкретного вида данных.
- Прочее:
 - теги,
 - ключевые слова,
 - другое.

4.1.4. Операции с данными

Система должна позволять осуществлять следующие операции с данными:

Первый этап:

- Экспорт для локальной работы,
- Визуализация,
- Копирование в пользовательское пространство.

Второй этап:

- Аналитические запросы,
- Моделирование:
 - Экономические модели,
 - ГИС-модели.

4.1.5. Экспорт

У пользователя должна быть возможность выгрузить данные из системы.

Нужно предусмотреть выгрузку в форматы: CSV, XLSX. Возможно появление других форматов. Выбор конкретного формата зависит от типа данных.

Необходимо предусмотреть пользовательский интерфейс написания запросов к данным.

4.1.6. Визуализация

Важной частью системы является визуализация.

У пользователя должна быть возможность строить графики, схемы, диаграммы, отображать информацию с геопривязкой на карте и т.д. Необходимо предусмотреть их экспорт в форматы PNG, PDF. В частности, в случае ГИС-данных должна быть возможность экспорта карт с нанесенными слоями.

4.1.7. Аналитические запросы

В системе должна быть предусмотрена возможность осуществлять аналитические запросы к данным, строить отчёты по шаблонам (предопределённым или пользовательским). Важно предусмотреть пользовательский интерфейс написания аналитических запросов и построения шаблонов отчётов.

4.1.8. Моделирование

Система должна позволять хранить, извлекать и визуализировать данные для экономического и геопространственного моделирования. Должна быть предусмотрена возможность строить и рассчитывать эти модели с использованием мощностей сервера.

Моделирование включает:

- Экономико-математические модели (система OMBEAR, КАМИН и др.),
- ГИС-моделирование (кластеризация, суммирование по области и др.),
- Модели анализа данных (поиск корреляционных, регрессионных и прочих взаимосвязей),
- Прогнозные модели (подбор оптимальной прогнозной модели на основе регрессионных моделей, машинного обучения, нейронных сетей).

Анализ существующих решений по реализации интерфейса пользователя привел к разделению интерфейса пользователя на две взаимосвязанные части: ГИС-интерфейс на базе NextGIS и Веб-интерфейс (рис. 4.1)

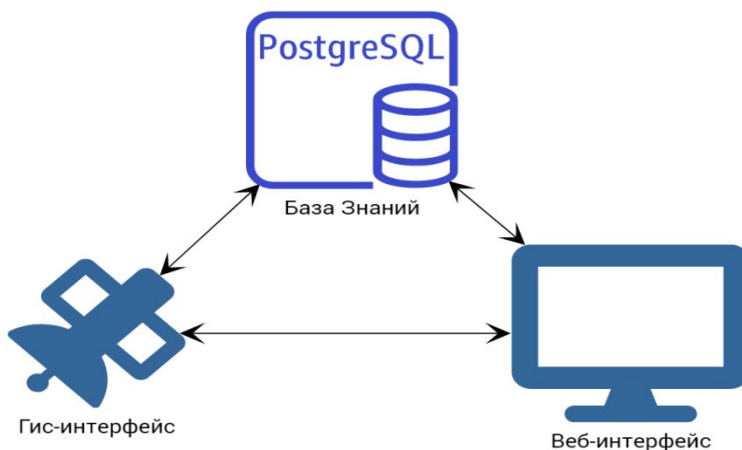


Рис. 4.1. Схема интерфейса пользователя Базы Знаний

Проблема ГИС-интерфейса NextGIS заключается в том, что он позволяет работать только с визуализацией, но не с моделированием. Поэтому он будет использоваться только для отображения больших, сложно обрабатываемых слоёв, например – автомобильных и железных дорог. ГИС-моделирование будет проводиться на основе системы Веб-интерфейса на языке R.

4.2. Веб-интерфейс Базы Знаний

Для реализации Веб-интерфейса БЗ был проведен анализ имеющихся технологий реализации Веб-интерфейсов. По результату анализа была выбрана платформа Shiny. Shiny – это пакет языка программирования R, который позволяет создавать интерактивные веб-приложения прямо из R. Он позволяет размещать автономные приложения на веб-странице или создавать информационные панели.

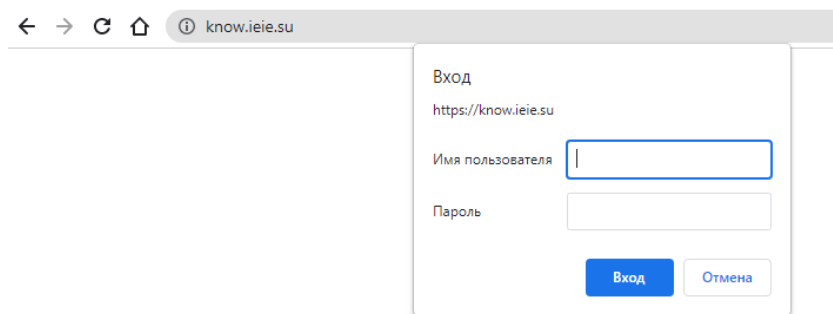
Базовыми достоинствами платформы Shiny, необходимыми для реализации веб-интерфейса базы знаний, являются:

- Гибкость и широкие возможности, удовлетворяющие технические требования;
- Сопряженность с языками программирования R, Python, SQL и html, на которых реализуется базовый модельный аппарат БЗ;
- Возможность установки на локальный сервер, сопряженность с БД и способность осуществлять серверные вычисления в самом интерфейсе;
- Совместимость с ГИС-интерфейсами;
- Широкие возможности отлаживания интерфейса локально на рабочих станциях для экономико-математического моделирования.

На данный момент реализованный веб-интерфейс включает в себя систему ГИС-моделирования, модельный блок и интерфейсы работы с данными, в том числе с панельными данными, данными инвестиционных проектов, данными компаний, таможенной статистикой и данными о технологиях. Для получения доступа к Базе Знаний ИЭОПП СО РАН необходимо пройти двойную авторизацию, как показано на рис. 4.2 и 4.3. Двойная ав-

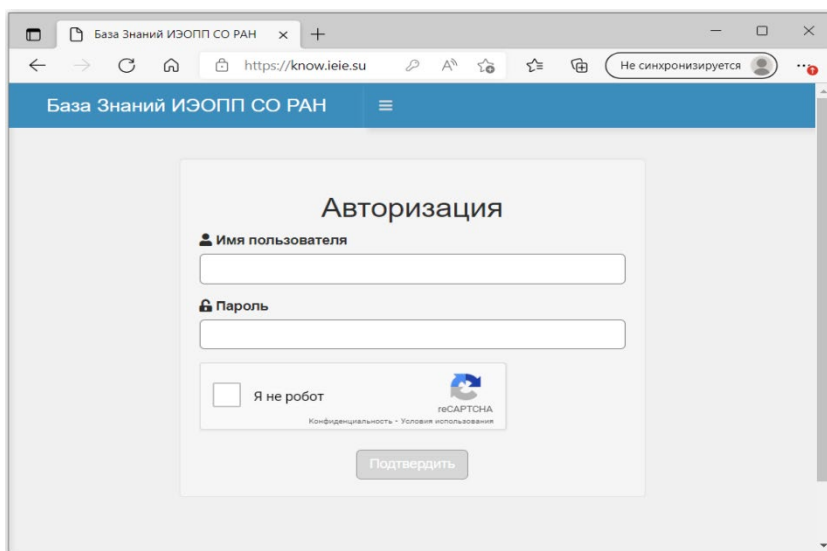
торизация призвана увеличить информационную безопасность системы. Для успешной авторизации необходимо ввести имя пользователя и пароль для системы Базы Знаний, а также пройти проверку с помощью reCAPTCHA.

При успешном прохождении авторизации пользователь получает доступ к Базе Знаний ИЭОПП СО РАН с помощью графического интерфейса, представленного на рис. 4.4.



The image shows a web browser window with the address bar displaying 'know.ieie.su'. A modal dialog box is centered on the screen. The dialog has a title 'Вход' (Login) and a URL 'https://know.ieie.su'. It contains two input fields: 'Имя пользователя' (Username) and 'Пароль' (Password). Below the password field are two buttons: 'Вход' (Login) in blue and 'Отмена' (Cancel) in white with a blue border.

Рис. 4.2. Страница авторизации 1



The image shows a web browser window with the title 'База Знаний ИЭОПП СО РАН'. The address bar shows 'https://know.ieie.su'. The page has a blue header with the site name and a menu icon. The main content area is titled 'Авторизация' (Authorization). It contains two input fields: 'Имя пользователя' (Username) and 'Пароль' (Password). Below the password field is a reCAPTCHA widget with the text 'Я не робот' (I am not a robot) and a 'Подтвердить' (Confirm) button. The reCAPTCHA logo and text 'reCAPTCHA Конфиденциальность - Условия использования' are also visible.

Рис. 4.3. Страница авторизации 2

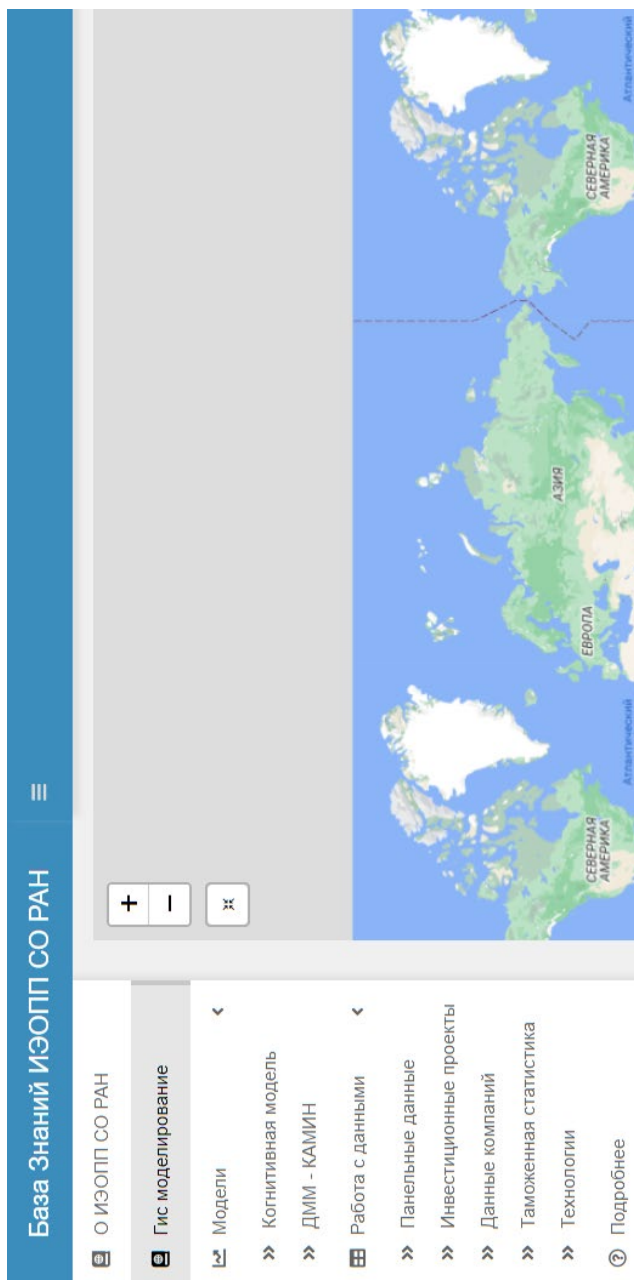


Рис. 4.4. Главный интерфейс Базы Знаний ИЗОПП СО РАН

Интерфейс Базы Знаний включает в себя общее меню в левой части экрана и содержимое выбранного раздела на основной части. Для скрытия меню можно воспользоваться соответствующей кнопкой в верхней части экрана (рис. 4.5).

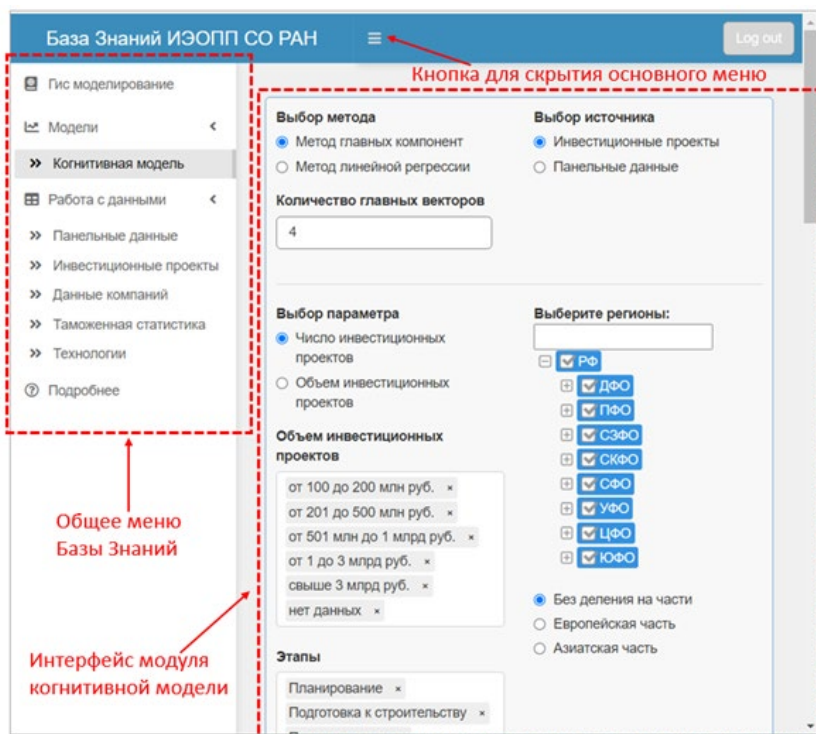


Рис. 4.5. Основные части главного интерфейса Базы Знаний

Далее будут описаны основные модули базы знаний.

4.2.1. ГИС-Моделирование

Интерфейс ГИС-моделирования состоит из двух частей: меню выбора группы слоёв и карты для отображения (рис. 4.6).

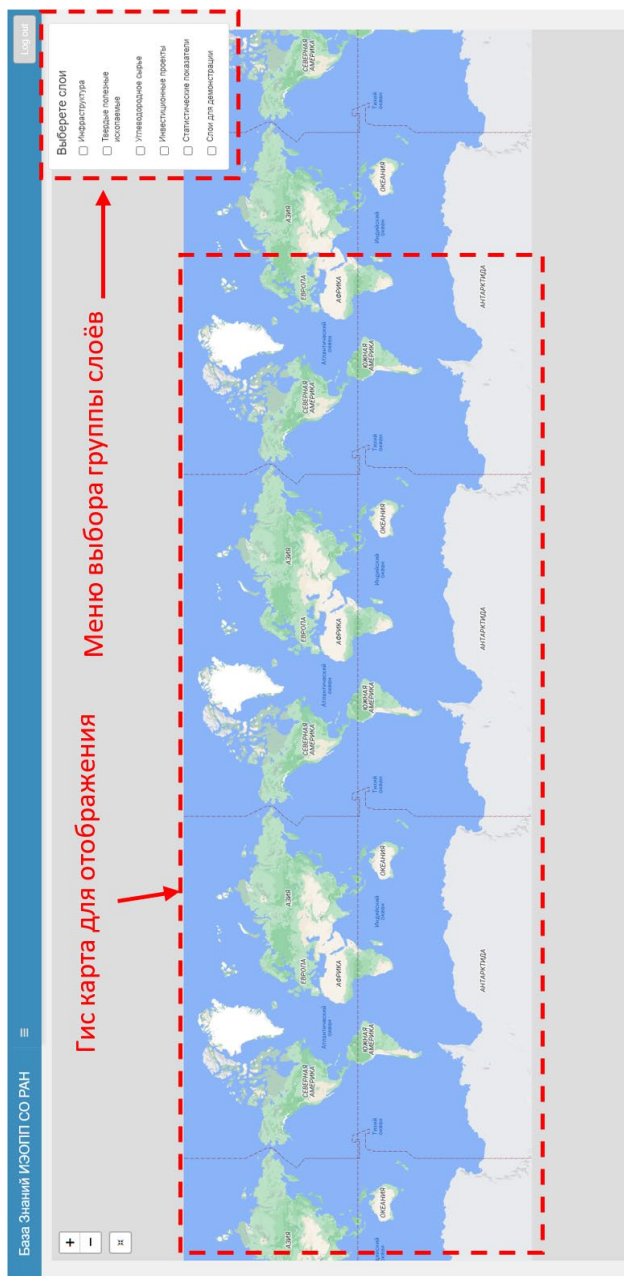


Рис. 4.6. Модуль ГИС-моделирования Базы Знаний

При выборе группы слоёв появляется дополнительное окно со списком слоёв в выбранном кластере (рис. 4.7 и 4.8).

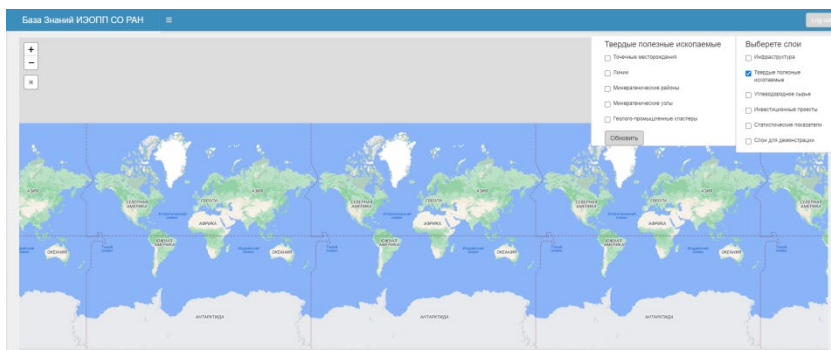


Рис. 4.7. Выбор ГИС-слоёв модуля ГИС-моделирования

Система Гис-интерфейса позволяет выбирать базовую карту (подложку), что видно на рисунках 4.7 и 4.8. Она может быть цветной или черно-белой, может быть базовой от Яндекс, OpenStreetMap или Google. В будущем планируется добавить в интерфейс возможность выбора подложки пользователем в виде всплывающего окна. Гис-интерфейс Базы Знаний является низкоуровневым, что позволяет не только отображать объекты на карте, но и внедрять элементы Гис-моделирования, такие как кластеризация, оптимизация, пространственная аналитика и др. (Более детально про текущее наполнение системы Гис-интерфейса будет описано в разделе 4.3.)

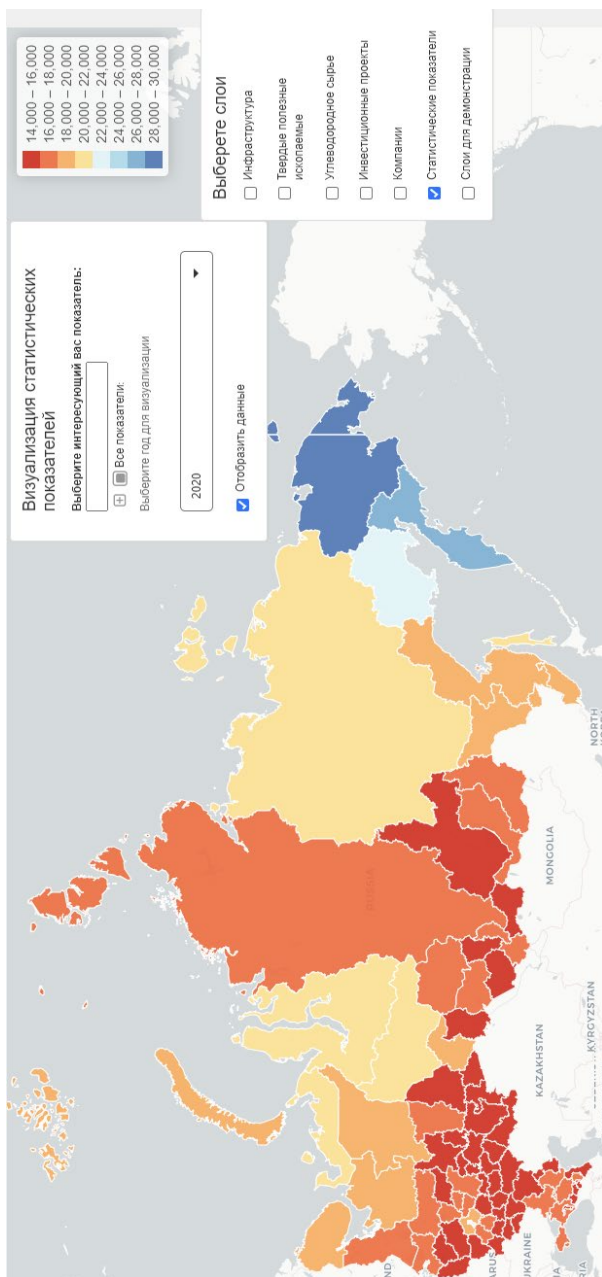


Рис. 4.8. ГИС-визуализация статистических показателей
(на примере показателя «Стоимость фиксированного набора потребительских товаров и услуг в рублях на 2020 год»)

4.2.2. Модели

Когнитивная модель

Цель инструментария когнитивного моделирования – построение модели анализа взаимосвязей в сложной системе взаимодействующих социально-экономических показателей, осуществление прогноза поведения системы в разных сценариях развития внешних воздействий и внутренних изменений, выработка предложений по управлению поведением системы для обеспечения ее устойчивого развития.

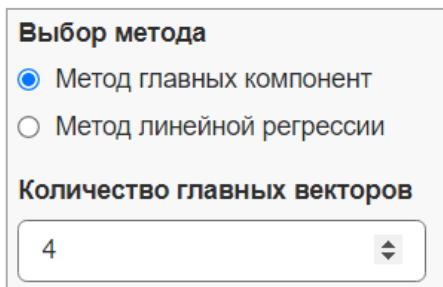
В интерфейсе модуля когнитивной модели можно выделить две панели, которые располагаются друг под другом. Вверху располагается панель задания параметров работы модели, ниже – результаты работ модели, как показано на рис. 4.9. Панель задания параметров также разделена на две части (рис. 4.9) – в верхней части располагаются общие параметры моделирования – выбор метода и задание его параметров, а также выбор источника исходных данных.

The screenshot displays the user interface of a cognitive model. The top section, titled 'Панель задания параметров моделирования' (Modeling parameter setting panel), includes a status bar with buttons like 'Приостановлен', 'Предпроектные проработки', 'Строительство', 'Введен в эксплуатацию', 'Проектирование', 'Модернизация', and 'Отменен'. Below this is a section for 'Отрасли' (Sectors) with a list of industry categories such as 'Спортивные объекты и сооружения', 'Деревообработка', 'Жилые объекты', 'Коммерческие объекты', 'Инфраструктура', 'Инженерные сети', 'Энергетика', 'Металлообработка', 'Пищевая промышленность', 'Медицинские объекты', 'АПК', 'Строительные материалы', 'Химическая промышленность', 'Переработка отходов', 'Машиностроение', 'Добыча и переработка', and 'фармацевтическая промышленность'. A radio button labeled 'Азиатская часть' is also present. The bottom section, titled 'Представление результатов' (Results presentation), contains a table with the following data:

Информация о ГВ	Таблица взаимосвязей	Импульсы	Сравнение по факторам для ГВ	
	Комп. 1	Комп. 2	Комп. 3	Комп. 4
1	0.766	0.602	0.223	0.032
2	-0.638	0.749	0.173	-0.032
3	-0.059	-0.274	0.956	-0.084
4	-0.05	-0.018	0.079	0.995

Рис. 4.9. Основные части интерфейса модуля когнитивной модели

В зависимости от выбранного метода анализа изменяется система выбора параметров модели, как показано на рис. 4.10 и 4.11.



Выбор метода

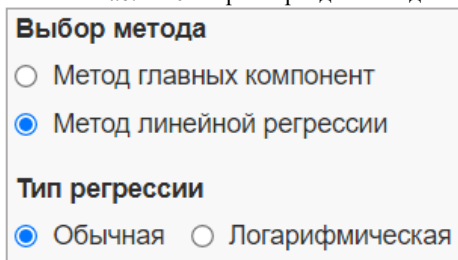
☒ Метод главных компонент

☐ Метод линейной регрессии

Количество главных векторов

4

Рис. 4.10. Параметры для метода главных компонент



Выбор метода

☐ Метод главных компонент

☒ Метод линейной регрессии

Тип регрессии

☒ Обычная ☐ Логарифмическая

Рис. 4.11. Параметры для линейной регрессии

Нижняя часть рис. 4.9. является динамично моделируемой и зависит от выбранного источника данных. В случае выбора в качестве источника данных базы данных инвестиционных проектов, пользователь может выбрать интересующие его этапы, на которых находятся инвестиционные проекты и отрасли. Обычно на шаге выбора этапов исключаются отмененные, приостановленные или введенные в эксплуатацию инвестиционные проекты. Фактором для модели будет являться объем финансирования инвестиционных проектов в выбранной отрасли по регионам (рис. 4.12). В случае выбора в качестве источника данных базы данных социально-экономических показателей пользователь может выбрать наименование интересующих его показателей из списка доступных и год, для которого будет проводиться анализ. Факторами для модели будут являться значения выбранных показателей на выбранный год по регионам.

Выбор метода

☒ Метод главных компонент
 ☐ Метод линейной регрессии

Выбор источника

☒ Инвестиционные проекты
 ☐ Панельные данные

Количество главных векторов

4

Выбор параметра

☒ Число инвестиционных проектов
 ☐ Объем инвестиционных проектов

Объем инвестиционных проектов

от 100 до 200 млн руб. ✕

от 201 до 500 млн руб. ✕

от 501 млн до 1 млрд руб. ✕

от 1 до 3 млрд руб. ✕

свыше 3 млрд руб. ✕

нет данных ✕

Выберите регионы:

☐ ☒ РФ

+ ☒ ДФО

+ ☒ ПФО

+ ☒ СЗФО

+ ☒ СКФО

+ ☒ СФО

+ ☒ УФО

+ ☒ ЦФО

+ ☒ ЮФО

☒ Без деления на части
 ☐ Европейская часть
 ☐ Азиатская часть

Этапы

Планирование ✕

Подготовка к строительству ✕

Приостановлен ✕

Предпроектные проработки ✕

Строительство ✕

Введен в эксплуатацию ✕

Проектирование ✕

Модернизация ✕

Отменен ✕

Рис. 4.12. Панель задания параметров модели

При изменении параметров происходит автоматический пересчет модели.

Область представления результатов, расположенная в нижней части интерфейса, разделена на вкладки (см. рис. 4.9).

После выбора пользователем вида модели и источника данных, с которыми он желает взаимодействовать, происходит расчет модели.

В случае использования метода главных компонент будут проходить следующие этапы моделирования. Сначала будет построена модель главных компонент (векторов) по выбранным данным, и построены факторы, количество которых пользователь выбрал при выборе модели. Для того чтобы модель работала, важно чтобы выбранное количество отраслей превышало вы-

бранное количество главных векторов (факторов). Факторы являются линейными комбинациями выбранных показателей и формируют таблицу взаимосвязей, которая является основной частью когнитивного моделирования. Как только таблица взаимосвязей будет построена, пользователь может её скорректировать и использовать для импульсного моделирования, которое представляет собой подачу импульсов в различные полученные главные вектора модели (факторы) и фиксирование отклика системы в целом.

В случае использования метода линейных регрессий строится система линейных уравнений и по полученным коэффициентам формируется таблица взаимосвязей. В этом случае каждая точка когнитивной модели является не безразмерным вектором, а конкретным показателем. После этого так же, как в методе главных компонент, происходит импульсное моделирование в выбранные пользователем показатели и фиксация отклика всей системы. Сама система линейных уравнений, из-за сложности процесса для пользователей, реализована без добавления независимых факторов, которые необходимы для правильного схождения системы линейных уравнений и использования 2- или 3-шагового обобщенного метода наименьших квадратов. Поэтому предполагается экспертная корректировка полученных коэффициентов в таблице взаимосвязей для корректной работы когнитивной модели.

Для демонстрации работы модуля используем метод главных компонент для базы данных инвестиционных проектов. Выберем демонстрационный набор инвестиционных проектов на всех этапах, кроме отмененного, приостановленного и введенного в эксплуатацию по следующим отраслям:

- Строительство и реконструкции бизнес-центров;
- Строительство жилых комплексов;
- Строительство гостиничных комплексов;
- Строительство и реконструкции школ и специальные учебные заведения;
- Горно-химическая промышленность, строительство и реконструкция рудников;
- Изделия из металла, строительство и реконструкции кузнечных и литейных цехов;

- Металлообрабатывающее оборудование, создание сборочных цехов;
- Нерудные материалы, разработки месторождений песка и строительство заводов;
- Строительство заводов по производству металлического проката;
- Строительство завода по производству металлоконструкций и арматуры;
- Строительство и реконструкции комплексов производства стройматериалов и изделий;
- Строительство и реконструкции заводов по производству металлических труб;
- Цветная металлургия, строительство и реконструкция горно-обогатительных комбинатов;
- Черная металлургия, строительство и реконструкция горно-обогатительных комбинатов.

Информация о ГВ		Таблица взаимосвязей		Импульсы		Сравнение по факторам для ГВ	
Содержание главных векторов							
Loadings:							
	RC1	RC2	RC3	RC4			
Бизнес-центры	0.953						
Горно-химическая промышленность				0.677			
Гостиничные комплексы	0.358		0.153	-0.148			
Жилые комплексы	0.974						
Изделия из металла				0.741			
Металлообрабатывающее оборудование				-0.237			
Нерудные материалы		0.838					
Прокат металлический		0.754					
Строительные металлоконструкции и изделия	0.249		0.953				
Стройматериалы и изделия		0.844					
Трубы металлические			0.988				
Цветная металлургия		0.518		-0.128			
Черная металлургия				0.272			
Школы и специальные учебные заведения	0.878		0.270				
	RC1	RC2	RC3	RC4			
SS loadings	2.844	2.272	1.996	1.186			
Proportion Var	0.203	0.162	0.143	0.085			
Cumulative Var	0.203	0.365	0.508	0.593			

Рис. 4.13. Пример таблицы отчета о главных векторах

В результате когнитивного моделирования мы сможем оценить влияние роста объемов финансирования металлургии на инвестиции в другие отрасли, в том числе инфраструктурные.

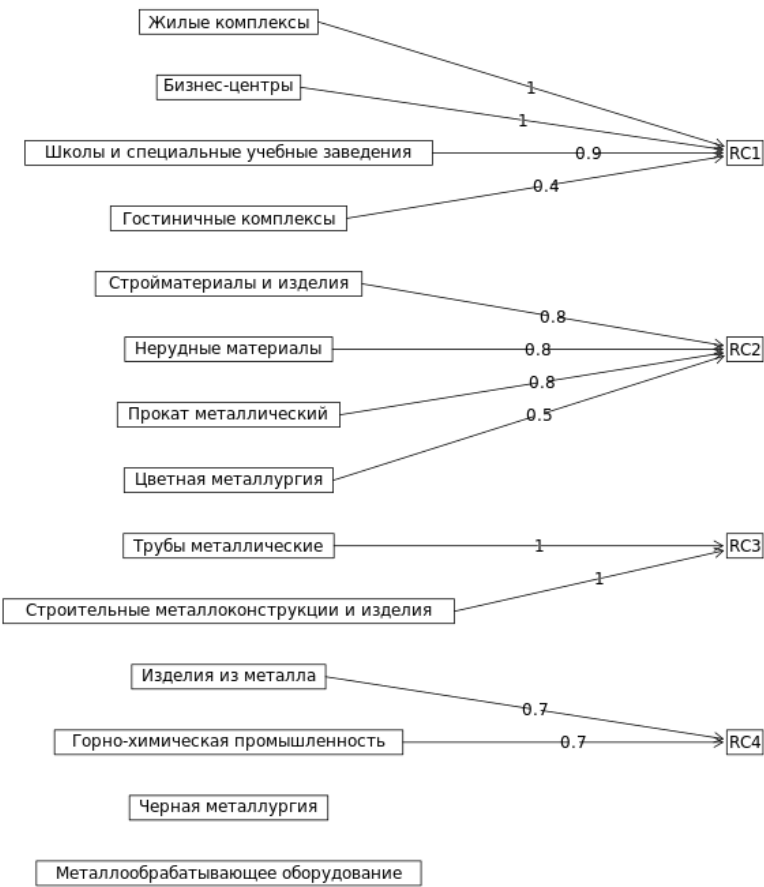


Рис. 4.14. Пример графика распределения отраслей по главным векторам

На первой вкладке «Информация о главных векторах» (при использовании метода главных компонент) представлен отчет о распределении выбранных факторов анализа по главным компонентам когнитивной модели. Сформированный отчет включает в себя таблицы и графики (рис. 4.13 и 4.14).

Полученные результаты показывают, как сформировались четыре фактора (главных векторов), отражающих принадлежность к ним анализируемых инвестиционных проектов. На рис. 4.14 видно, какие проекты имеют наибольший вес в каких факторах, а на рис. 4.13 видны конкретные коэффициенты. В первый фактор вошли проекты из категории «качества жизни»: строительство и реконструкция бизнес-центров, гостиничных и жилых комплексов и школ. Второй фактор представлен проектами строительства предприятий по производству нерудных материалов, металлопроката, стройматериалов и проектами в отрасли цветной металлургии. Третий фактор представлен проектами создания заводов строительных металлоконструкций и металлических труб. Четвертый фактор характеризуется проектами строительства рудников горно-химической промышленности и строительства и реконструкции кузнечных и литейных цехов. Интересно, что проекты черной металлургии и производства металлообрабатывающего оборудования не вошли с большим весом ни в один из сформированных факторов. Первичный анализ показал, что инвестиционных проектов создания сборочных цехов металлообрабатывающего оборудования в анализируемой выборке очень мало, и их количество не позволяет получить высокую объясняющую способность сформированного фактора, а фактор черной металлургии появляется при добавлении еще одного главного вектора, но он также имеет низкую объясняющую способность.

По описательной таблице на рис. 4.13 видно, что сформированные четыре фактора в совокупности описывают 59,3% дисперсии всей выборки. Первый фактор имеет наибольшее влияние в системе взаимодействующих факторов и объясняет 20,3% общей дисперсии.

Одним из результатов использования метода главных векторов, помимо формирования латентных факторов, является построение таблицы их взаимосвязей, которая используется для построения взвешенного ориентированного графа, представляющего, по сути, когнитивную модель сложной системы взаимодействующих социально-экономических переменных (когнитивную карту).

На следующей вкладке «Таблица взаимосвязей» расположена редактируемая таблица взаимосвязей когнитивной карты. В зависимости от выбранного метода анализа ее вид может отличаться: в случае метода главных компонент она отображает взаимосвязи главных компонент (факторов) (рис. 4.15), а в случае линейной регрессии описывает взаимосвязи самих первичных показателей.

Информация о ГВ		Таблица взаимосвязей	Импульсы	Сравнение по факторам для ГВ	
		Комп. 1	Комп. 2	Комп. 3	Комп. 4
1		0.885	-0.028	0.464	-0.006
2		0.041	0.999	-0.018	-0.022
3		-0.463	0.035	0.885	0.028
4		0.021	0.021	-0.021	0.999

Рис. 4.15. Вкладка с редактируемой таблицей взаимосвязей

В таблице видно, что главные вектора получились почти независимыми и влияют в основном только сами на себя, кроме положительного влияния первого фактора на третий и, наоборот, отрицательного влияния третьего фактора на первый. Полученная таблица позволяет провести импульсное моделирование и оценить влияние подаваемого импульса в тот или иной фактор на всю систему или любой из факторов, выбранный в качестве целевого для всей системы, например фактор «Качество жизни».

На третьей вкладке «Импульсы» пользователь может задавать импульсы для моделирования развития системы и просматривать отчет о результатах моделирования, который состоит из таблиц, графиков и диаграмм (рис. 4.16–4.20).

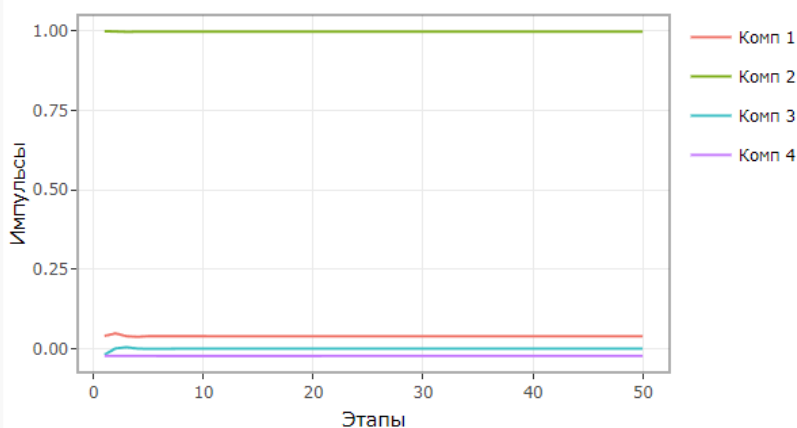
Импульсный вектор

0,1,0,0

Таблица эффектов после 50 итераций

Комп 1	Комп 2	Комп 3	Комп 4
0.04	1.00	0.00	-0.02

Импульсы по главным векторам



Результаты моделирования

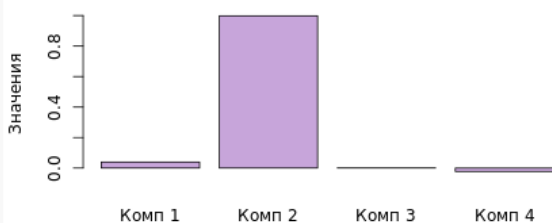


Рис. 4.16. Распределение импульса во второй фактор

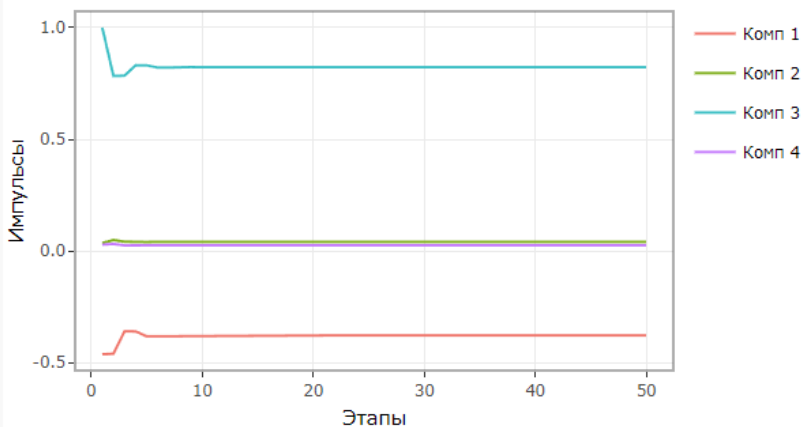
Импульсный вектор

0,0,1,0

Таблица эффектов после 50 итераций

Комп 1	Комп 2	Комп 3	Комп 4
-0.38	0.04	0.82	0.03

Импульсы по главным векторам



Результаты моделирования

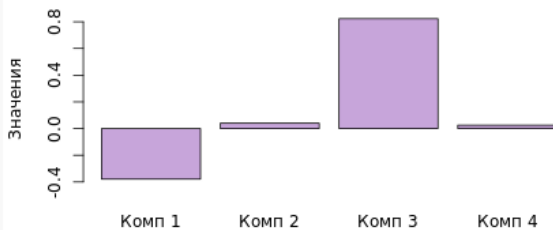


Рис. 4.17. Распределение импульса в третий фактор

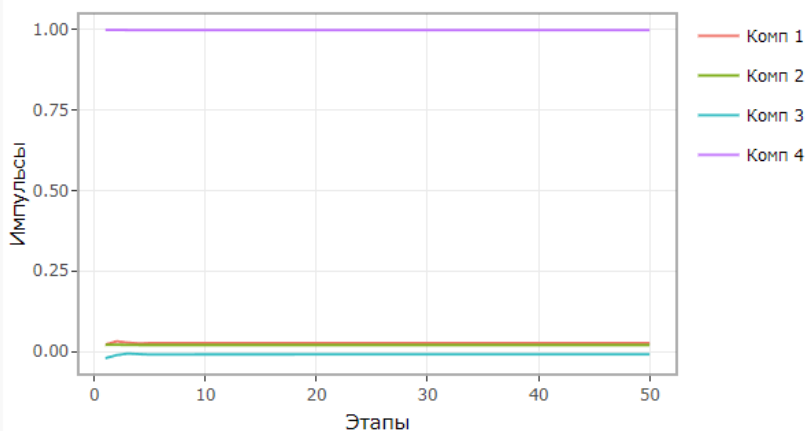
Импульсный вектор

0,0,0,1

Таблица эффектов после 50 итераций

Комп 1	Комп 2	Комп 3	Комп 4
0.03	0.02	-0.01	1.00

Импульсы по главным векторам



Результаты моделирования

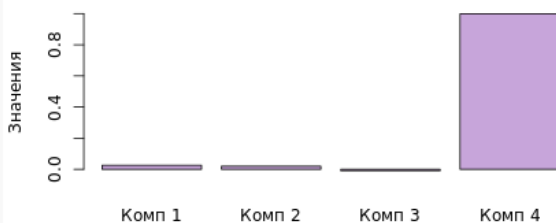


Рис. 4.18. Распределение импульса в четвёртый фактор

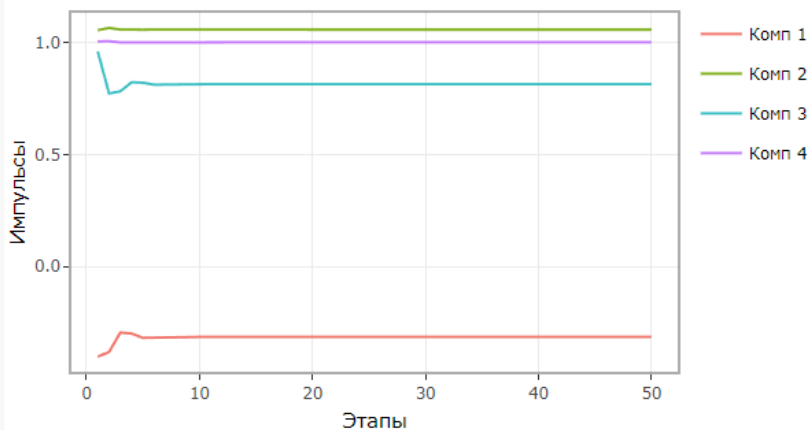
Импульсный вектор

0,1,1,1

Таблица эффектов после 50 итераций

Комп 1	Комп 2	Комп 3	Комп 4
-0.31	1.06	0.81	1.00

Импульсы по главным векторам



Результаты моделирования

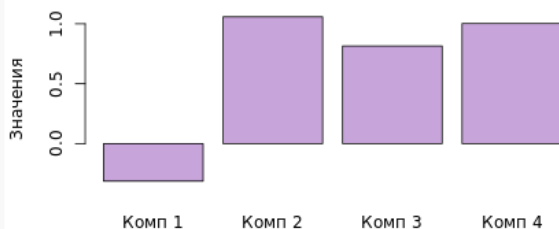


Рис. 4.19. Распределение импульса в факторы 2–4

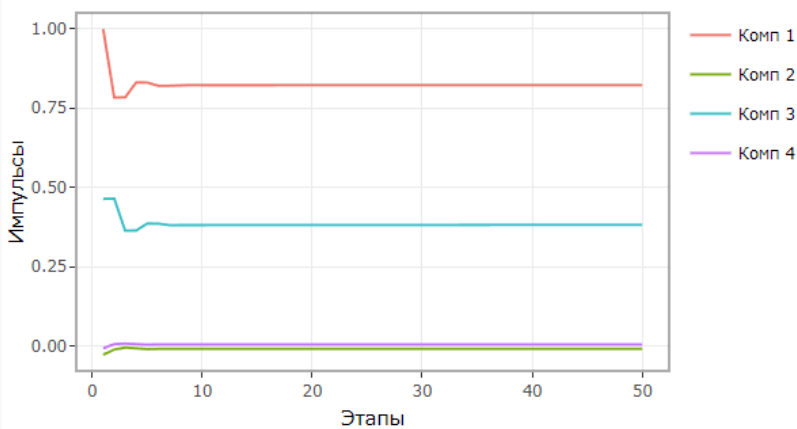
Импульсный вектор

1,0,0,0

Таблица эффектов после 50 итераций

Комп 1	Комп 2	Комп 3	Комп 4
0.82	-0.01	0.38	0.00

Импульсы по главным векторам



Результаты моделирования

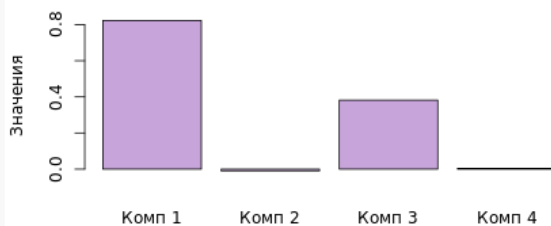


Рис. 4.20. Распределение импульса в первый фактор

Мы задаем импульсы по очереди в каждый фактор и оцениваем его влияние на целевой («Качество жизни»). Мы можем осуществить это, изменяя импульсный вектор, который включает импульсные веса в каждую вершину, записанные через запятую. Например, если мы хотим подать импульс во второй фактор, то нужно записать “0,1,0,0” или если мы хотим подать импульсы в три фактора со второго по четвертый, то нужно записать “0,1,1,1”. Полученные результаты моделирования записаны в разных форматах для удобства анализа. В таблице и на графике результатов моделирования отображаются итоговые значения импульса по векторам. На центральном графике видно, как ведут себя коэффициенты на разных итерациях когнитивной модели. Если коэффициенты сходятся, то система стабильна, если сходимости нет, то модель построена не корректно и требует корректировки.

Результат моделирования демонстрирует, что импульсы во второй и четвертые вектора дают минимальный прирост первому целевому фактору, а третий фактор влияет на прирост целевого фактора отрицательно. Это свидетельствует о том, что инвестиции в строительство заводов по производству металлических труб и строительных металлоконструкций отрицательно влияют на развитие жилищных и гостиничных комплексов, бизнес-центров и школ в регионе, что требует дальнейшего анализа.

Одновременный импульс во все отрасли второго, третьего и четвертого факторов не дают положительного эффекта для первого, но импульс в первый фактор обеспечивает положительное развитие для третьего. Это логично, потому что для строительства инфраструктурных объектов необходимы строительные материалы.

Представленный пример демонстрирует модельные и вычислительные возможности системы, но более содержательный анализ использования когнитивного моделирования представлен в главе 5.

ДММ-КАМИН

Базовое описание модели ДММ-КАМИН представлено в разделе 3.1, а содержательное описание результатов в разделе 5.4.

Цель модельного комплекса – прогнозирование динамики макроэкономических показателей экономики в зависимости от изменения базовых показателей, таких как объемы инвестирования, матрицы материалоёмкости и др.

Данный модуль предоставляет следующие возможности:

- Выполнение оптимизационной модели ДММ-КАМИН. Для запуска вычислений нужно загрузить персональные исходные данные в формате *xlsx*, нажать кнопку «Сохранить» и кнопку «Запустить стандартную оптимизацию».
- Редактирование данных. Для редактирования необходимо дважды щелкнуть на ту цифру в таблице, которую нужно изменить, и ввести новое значение (рис. 4.21).
- Отображение данных в табличном виде. Для просмотра требуемой таблицы нужно выбрать нужные данные из списка.
- Визуализация результатов, с учетом сценарных расчетов.

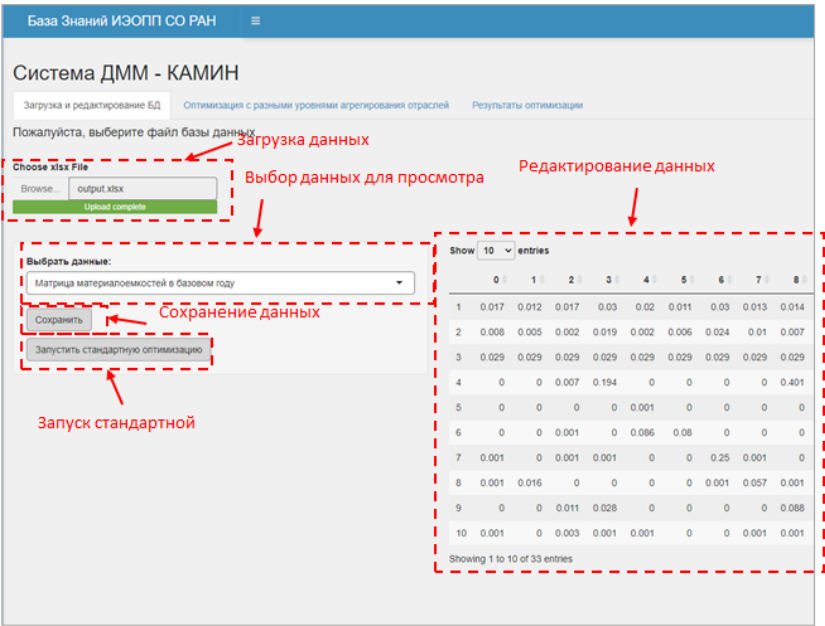


Рис. 4.21. Запуск оптимизации системы ДММ-КАМИН

Проведем демонстрацию работы интерфейса на простом примере. Как повлияет на макроэкономическую динамику снижение экспорта ресурсов в экономике. Для этого загрузим тестовую базу данных 2019 г. в формате *xlsx*. В базе данных на разных листах содержится вся необходимая статистика для работы модели. Далее нужно сохранить данные и запустить стандартную оптимизацию. Результаты оптимизации будут отображены в разделе «Результаты оптимизации» в виде прогнозов до 2035 г. Таким образом, реализуется базовый прогноз. Сохраним результат под именем «Базовый сценарий».

Во вкладке «Результаты оптимизации» можно:

- анализировать таблицы результатов оптимизации посредством выбора нужного показателя из списка;
- выгрузить результаты оптимизации в виде *xlsx* файла;
- задать название оптимизации для сохранения и сохранить результат для визуализации результатов.
- выбрать названия оптимизаций для визуализации и нажать команду «Построить график» для построения графика.

Чтобы оценить эффект снижения экспорта ресурсной группы, в разделе «Загрузка и редактирование БД» зайдём в раздел «Сальдо (экспорт-импорт) по годам» и уменьшим экспорт ресурсной группы с 6087.771 на 4000 на ближайшие 5 лет (2020–2024 гг.). Сохраним полученные результаты и запустим оптимизацию. В разделе «Результаты оптимизации» сохраним новые результаты под именем «Снижение экспорта», выберем файлы для добавления на график «Базовый сценарий» и «Снижение экспорта», и нажмём «Построить график». В результате будет построен график прогнозов по двум сценариям. Для визуализации можно выбирать параметры «Валовый внутренний продукт», «Валовый выпуск» и «Численность занятых» в абсолютных значениях или темпах роста. На рис. 4.22 отображён пример с визуализацией ВВП в абсолютных значениях. Анализ графика показывает, что более сильная просадка ВВП в 2020 г., но после 2025 г. происходит возврат на базовую траекторию.



Рис. 4.22. Визуализация разных прогнозных сценариев в системе ДММ-КАМИН

Для расширения перечня показателей в разделе «Результаты оптимизации» можно выполнить настройку уровней агрегирования отраслей на вкладке «Оптимизация с разными уровнями агрегирования отраслей» (рис. 4.23). Так, в результаты можно добавить показатели «Основные фонды», «Фондоемкость валового выпуска», «Возмещение выбытия основных фондов», «Нормы возмещения выбытия основных фондов», «Приросты основных фондов», «Вводы основных фондов» и «Инвестиции в основной капитал» указав нужные уровни агрегирования. После нажатия «Запустить оптимизацию», набор доступных для анализа показателей в разделе «Результаты оптимизации» увеличится.

База Знаний ИЗОПП СО РАН

Система ДММ - КАМИН

Загрузка и редактирование БД Оптимизация с разными уровнями агрегирования отраслей Результаты оптимизации

Сохранить отчет

Запустить оптимизацию

- Отраслевые показатели
 - Валовой выпуск: 2 уровень
 - Валовой внутренний продукт: 2 уровень
 - Численность занятых: 2 уровень
- Сырьевые показатели
 - Материальные затраты: Для отраслей - 2, для видов МЗ - 0
 - Материалоемкость валового выпуска: Для отраслей - 2, для видов МЗ - 0
- Инвестиционные показатели
 - Основные фонды: Для отраслей - 1, для видов ОПФ - 0

Рис. 4.23. Запуск оптимизации с выбором разных уровней агрегирования отраслей

Реализованная система ДММ-КАМИН позволяет визуализировать разные прогнозные сценарии и является инструментом оценок влияния различных шоков, в том числе и при реализации кластеров инвестиционных проектов, на долгосрочную динамику макроэкономических показателей. Ключевой точкой для реализации модели является сбор данных и правильное внесение изменений для реализации сценарных прогнозов.

4.2.3. Работа с данными

Панельные данные

Модуль «Панельные данные» предназначен для загрузки и визуализации социально-экономических показателей.

Интерфейс работы с панельными данными позволяет выбрать пул необходимых показателей с возможностью текстового поиска, выбора регионов и временного периода (рис. 4.24).

Интерфейс позволяет просмотреть данные визуально в табличном виде, загрузить в виде *xlsx* и провести базовую аналитику и визуализацию. Визуализация представляет собой локальную интерактивную ГИС-карту (рис. 4.25).



Рис. 4.24. Выгрузка данных в модуле «Панельные данные»

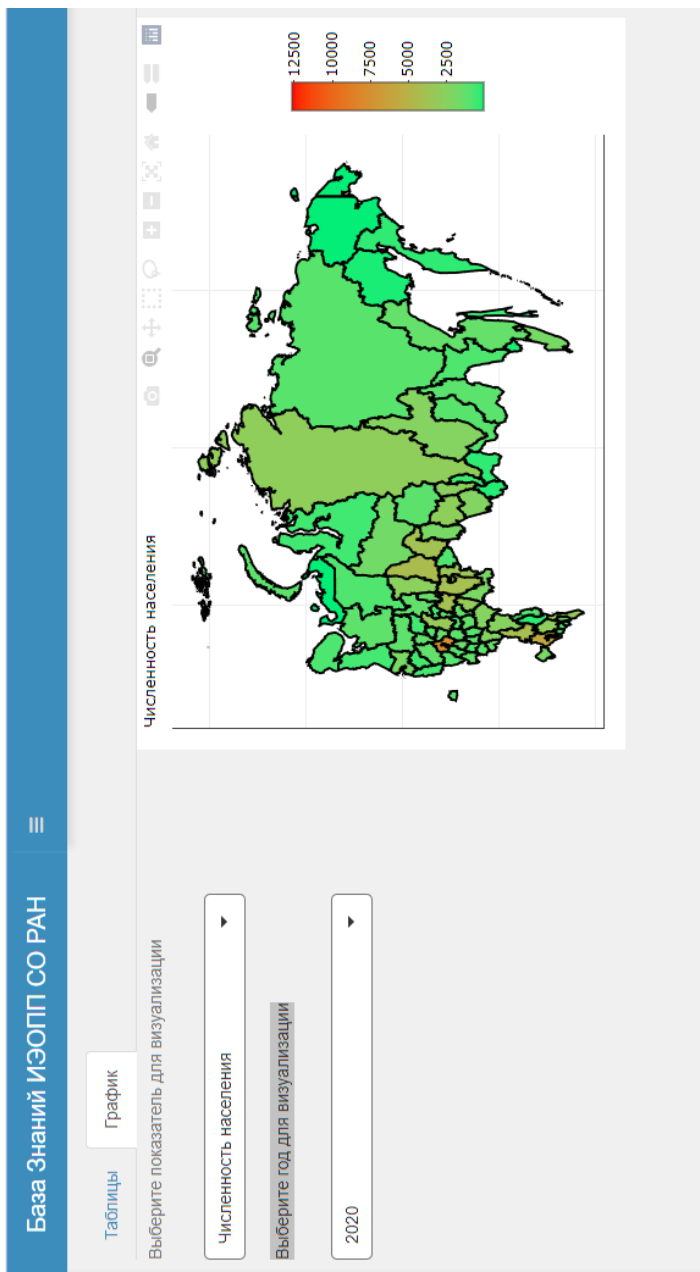


Рис. 4.25. Визуализация данных в модуле «Панельные данные»

Как пример можно описать кейс, когда в панельной регрессионной модели в разрезе регионов, нужно добавить фактор, описывающий цифровую инфраструктуру регионов. Такими факторами, могут быть факторы, связанные с наличием компьютеров или интернетом. Для поиска возможных факторов воспользуемся полнотекстовым поиском и введем слова компьютер и интернет (рис. 4.26 и 4.27). В результате появится список факторов, удовлетворяющих запросу. Из полученного списка, после анализа было принято решение выбрать в качестве регрессора – Число персональных компьютеров с доступом к сети Интернет на 100 работников – потому что он удовлетворяет всем критериям и уже был рассчитан на душу населения, что снимало вопрос гетероскедастичности и сравнимости регионов.

Выберите интересующие вас показатели (макс. 20):

компьютер

☐ Все показатели:

- ☐ ТРАНСПОРТ
- ☐ ИНВЕСТИЦИИ
- ☐ НАСЕЛЕНИЕ
- ☐ СТРОИТЕЛЬСТВО
- ☐ КУЛЬТУРА, ОТДЫХ И ТУРИЗМ
- ☐ ЗДРАВООХРАНЕНИЕ

☐ **ИНФОРМАЦИОННЫЕ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ**

- ☐ Использование информационных и коммуникационных технологий в организациях
 - ☐ **Организации, использовавшие персональные компьютеры**
 - ☐ Организации, использовавшие серверы
 - ☐ Организации, использовавшие локальные вычислительные сети
 - ☐ Организации, использовавшие глобальные информационные сети
 - ☐ Использование сети Интернет в организациях
 - ☐ Организации, имевшие веб-сайт
- ☐ **Число персональных компьютеров на 100 работников**
 - ☐ Всего
 - ☒ **Число персональных компьютеров с доступом к сети Интернет**
 - ☐ Использование электронного документооборота в организациях
- ☐ **Использование персональных компьютеров и сети Интернет в домашних хозяйствах**
 - ☐ **Удельный вес домашних хозяйств, имевших персональный компьютер**
 - ☐ Удельный вес домашних хозяйств, имевших доступ к сети Интернет
 - ☐ Удельный вес домашних хозяйств, имевших широкополосный доступ к сети Интернет

Рис. 4.26. Поиск показателей по запросу «Компьютер»

Выберите интересующие вас показатели (макс. 20):

интернет

☐ Все показатели:

- ☐ ТРАНСПОРТ
- ☐ ИНВЕСТИЦИИ
- ☐ НАСЕЛЕНИЕ
- ☐ СТРОИТЕЛЬСТВО
- ☐ КУЛЬТУРА, ОТДЫХ И ТУРИЗМ
- ☐ ЗДРАВООХРАНЕНИЕ
- ☒ ИНФОРМАЦИОННЫЕ И КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
 - ☐ Использование информационных и коммуникационных технологий в организациях
 - ☒ **Использование сети Интернет в организациях**
 - ☐ Организации, использовавшие сеть Интернет
 - ☐ Организации, использовавшие широкополосный доступ к сети Интернет
 - ☐ Организации, имевшие веб-сайт
- ☒ Число персональных компьютеров на 100 работников
 - ☐ Всего
 - ☒ **Число персональных компьютеров с доступом к сети Интернет**
 - ☐ Использование электронного документооборота в организациях
- ☒ **Использование персональных компьютеров и сети Интернет в домашних хозяйствах**
 - ☐ Удельный вес домашних хозяйств, имевших персональный компьютер
 - ☐ Удельный вес домашних хозяйств, имевших доступ к сети Интернет
 - ☐ Удельный вес домашних хозяйств, имевших широкополосный доступ к сети Интернет
- ☒ **Использование сети Интернет населением**
 - ☐ Население, использовавшее сеть Интернет
 - ☐ Население, использовавшее сеть Интернет каждый день или почти каждый день
 - ☐ Число подключенных абонентских устройств мобильной связи на 1000 человек населения
- ☒ **Численность активных абонентов фиксированного и мобильного широкополосного доступа к сети Интернет на 100 человек населения**
 - ☐ Численность активных абонентов фиксированного широкополосного доступа к сети Интернет
 - ☐ Численность активных абонентов мобильного широкополосного доступа к сети Интернет

Рис. 4.27. Поиск показателей по запросу «Интернет»

После этого выделяем нужный показатель, осуществляем предварительный просмотр и загружаем данные, нажатием кнопки «Загрузить xls». Так же можно сравнить регионы на карте, зайдя в раздел «График», где автоматически будет отображаться выбранный показатель на последний доступный год.

Инвестиционные проекты

Модуль «Инвестиционные проекты» предназначен для анализа инвестиционной активности на территории Российской Федерации. В разделе/вкладке «Карточки инвестиционных проектов» реализована возможность работы с таблицей инвестиционных проектов с возможностью сортировки и полнотекстового поиска. В случае наличия карточки инвестиционного проекта, можно нажать на кнопку «Загружен» в столбце «Статус» для загрузки карточки в формате pdf (рис. 4.28).

База Знаний ИЗОПП СО РАН

Карточки инвестиционных проектов

Инвестиционные проекты

Show 10 entries

Полнотекстовый поиск

Агрегированная информация по инвестиционным проектам

Таблица инвестиционных проектов

Загрузка карточки инвестиционного проекта

Search: СПГ Сахалин

Id	Статус	Название	Стадия	Регион	Отрасль	Объем инвестиций в млрд.	Собственность	long	lat
31399	Загружен	Строительство третьей технологической линии по производству СПГ	Подготовка к строительству	Сахалинская область	СПГ	384	Част	142.751	50.1509
31175	Не загружен	Строительство завода по сжижению природного газа	Приостановлен (на стадии проектирования проработки)	Сахалинская область	СПГ	19	Част	142.949	53.589
14128	Загружен	Строительство комплекса по производству СПГ (1-я очередь)	Предпроектные проработки	Сахалинская область	СПГ	8.4	Компания с госучастием	143.101	49.2214
174	Загружен	Строительство завода по сжижению природного газа	Проектирование	Сахалинская область	СПГ	6	Част	142.751	50.1509
1295	Загружен	Строительство мини-завода по сжижению природного газа	Введен в эксплуатацию	Сахалинская область	СПГ		Част	142.613	46.9538

Рис. 4.28. Работа с таблицей инвестиционных проектов

В панели «Агрегированная информация по инвестиционным проектам» можно агрегировать данные инвестиционных проектов для последующего использования в модельном аппарате (рис. 4.29).

The screenshot shows a web interface with two tabs: 'Карточки инвестиционных проектов' and 'Агрегированная информация по инвестиционным проектам'. The active tab contains several filter sections:

- Регионы:** Амурская область, Калининградская область, Краснодарский край, Ленинградская область, Москва, Московская область.
- Этапы:** Модернизация, Строительство (нулевой цикл), Строительство (реконструкция).
- Отрасли:** АБК, инженерные, технические, лабораторные объекты, Авиастроение и авиаремонт, Автодороги и мосты, Автомобилестроение, Агропромышленные комплексы, центры и парки.
- Собственность:** Гос., ГЧП, Компании с госучастием, Част.

At the bottom of the filter panel is a button labeled 'Загрузить xlsx' with a download icon.

Рис. 4.29. Работа с агрегированными данными инвестиционных проектов

Данные компаний

Модуль «Данные компаний» предназначен для анализа компаний и их бухгалтерской отчетности на территории Российской Федерации. В разделе «Информация о компаниях» реализована функциональность выгрузки таблицы информации о компаниях и их бухгалтерской отчетности (рис. 4.30).

О ИЭОПП СО РАН

Гис моделирование

Модели

» Когнитивная модель

» ДММ - КАМИН

Работа с данными

» Панельные данные

» Инвестиционные проекты

» Данные компаний

» Таможенная статистика

» Технологии

Подробнее

Информация о компаниях

Выбор параметров

Выберите регионы:

Брянская область

Выберите нужные кода ОКВЭД

01.11.2 × 01.46 × 01.45.1 × 01.42 ×
01.43.1 × 01.47 × 01.47.11 × 01.41.1 ×

☐ ОКВЭД только основной деятельности

☐ Действующие компании

☒ Добавить в таблицу данные финансовой отчетности

Выберите год для финансовой отчетности:

2005 2020

2005 2007 2009 2011 2013 2015 2017 2019 2020

Выберите нужные коды параметров финансовой отчетности

1 ×

Скачать каталог параметров финансовой отчетности

Загрузить xlsx

Скачивание может занять достаточно долгое время

Рис. 4.30. Работа с модулем «Данные компаний»

Таможенная статистика

Модуль «Таможенная статистика» предназначен для анализа таможенных данных регионов Российской Федерации. В разделе/вкладке можно выгрузить таблицу таможенной статистики (рис. 4.31). Интерфейс работы с данными таможенной статистики представляет собой систему выгрузки и визуализации данных с дискриминацией по региону, экспорту-импорту, кодам ТН ВЭД и годам. Также предусмотрено скачивание справочника ТН ВЭД кодов для возможности их детальной проработки.

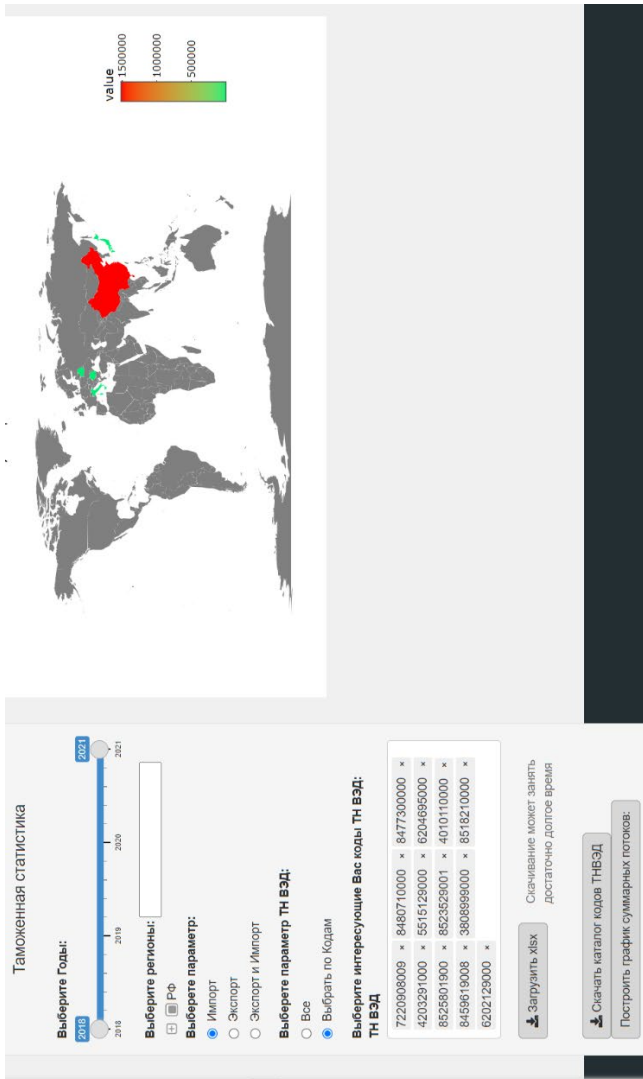


Рис. 4.31. Работа с модулем «Таможенная статистика»

Технологии

Модуль «Технологии» создан для анализа известных технологий производства разной продукции. В процессе анализа инвестиционных проектов и развития территорий исследователям постоянно приходится сталкиваться с проблемой учета текущих технологических процессов создания продукта. ИЭОПП СО РАН совместно с Институтом Катализа СО РАН были собраны и структурированы данные по текущим технологиям производства базовых продуктов с указанием входного, выходного продукта, необходимых условий производства, компаний, реализующих технологию и документы, где технология описана. В разделе можно работать с таблицей технологий производства с возможностью сортировки и полнотекстового поиска. В таблице указано расположение информации со ссылкой на страницы отчетов и справочников, которые можно скачать, выбрав нужный справочник из списка «Выберите технологический справочник» и нажав «Скачать технологический справочник» (рис. 4.32).

4.3. ГИС-интерфейс Базы Знаний

В рамках выбора интерфейса в качестве оптимальной платформы хранения ГИС-данных был выбран программный продукт NextGis. Сам ГИС-интерфейс является частью веб-интерфейса и создается на базе пакета Leaflet. Подобная структура позволяет использовать существующие в языках программирования методы моделирования ГИС информации и создавать собственные слои данных.

На текущий момент ГИС-интерфейс имеет следующие блоки данных:

1. Инфраструктура (рис. 4.33).
 - 1.1. Административные районы,
 - 1.2. Автомобильные дороги,
 - 1.3. Аэропорты,
 - 1.4. Газопроводы,
 - 1.5. Железные дороги,
 - 1.6. Нефтепроводы,
 - 1.7. Речные порты.
2. Твердые полезные ископаемые (рис. 4.34)
 - 2.1. Точечные месторождения,
 - 2.2. Золотые жилы (линии),
 - 2.3. Минерагенические районы,
 - 2.4. Минерагенические узлы,
 - 2.5. Геолого-промышленные кластеры.
3. Углеводородное сырье (рис. 4.35)
 - 3.1. Лицензирование,
 - 3.2. Месторождения,
 - 3.3. Скважины,
 - 3.4. Выявленные структуры,
 - 3.5. Подготовленные структуры,
 - 3.6. Кластеры нефтегазодобычи,
 - 3.7. Полигоны.
4. Инвестиционные проекты (рис. 4.36)
5. Данные о компаниях (рис. 4.37)
6. Статистические показатели (см. рис. 4.8)
7. Слои для демонстрации (рис. 4.38). Представляют собой примеры возможности создания и отрисовки собственных ГИС-слоёв.

- 7.1. Точки для демонстрации по черной металлургии,
- 7.2. Транспортная стратегия, железные дороги,
- 7.3. Транспортная стратегия, автомобильные дороги,
- 7.4. Транспортная стратегия, порты.

Для ряда слоёв есть возможность выбора дискриминации:

- Для слоя «Точечные месторождения» можно выбрать вид твердых полезных ископаемых для визуализации (рис. 4.34),
- Для «Инвестиционные проекты» можно выбрать отрасль инвестиционных проектов для визуализации (рис. 4.36),
- Для «Статистические показатели» можно выбрать статистический показатель и год (рис. 4.39).

В ходе реализации блока инвестиционных проектов и данных о компаниях была использована кластеризация для оптимизации отображения данных. Использование пакета Leaflet позволяет создавать собственные слои, что было проделано при описании транспортной стратегии развития Азиатской России.

В рамках создания ГИС-интерфейса происходит разработка геопространственных технологий в сфере обеспечения пространственной организации территорий и выработки оптимальных пространственных решений по их развитию и функционированию. Планируется создание интерактивной системы, включающей комплекс математических моделей, в том числе моделей оптимального размещения, поиска ближайших месторождений и оптимальных путей доставки. Интерфейс планируется дополнять системой детального выбора данных для анализа и визуализации, что уже частично реализовано. Планируется создание индивидуального пространства пользователя с возможностью создания собственных карт и их визуализации.

Особой точкой дальнейшего развития системы является пространственная аналитика. Предполагается создание интерфейса, где пользователь может нарисовать на карте территорию, выбрать оквэды, отрасли и показатели и получить всю необходимую информацию про выбранную область (рис. 4.40).

Например, сколько компаний с выбранными оквэдами зарегистрированы в этой области, какие у них совокупные выручка, прибыль и затраты на труд. Как распределена выручка или прибыль по оквэд у этих компаний. Где концентрируются в этой области компании с наибольшей выручкой. Какие месторождения есть на выбранной территории, с каким количеством запасов и на какой стадии разработки. Какие инвестиционные проекты реализуются на выбранной территории и на какой стадии они находятся. Какое распределение инвестиционных потоков идет по отраслям и где основная концентрация инвестиционных проектов на территории. На данный момент, система находится на медианном этапе разработки и по еще не внедрена в интерфейс, но пример подобного анализа представлен в разделе 5.5. Основной проблемой, с которой мы столкнулись, является привязка компании к месту регистрации, а не месту экономической деятельности. В результате большинство крупных компаний концентрируются в городе Москва, но фактически работают на территории всей Российской Федерации. Для решения этой проблемы планируется создание дополнительной базы данных по 1000 крупнейших компаний с привязкой точек их экономических интересов. Это позволит увеличить качество анализ и убрать пространственное смещение данных, но потребует больших усилий в создании и поддержании данного блока в актуальном состоянии.

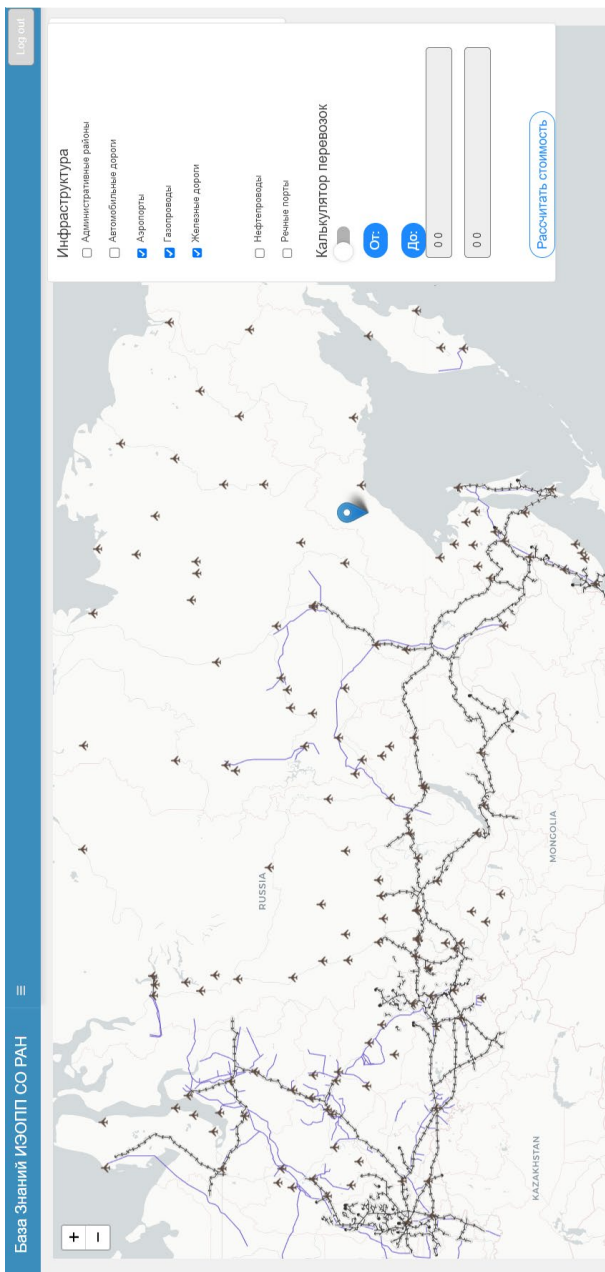


Рис. 4.33. ГИС-интерфейс: Инфраструктура

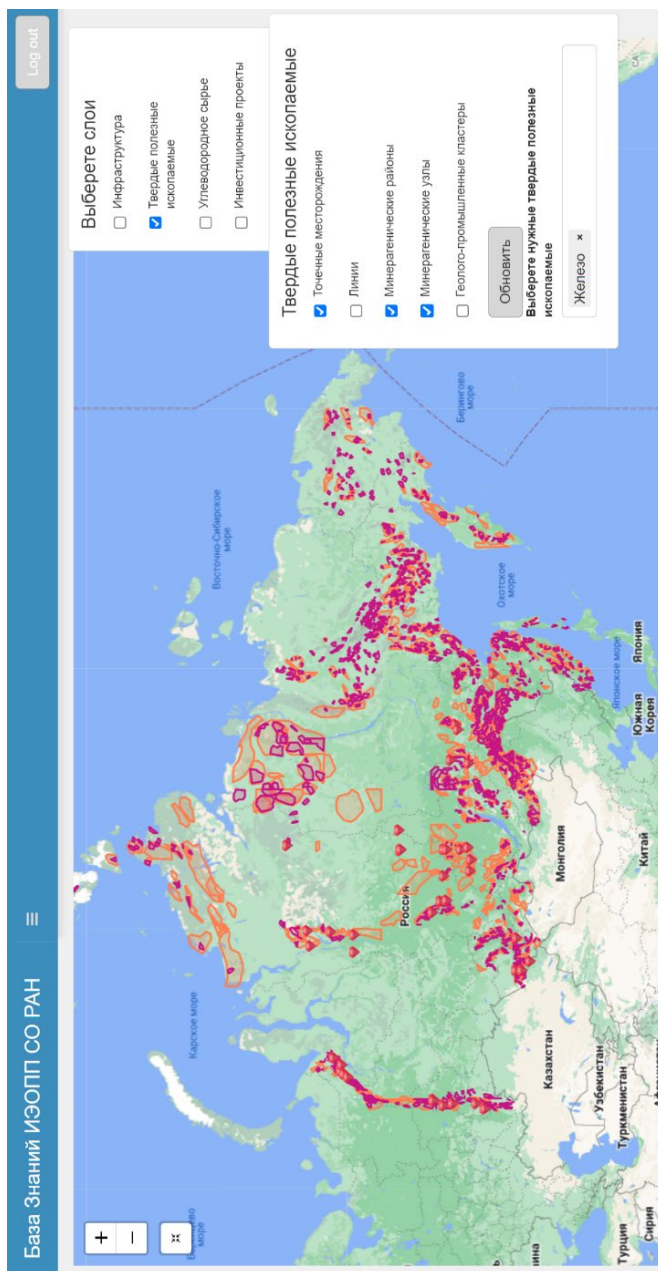


Рис. 4.34. ГИС-интерфейс: Твердые полезные ископаемые

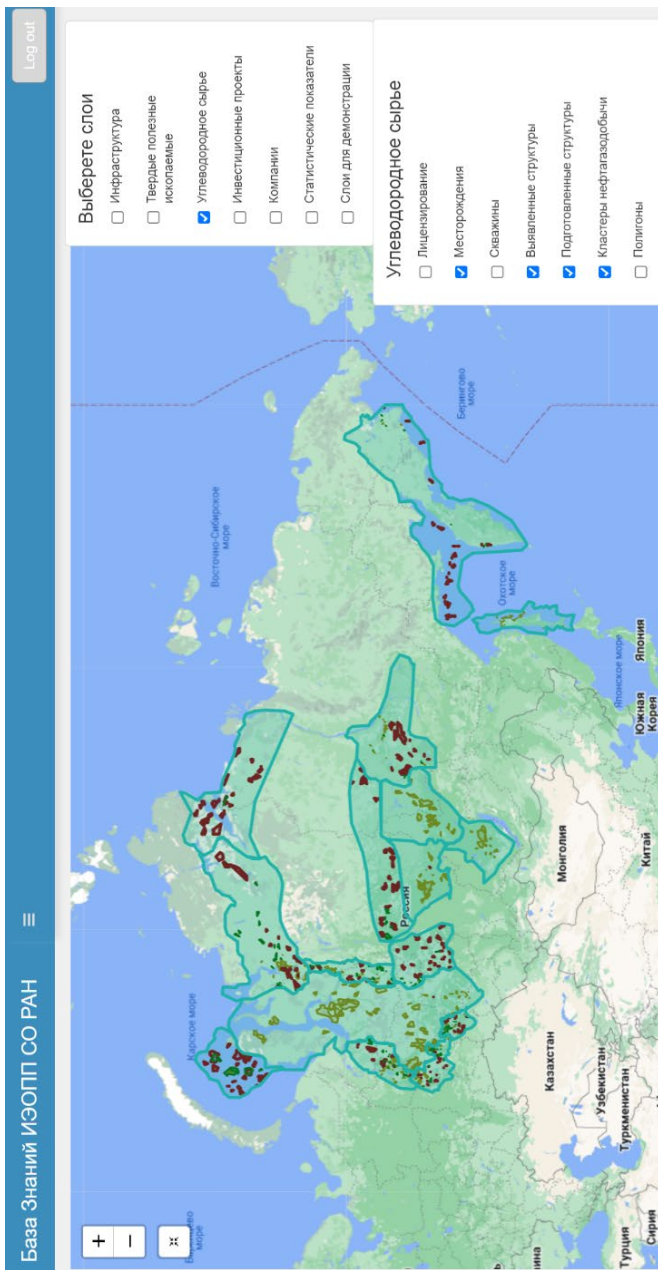


Рис. 4.35. ГИС-интерфейс: Углеводородное сырье

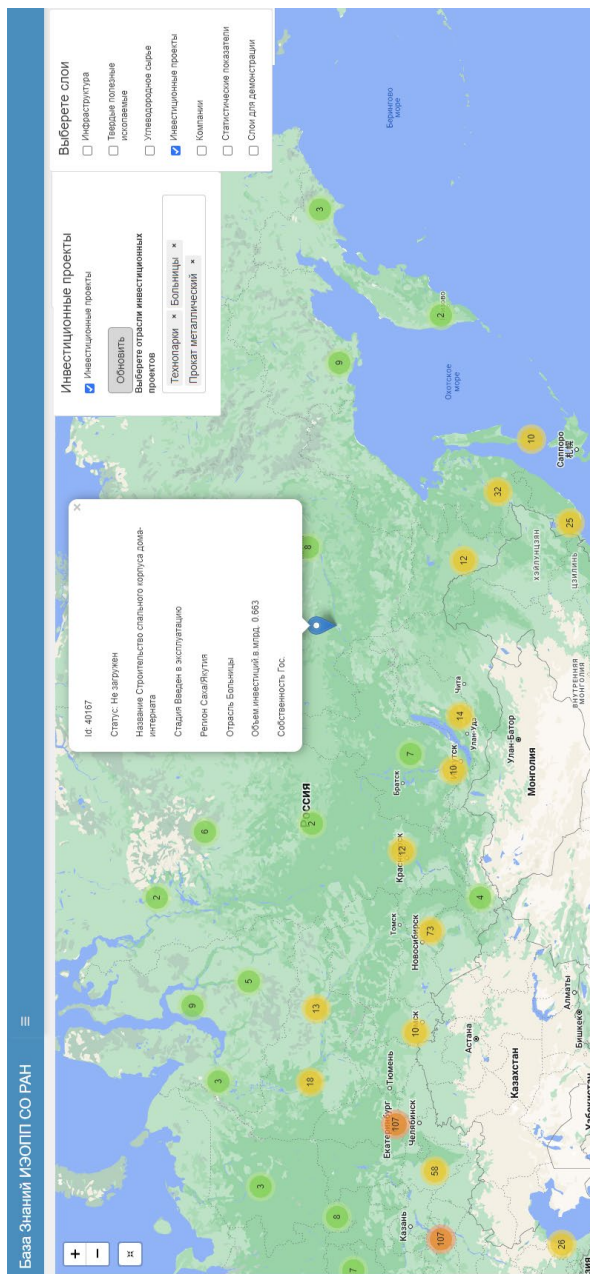


Рис. 4.36. ГИС-интерфейс: Инвестиционные проекты

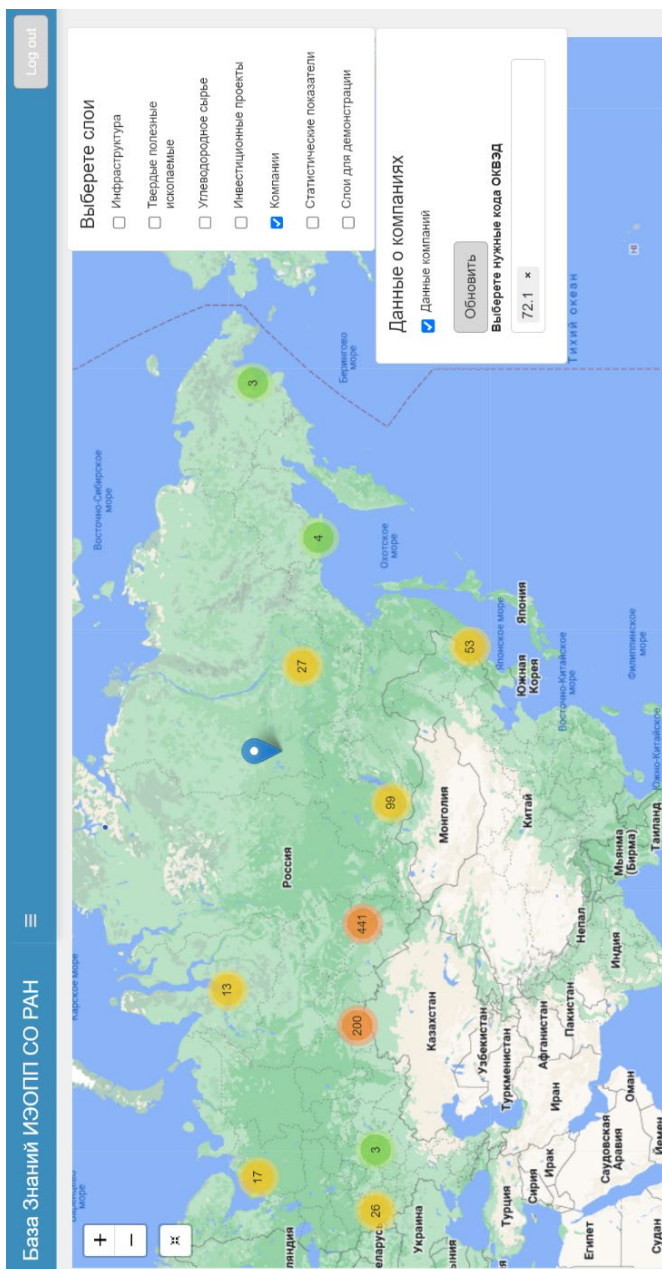


Рис. 4.37. ГИС-интерфейс: Данные о компаниях, ОКВЭД 72.1

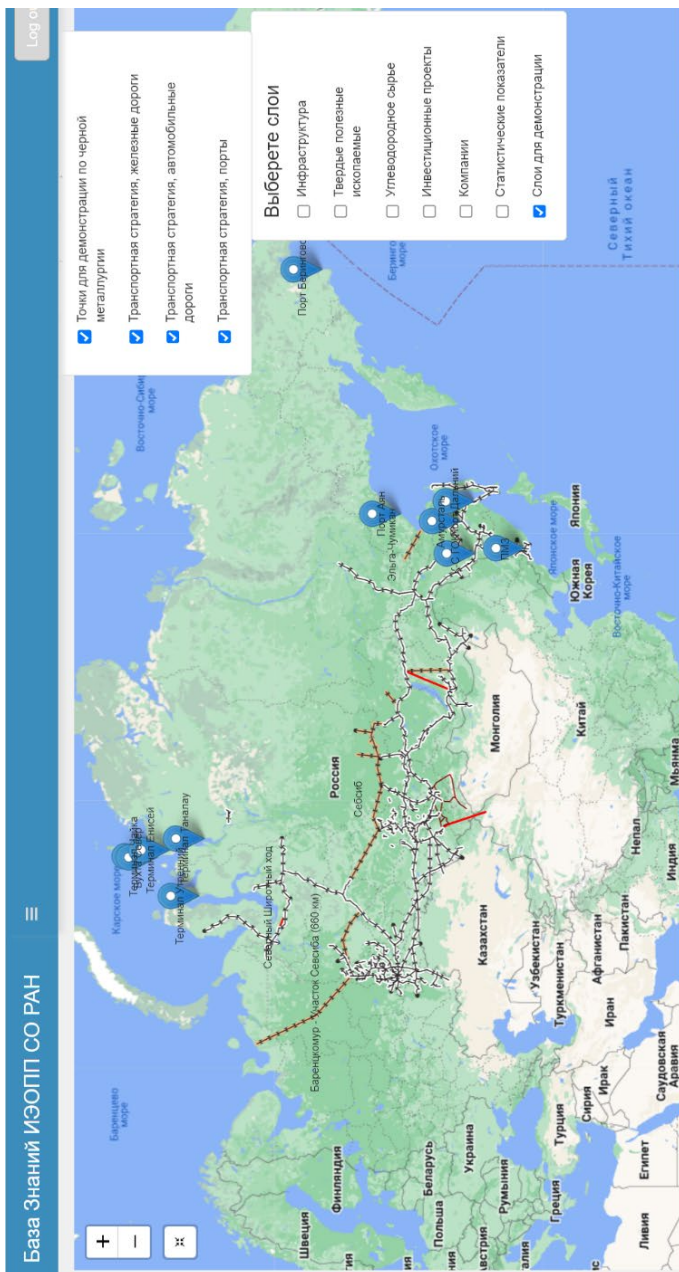
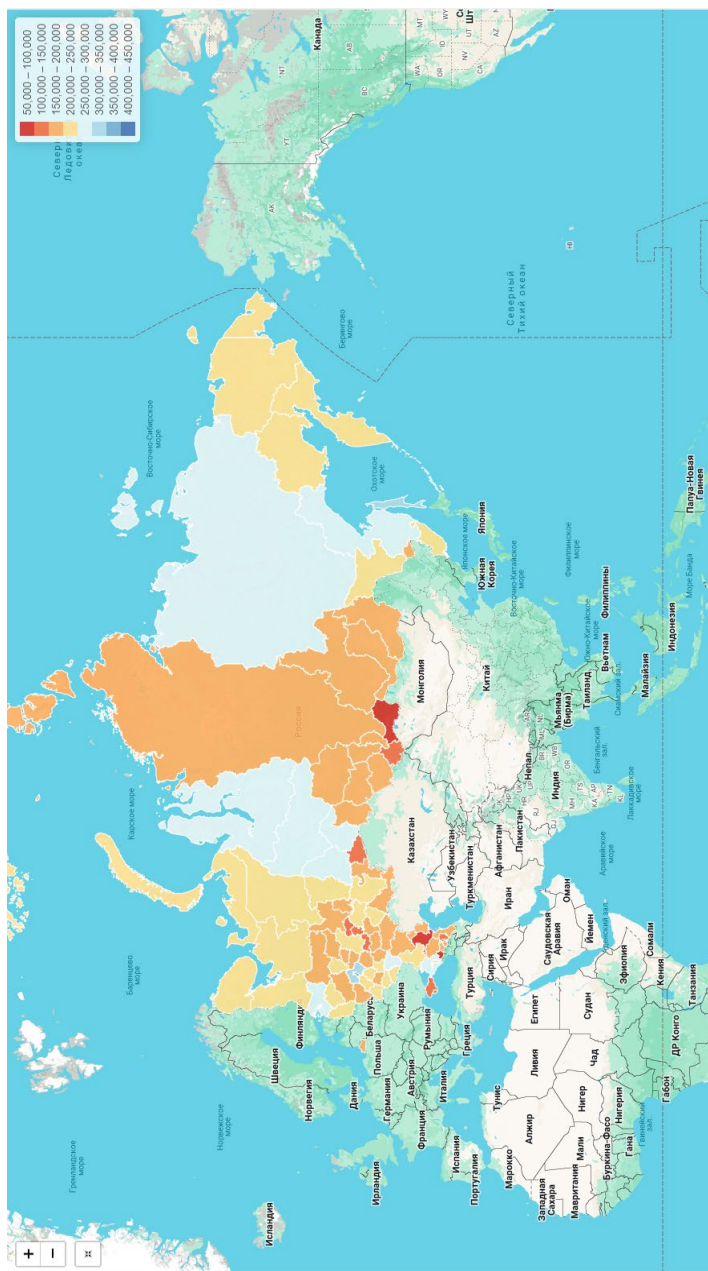


Рис. 4.38. ГИС-интерфейс: Транспортная стратегия, слои для демонстрации



Глава 5

ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ПЛАТФОРМЕННЫХ РЕШЕНИЙ

В данной главе приводятся примеры использования разработанной базы знаний для решения экономических исследовательских задач. Более содержательно описанные примеры представлены в опубликованных монографиях «Новый импульс Азиатской России: источники и средства развития» под ред. В.А. Крюкова, Н.И. Суслова, «Модели и методы прогнозирования: Азиатская Россия в экономике страны» под ред. А.О. Баранова, В.И. Суслова и «Анализ и оценка процессов создания и развития в Азиатской России транспортной магистральной сети различного назначения» под ред. А.А. Широва, О.В. Тарасовой [Новый импульс..., 2023; Модели..., 2023]

5.1. Оценка инвестиционного импульса с использованием когнитивной модели базы знаний

В разделе описан пример использования модуля когнитивного моделирования для оценки влияния инвестиционного импульса, представленного в статье М.А. Ягольнищера, М.А. Овсянниковой и А.В. Костина [Ягольнищера и др., 2022, с. 51–65].

Инвестиции – важнейшая экономическая категория, они являются значимыми как на макро-, так и на микроуровне, и, в первую очередь, для простого и расширенного воспроизводства, структурных преобразований, получения максимальной прибыли и на этой основе – решения многих социальных проблем.

Эффект мультипликации инвестиций возникает тогда, когда рост потребительского спроса одной отрасли стимулирует расширение производств, связанных с нею отраслей, что по цепочке стимулирует спрос на продукцию иных отраслей.

Для решения задачи стабилизации и оживления экономической конъюнктуры должны быть определены отрасли – мультипликаторы экономического роста. При ограниченности инвести-

ционных средств важно всесторонне обосновать приоритетные направления с учетом их воздействия на макроструктуру в целом. При этом учитываются не только экономические, но и социальные аспекты.

Целью исследования является оценка мультипликативного эффекта инвестиционных проектов, планируемых и реализующихся на территориях Российской Федерации. Исходной информацией для проведения оценки послужили данные цифровой платформы «Инвестиционные проекты»¹.

Выстраивание цепочек инвестиционных проектов, формирующих потоки дополнительной стоимости, может создать импульсы роста за счет их синергетических эффектов. Так цепочки выстраиваются в мезоэкономические системы, такие как кластеры и сети и являются возможными драйверами будущего развития регионов.

Рассматривались инвестиционные проекты следующих отраслей: деревообработка, металлообработка, энергетика, добыча и переработка полезных ископаемых, химическая промышленность, машиностроение, строительные материалы. Также рассматривались проекты, объединенные в один фактор «качество жизнеобеспечения», куда вошли проекты, реализуемые в таких сферах, как жилые объекты, инженерные сети, инфраструктура, медицинские объекты, социальные объекты, переработка отходов. При этом в оценке участвовали инвестиционные проекты, находящиеся на стадиях планирования, предпроектных проработок, проектирования, подготовки к строительству, строительства, модернизации, ввода в эксплуатацию и реализуемых в рамках различных форм собственности (государственной, частной, государственно-частного партнерства).

¹ Право использования Программного обеспечения «Инвестиционные проекты» предоставлено ООО «ПКР Аналитика» по Лицензионному договору №119-11/21, заключенному при финансовой поддержке Минобрнауки РФ в рамках выполнения работ по крупному научному проекту «Социально-экономическое развитие Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий» (регистрационный номер заявки 2020-1902-01-377).

Выделение информативной системы факторов для построения когнитивной модели проводилось методом многомерного факторного статистического анализа. Анализ исходных данных проводился методом главных компонент (PCA, Principal Component Analysis). Данный метод представляет собой стандартный подход для анализа данных, в котором исходные данные проецируются на гиперплоскость меньшей размерности. Идея метода главных компонент была сформулирована К. Пирсоном в начале XX века [Pearson, 1901], после чего в тридцатых годах Г. Хотеллинг разработал схожий метод, делая упор на идею взятия линейных комбинаций переменных [Hotelling, 1933]. Сильной стороной метода можно считать возможность выявления неочевидных, и скрытых поначалу, закономерностей и взаимосвязей между исходными переменными. Также он позволяет понижать размерность пространства данных и строить ортогональный базис факторов.

Для реализации в модуле «Когнитивные модели» были выбраны: модель – метод главных векторов, данные – действующие инвестиционные проекты, количество главных векторов – 4, территория – вся территория РФ.

В табл. 5.1 приведены результат использования метода главных компонент. Метод факторного анализа позволил выделить четыре главных фактора, объясняющих 94% выборочной вариации исходных переменных (45%, 19%, 15%, 15%). Факторы имеют достаточно прозрачную содержательную интерпретацию.

Фактор Ф1 формируется инвестициями в объекты энергетики и химической промышленности. И это вполне закономерно, так как химическая промышленность – самая энергоемкая отрасль народного хозяйства. Выпуская 7% всей промышленной продукции, она потребляет 20% энергии.

Фактор Ф2 в основном определяется добычей и переработкой полезных ископаемых, которые связаны с дальнейшими переделами, в частности отраслями машиностроения, металлообработки, нефте-, газо- и углехимии.

Фактор Ф3 представлен комплексной переменной «качество жизнеобеспечения» и «строительные материалы». И это закономерно, поскольку все объекты, связанные с жизнеобеспечением –

это объекты строительства, их достаточно много, большинство из них входит в перечень национальных проектов.

И, наконец, фактор Ф4 образован инвестиционными проектами сферы дерево- и металлообработки. Что касается металлообработки, то это достаточно простые металлические изделия (метизы, прокат металлический, промышленная арматура, строительные металлоконструкции и изделия, трубы металлические и др.), большинство из которых производится непосредственно на металлургических заводах и используется в строительстве.

Таблица 5.1

Факторы и факторные нагрузки

Переменная	Фактор / процент объясненной вариации выборки			
	Ф1/45	Ф2/19	Ф3/15	Ф4/15
Деревообработка	-0,191	0,365	0,048	0,903
Энергетика	0,935	0,192	-0,039	-0,105
Добыча и переработка	0,124	0,943	0,083	0,275
Металлообработка	0,532	-0,119	0,221	0,768
Химическая промышленность	0,884	0,238	0,185	0,189
Машиностроение	0,390	0,860	0,305	-0,040
Строительные материалы	0,322	0,104	0,834	0,283
Качество жизни	-0,093	0,183	0,927	-0,023

Примечание. Получено авторами в результате статистической обработки данных цифровой платформы «Инвестиционные проекты».

Применение факторного анализа позволило представить анализируемую систему показателей обобщенным набором факторов, связь между которыми задается матрицей преобразования исходных переменных. На рис. 5.1 представлена такая система взаимодействующих факторов.

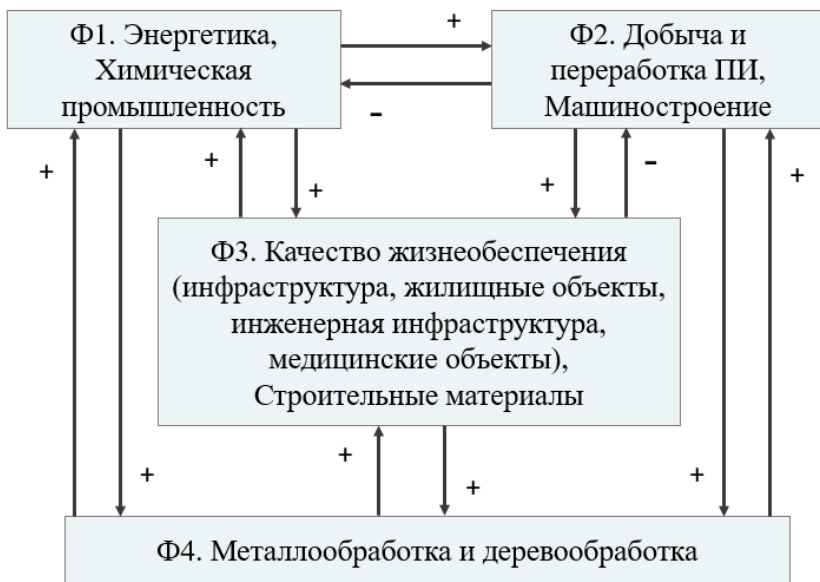


Рис. 5.1. Когнитивная схема взаимодействующих факторов

Эту систему взаимодействующих факторов принято называть когнитивной схемой. Идея когнитивного моделирования на основе когнитивной схемы приписывается американскому социологу и политологу Р. Аксельроду [Axelrod, 2015], который ввел понятие когнитивного моделирования, применяемого для проведения анализа и прогнозирования. В своих исследованиях он развил существовавшие ранее идеи построения когнитивных схем для формального представления знаний, как отмечает Гинис Л.А. [Гинис, 2005].

Отметим несколько важных моментов когнитивной схемы. Во-первых, это система взаимодействующих факторов, и она может быть названа сложной системой. Во-вторых, связи в системе не являются симметричными. Так Ф1 положительно влияет на Ф2, т.е. развитие энергетической структуры с большей вероятностью приводит к освоению полезных ископаемых в районах нового освоения, что не всегда обеспечивает обратное. Аналогичное заключение можно сделать и относительно взаимовлияний факторов Ф2 и Ф3. В-третьих, когнитивная схема представляет не что

иное, как ориентированный граф, поскольку дуги имеют направление, а при наличии весов влияния ориентированный граф представляет из себя взвешенный ориентированный граф, исследуя который можно решить и задачу прогнозирования.

Как отмечается в работе Морозовой М.Е., Шмат В.В. [Морозова, Шмат, 2017], с использованием когнитивной схемы можно проводить следующие исследования (их также можно считать стадиями когнитивного моделирования):

- Статический анализ: моделирование самостоятельного развития системы без управляющих воздействий. На данном этапе рассматривается строение системы, находящейся в состоянии покоя, т.е. без подачи импульса. Данный этап состоит в составлении когнитивной схемы, ее проверки и модификации, чтобы она в достаточной мере описывала исследуемую систему.
- Динамический анализ: изучение управляемого развития ситуации – исследователь направляет определенные импульсы (воздействия) на факторы. С помощью направления импульсов разной силы на разные факторы исследуются произошедшие изменения, т.е. отклики системы. Варьируя вносимые изменения, ученый может лучше исследовать моделируемую систему, проверить свои гипотезы и выработать новые. На этом этапе можно ответить на поставленный исследователем вопрос «Что будет, если ...?».

В базе знаний после использования метода главных компонент полученную когнитивную таблицу можно просмотреть во вкладке «Таблица взаимосвязей» (рис. 5.2).

Информация о ГВ		Таблица взаимосвязей		Импульсы		Сравнение по факторам для ГВ	
		Комп. 1		Комп. 2		Комп. 3	Комп. 4
1		0.597		0.543		0.356	0.376
2		-0.468		0.191		0.397	0.464
3		0.22		-0.401		-0.089	0.512
4		0.066		0.482		0.792	-0.369

Рис. 5.2. Когнитивная таблица взаимодействующих факторов в базе знаний

В данной работе был проведен как статический, так и динамический анализ. Результатом выполнения статического анализа является составленная когнитивная таблица (см. рис. 5.2). Процедура проведения и полученные результаты динамического анализа представлены далее.

Представляет интерес и задача импульсного моделирования, решаемая с помощью ориентированного графа. Например, задавая значения возмущений в какой-либо вершине графа, можно определить изменения в других его вершинах. Здесь в отличие от задачи анализа, возможно решить задачу прогноза – определить, к чему приведет воздействие в сложной системе, заданное изменениями одного или нескольких управляющих факторов. Таким образом, проводится моделирование распространения управляющего воздействия по системе, заданной когнитивной схемой, в результате которого изменяется состояние системы (значения факторов). Подобная динамическая задача решается с использованием импульсного процесса [Робертс, 1986]. Смысл его заключается в следующем.

Рассмотрим сущность импульсного процесса с математической стороны, основываясь на работе Захаровой Е.Н. [Захарова, 2007]. Изначально система находится в состоянии покоя и задается таблицей взаимосвязей (матрицей смежности) $A_G = [a_{ij}]_{k \times k}$, $i, j = 1, 2, \dots, k$. Также определены начальные значения факторов, т.е. задан вектор начального состояния $X(0) = (x_1(0), \dots, x_k(0))$. Для изучения системы исследователь подает начальный импульс, который переводит систему из начального состояния $X(0)$ в новое состояние $X(1) = (x_1(1), \dots, x_k(1))$. Подаваемый в систему импульс – по сути, вектор приращений значений факторов, имеет вид $P(0) = (p_1(0), \dots, p_k(0))$, где $p_i(0) = x_i(1) - x_i(0)$ – начальный импульс фактора i , который переводит значение фактора из начального состояния $X(0)$ в следующее состояние $X(1)$.

Для прогнозирования развития системы используются уравнения вида $P(t+1) = A_G P(t)$, где $P(t+1)$ и $P(t)$ – это вектора приращений значений факторов в последовательные дискретные моменты t и $t+1$, а их совокупность называется импульсным процессом. Состояния системы $X(0), X(1), \dots, X(n)$ отображают динамику изменений системы при управляющих воздействиях $P(0)$. Компоненты $X(n)$ характеризуют прогнозируемые значения фактора, и для их вычисления используется формула $x_i(n) = x_i(0) + \sum_{t=0}^n p_i(t)$, т.е. производится суммирование начального значения фактора с накопленными приращениями значений фактора за n итераций (тактов) моделирования.

Также можно выделить ряд особенностей проводимого импульсного моделирования, упоминаемых в статье Белан А.К., Шмат В.В. [Белан, Шмат, 2015]:

- С содержательной стороны, система, находящаяся изначально в состоянии покоя, подвергается некоторому воздействию (импульсу), после чего, из-за распространения импульса по системе, она приходит в движение. При этом корректируются значения факторов. Когда распространение импульса затихает, система приходит в новое равновесное состояние, которое является отражением прогноза сценария развития системы. В случае взрывного процесса равновесие недостижимо, но в экономике взрывные процессы встречаются крайне редко.

- Стандартным техническим решением когнитивного моделирования является рекуррентное вычисление $X(1), X(2), \dots, X(n)$. Причем количество итераций n устанавливается достаточное для проверки сходимости вычислений, т.е. нужно понять, получаются ли стабильные значения, демонстрирующие характер изменения системы.

- Одна итерация моделирования (переход от $X(i)$ к $X(i+1)$) обычно не соотносится с конкретным временным промежутком. Однако его можно вычислить посредством качественного анализа реальных условий и особенностей функционирования моделируемой системы.

В работе используется описанная выше процедура импульсного моделирования для расчета четырех сценариев развития системы на основании разного распределения инвестиций.

Для проведения импульсного моделирования в базе знаний нужно во вкладке «Импульсы» задать импульсный вектор, показывающий, в какие факторы мы будем подавать импульс. Например, для подачи импульса во второй фактор нужно задать вектор $[0,1,0,0]$ (рис. 5.3). В итоге будет отображен результат 50 итераций схождения модели. Если модель не сходится в точку, значит подобранная когнитивная схема нестабильна.

Импульсный вектор

0,1,0,0

Таблица эффектов после 50 итераций

Комп. 1	Комп. 2	Комп. 3	Комп. 4
-0.15	0.94	1.04	0.91

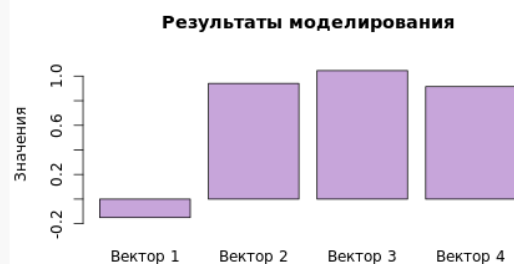
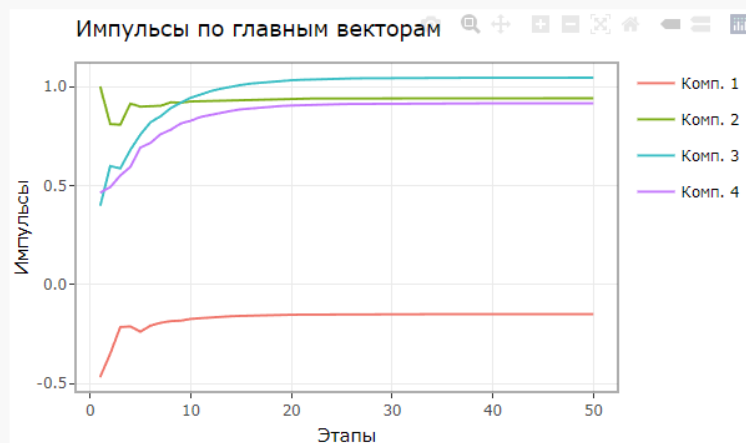


Рис. 5.3. Отображение результата импульсного моделирования в базе знаний

Напомним, что в исследовании участвуют актуальные к настоящему времени инвестиционные проекты цифровой платформы «Инвестиционные проекты», в перечисленных ранее отраслях на территории Российской Федерации. Когнитивная модель представлена взвешенным ориентированным графом, схема которого изображена на рис. 5.1. Импульсные воздействия подаются в определенные вершины графа согласно рассматриваемым ниже сценариям. Поскольку все измерители в вершинах графа сопоставимы (инвестиции представлены в одних денежных единицах), то и результаты воздействий также сопоставимы.

Рассмотрим сценарии инвестиционного поведения.

Сценарий 1. Инвестиции осуществляются преимущественно в сферу добычи и переработки полезных ископаемых.

В когнитивной модели импульс подается в вершину, представленную фактором Ф2. Поскольку все вершины взаимосвязаны, то результат импульса скажется в разных вершинах по-разному. В качестве целевой рассмотрим вершину Ф3 (качество жизнеобеспечения), поскольку и в национальных, и в региональных программах социально-экономического развития эти проблемы являются наиважнейшими. Мультипликативный эффект от единичного импульса в вершину Ф2 представлен на рис. 5.4. Имитационный процесс когнитивной модели сходится уже на 15 шаге. Решение свидетельствует, что единичный прирост инвестиций в добычу и обогащение ПИ приводит к приросту инвестиций в качество жизни в размере 1,03.

Сценарий 2. Преимущественно инвестиции осуществляются в энергетику и химическую промышленность.

В когнитивной модели единичный импульс подается в вершину Ф1. Сходимость к решению наблюдается на 17 шаге (см. рис. 5.4). При этом мультипликативный эффект от проектов в сферах энергетики и химической промышленности составляет 2,37, что более чем вдвое превышает мультиплицирующий эффект инвестиций в добычу и переработку ПИ.

Сценарий 3. Преимущественное инвестирование в проекты дерево- и металлообработки.

В когнитивной модели импульс подается в вершину Ф4. Сходимость итерационного процесса наблюдается к 17 шагу (см. рис. 5.4). При этом единичный прирост инвестиций в проекты дерево- и металлообработки мультиплицируют прирост инвестиций в проекты качества жизни и строительства на уровне 2,56.

Сценарий 4. Развитие промышленного комплекса.

Сценарий предполагает одновременное импульсное воздействие в вершины Ф1, Ф2, Ф4 (инвестиции растут в энергетику, химическую промышленность, добычу и переработку ПИ, металло- и деревообработку, машиностроительные проекты. Это приводит к мультипликации инвестиций в строительство, инфраструктуру, качество жизни до уровня 5,96 (см. рис. 5.4). Т.е. 3 условных единицы прироста в инвестирование проектов промышленного комплекса дают практически 2-кратный прирост инвестиций в обобщенный показатель качества жизни.

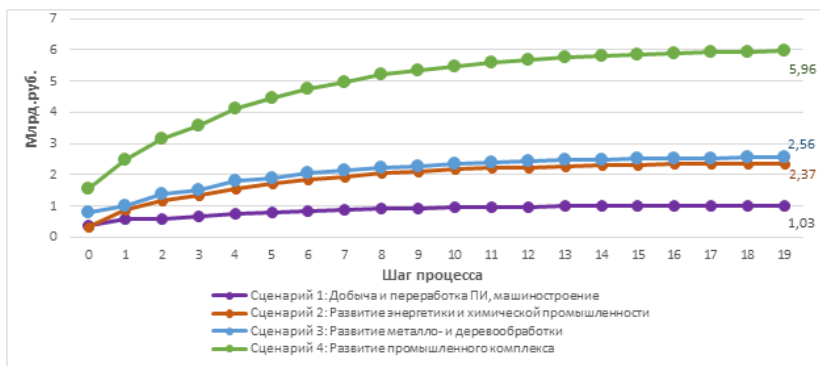


Рис. 5.4. Влияние различных сценариев распределения прироста инвестиций на прирост инвестиций в проекты качества жизни и строительства

Если еще и дополнительно наращивать инвестиции в проекты транспортной, коммунальной, социальной инфраструктуры, то, как показывают расчеты, эффект приращения инвестиций в проекты качества жизни может составить 8,38, т.е. более чем вдвое.

5.2. Влияние цифровизации городской среды на высокотехнологичные компании

В разделе описан пример использования блока информационного обеспечения для оценки взаимного влияния высокотехнологичных и наукоемких компаний (ВТиНК) и цифровизации городской среды. Результаты исследования представлены в статье Костиной Е.А. и Костина А.В. [Костина, Костин, 2023].

Для выстраивания взаимосвязей в исследовании появилась необходимость анализа зависимости результатов деятельности ВТиНК и факторов, описывающих уровень развития цифровизации городской среды. В качестве последнего использовался ежегодно рассчитываемый показатель IQ города, но он привязан именно к городам, а не регионам, поэтому для их изучения необходимо было сделать аналитический срез деятельности ВТиНК в границах определенного количества городов.

Возможны несколько путей решения задачи через базу знаний.

Первый путь заключается в том, что в разделе «Данные компаний» нужно скачать данные с использованием ограничений в виде выбора ОКВЭД для ВТиНК. В системе выбираются только действующие компании и добавляются таблицы показателей финансовой отчетности за весь временной период с 2005 г. Для привязки компаний к конкретному городу можно использовать или ГИС-координаты и проверку нахождения компании в полигоне города, или поиск названия города в адресе компании. Так же были убраны и просмотрены сомнительные компании, у которых были выявлены ошибки в отчетности, описанные в разделе 2.1. Результатом использованных процедур является сформированная Excel таблица данных компаний с привязкой к нужным городам и финансовыми показателями. Далее использовалась функция агрегации суммой в языке программирования R, итогом применения которой явились суммарные показатели по городам и годам.

Второй путь будет возможен после последнего обновления базы знаний. Он заключается в использовании пространственного анализа. Для этого в блоке «ГИС-моделирование» нужно

будет добавить компании на карту с такими же ограничениями, что и в первом решении. После этого выделить область вокруг нужных городов и нажать «Построить аналитический отчет», в результате будет получен аналитический срез по выбранной территории, в том числе необходимые для регрессионного анализа суммарные показатели. После анализа каждого города и списка формируется база данных, идентичная полученной первым путем.

Полученные данные использовались в описательном и регрессионном анализе для выявления взаимосвязи финансовых показателей ВТиНК компаний и различных факторов, в том числе показателя IQ городов. Проведенные расчеты подтвердили, что предполагаемая зависимость между цифровизацией городской среды и успешной деятельностью ВТиНК оказалась значима, в большей степени – для крупнейших городов.

5.3. Применение базы знаний для комплексной оценки возможностей реализации цепочек инвестиционных проектов на примере развития черной металлургии в Азиатской России

В разделе описан пример использования блока ГИС-моделирования для оценки свойств цепочек инвестиционных проектов на примере развития черной металлургии Азиатской России. Результаты исследования представлены в работах Петрова С.П. [Петров, 2022; 2023]. В них проанализирована цепочка инвестиционных проектов: создания производства по технологии прямого восстановления железа (ПВЖ) на базе Кимканово-Сутарского ГОКа – расширения производства, на ООО «Амурсталь» – строительства металлургического завода при ССК «Звезда» с годовой потребностью стали 330–350 тыс. т. Для первичного анализа требовалось изучить возможности выстраивания подобной цепочки.

Анализ инвестиционных проектов показывает наличие проекта на 164 млрд руб. строительства металлургического завода при ССК «Звезда» в городе Большой Камень. В карточке проекта описано, что по состоянию на начало 2022 г. проект

находится на стадии предпроектных проработок. Решаются вопросы, связанные со строительством электростанции, созданием мощностей для газификации завода и освоением водной акватории (рис. 5.5). Также в базе данных есть проект создания производства горячебрикетированного железа на базе Кимкано-Сутарского горно-обогатительного комбината на сумму 26 млрд руб. По состоянию на начало 2022 г. проект находится на стадии планирования.

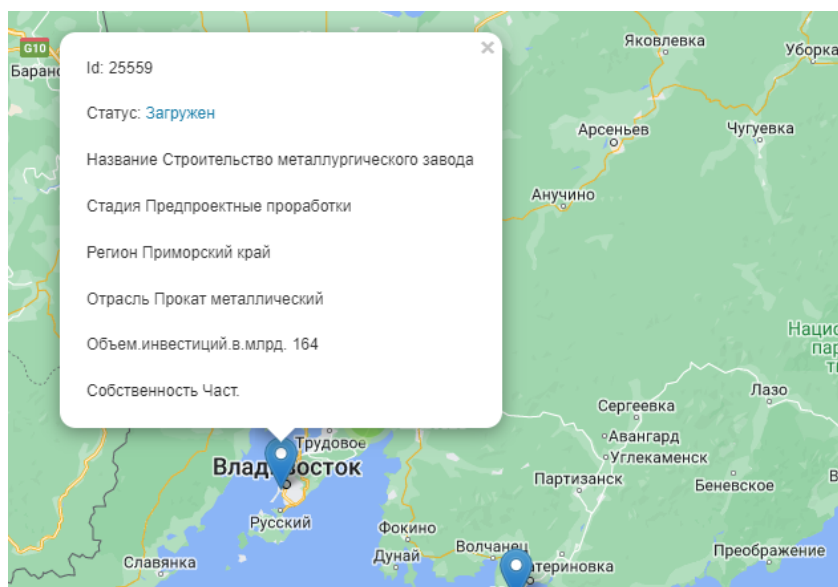


Рис. 5.5. Анализ инвестиционного проекта в блоке “ГИС-моделирование”

Требовался анализ транспортной связанности агентов и возможности их газификации. Для этого была построена карта железных дорог и газопровода (рис. 5.6). Видно, что все агенты связаны существующей железной дорогой. Действующий газопровод идет через Комсомольск-на-Амуре и Владивосток, поэтому возможна газификация ООО «Амурсталь» и металлургического завода при ССК «Звезда», но Кимкано-Сутарский ГОК не газифицирован. Ближайшие к нему линии газопровода в Хабаровске

и Благовещенске, что демонстрирует наличие проблемы, требующей дополнительного анализа. В качестве способа решения проблемы был рассмотрен вариант организации производства ПВЖ по технологии, позволяющей вместо природного газа использовать уголь.

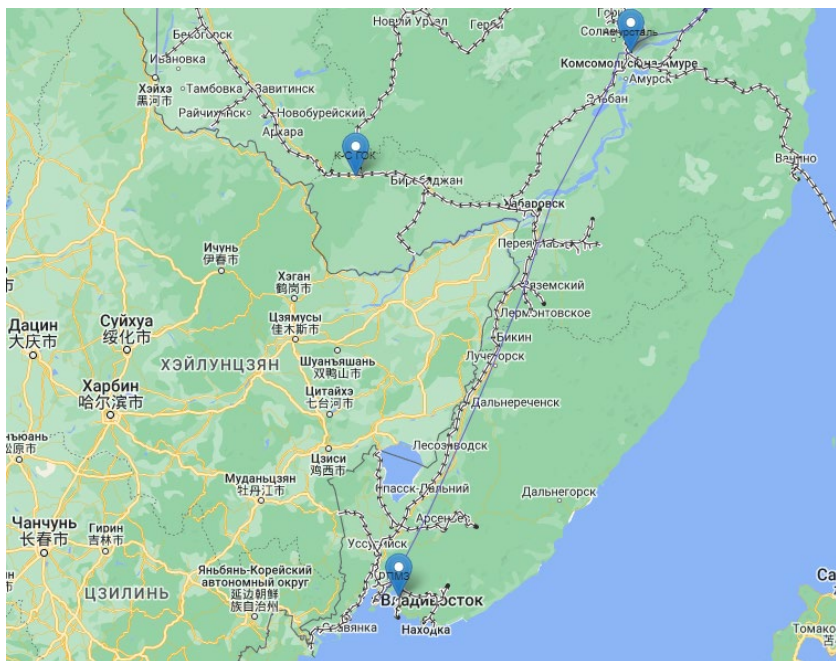


Рис. 5.6. Газопроводы и железные дороги в блоке “ГИС-моделирование”

Анализ угольных месторождений показал наличие вблизи Кимкано-Сутарского ГОКа запасов угля в районе Чегдомына (рис. 5.7) к которому проведена железная дорога. Калькулятор перевозок, построенный на основе АРИ РЖД позволил просчитать стоимость транспортировки угля до Кимкано-Сутарского ГОКа.

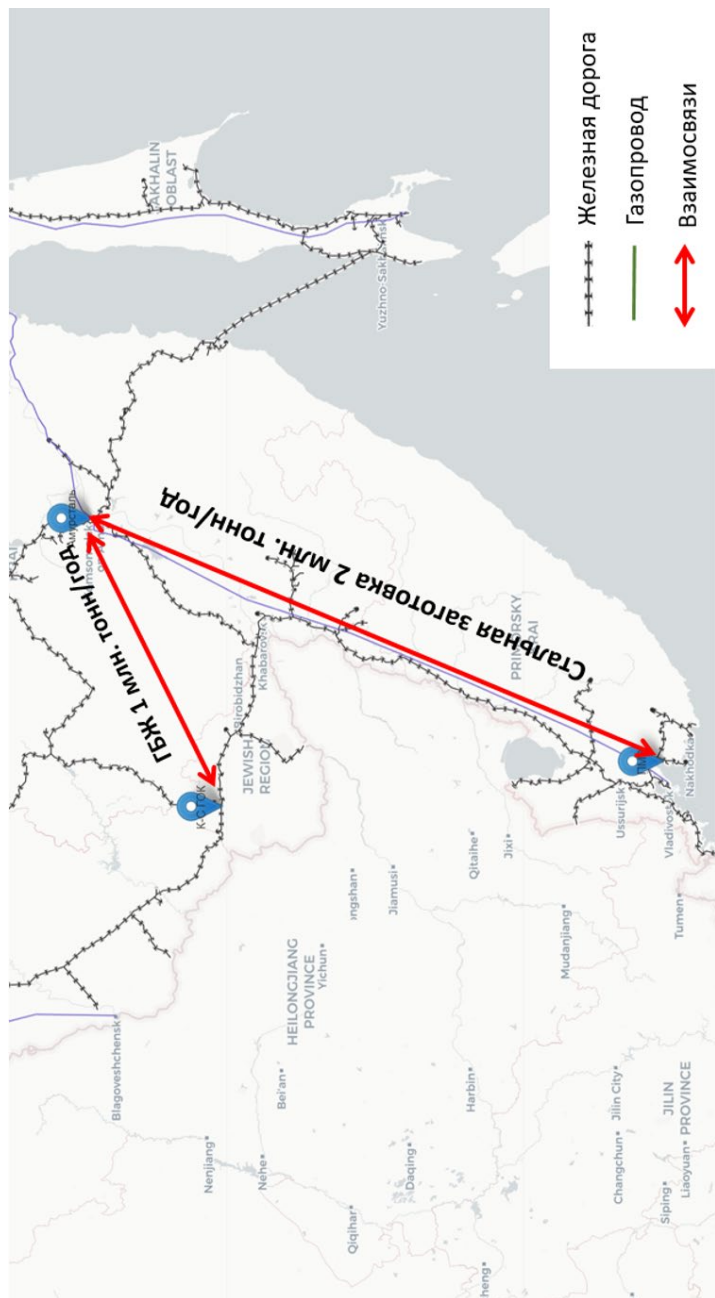


Рис. 5.7. Цепочка инвестиционных проектов черной металлургии

Изучение линий электропередач показывает наличие электрификации агентов и точки генерации электроэнергии. Для анализа возможностей увеличения потребления электроэнергии и выпуска железа Кимкано-Сутарским ГОКом в блоке «Таможенная статистика» был изучен экспорт по этим категориям. Анализ показал большой объем экспорта руды и железных концентратов из ЕАО и электроэнергии из Амурской области.

Исследования Петрова С.П. [Петров, 2022] показали высокую перспективность формирования металлургического производства на Дальнем Востоке на основе ПВЖ (см. рис. 5.7).

Рассмотренный кейс показывает, что использование Базы знаний ИЭОПП СО РАН, позволяет проводить первичный анализ инвестиционных проектов и обнаружить на начальных этапах возможные проблемы, связанные с расположением ресурсов, инфраструктурных объектов, проектируемых предприятий и логистических цепочек.

5.4. Оценки инвестиционного импульса с использованием прогностического комплекса ДММ-КАМИН

Система ДММ-КАМИН, описанная в разделе 3.1 и 4.2 и внедренная в Базу знаний, позволяет оценивать долгосрочное влияние реализации инвестиционных проектов на макроэкономические показатели. Результаты межотраслевых взаимодействий в рамках ДММ-КАМИН позволяют провести детальный анализ не только для агрегированных показателей на уровне национальной экономики, но и проанализировать отраслевой срез связанных с реализацией проектов изменений. Подобная работа была проведена сотрудниками ИЭОПП СО РАН д.э.н. Барановым А.О. и д.т.н. Павловым В.Н. Были оценены эффекты от реализации инвестиционного проекта в отрасли «Химическое производство» в рамках развития импортозамещения. Объемы предполагаемых инвестиций, описанные в базе данных инвестиционных проектов, составили 22 млрд руб.

На первых этапах был проведен сбор данных для моделирования и рассчитаны первичные эффекты от реализации инвестиционного проекта на показатели роста выпуска базовой продукции анализируемого проекта. Первоначально был выполнен прогноз без учета реализации инвестиционного проекта для сравнения результатов. Затем с учетом изменений в параметры объема инвестиций в основной капитал и объема выпуска химической отрасли реализована повторная оптимизация и получен прогноз на 2023–2035 годы, учитывающий реализацию проекта. Полученные результаты можно описать как «Базовый сценарий» и «Сценарий с учетом инвестиционного импульса».

Расчеты показали, что реализация инвестиционного проекта увеличит валовый выпуск отрасли химического производства на 236 млрд руб. суммарно за период с 2023 по 2035 год, а в 2035 г. разница валового выпуска отрасли в варианте с проектом по сравнению с вариантом без проекта составит 26 млрд руб. (рис. 5.8). С учетом строительного лага основной эффект будет достигнут после 2026 г.

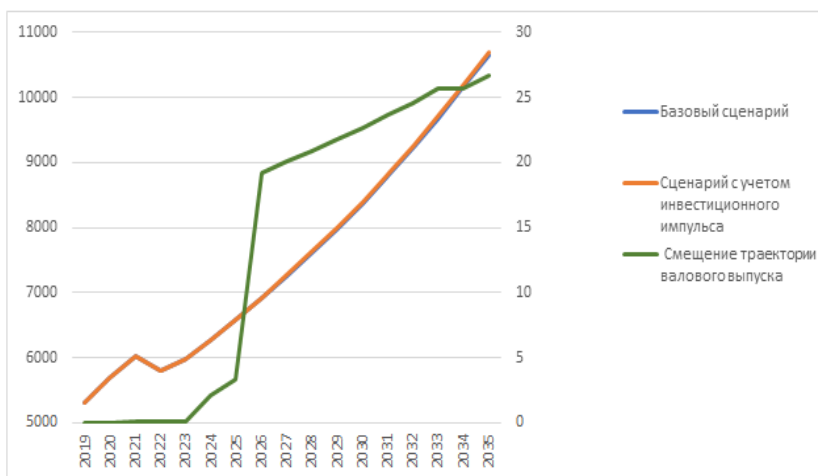


Рис. 5.8. Валовый выпуск отрасли химического производства в сопоставимых ценах с учетом инвестиционного импульса и без него. Разница сценариев отобразена на вспомогательной оси, млрд руб.

Для экономики в целом, с учетом межотраслевых взаимодействий, эффект будет более существенным. Валовая добавленная стоимость РФ по сравнению с базовым сценарием вырастет на 496 млрд руб. суммарно за период с 2023 по 2035 год, а в 2035 г. разница ВВП составит 41 млрд руб. (рис. 5.9). Эффект проявится уже с 2023 г. с учетом задействования строительной отрасли в реализации инвестиционного проекта.

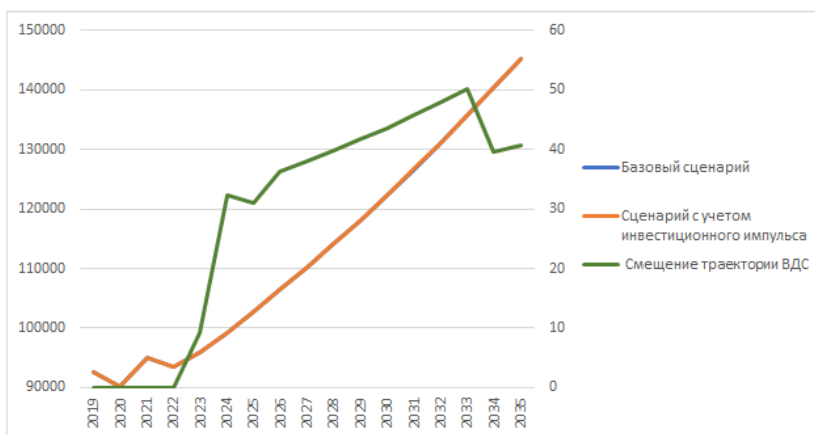


Рис. 5.9. Валовая добавленная стоимость РФ в сопоставимых ценах с учетом инвестиционного импульса и без него.

Разница сценариев отображена на вспомогательной оси, млрд руб.

Анализ расчетов показывает, что на уровне страны инвестиционный проект существенно не изменяет траекторию абсолютных макропоказателей. Но анализ мультипликативных эффектов показывает, что мультипликатор инвестиций, связанных с реализацией проекта, по ВВП составляет 23 раза, что является хорошим результатом. Также система ДММ-КАМИН позволяет сделать разбиение роста валового выпуска и ВВП по отраслям и выделить отрасли, в которых имел место наибольший эффект от реализации проекта. Таковыми отраслями стали (по уровню убывания эффекта):

- «Химическое производство»,
- «Оптовая и розничная торговля, ремонт, гостиницы и рестораны»,

- «Операции с недвижимым имуществом и предоставление услуг (за исключением НИОКР)»,
- «Государственное управление и обеспечение военной безопасности»,
- «Предоставление прочих коммунальных, социальных и персональных услуг»,
- «Производство нефтепродуктов»,
- «Обеспечение электрической энергией, газом и паром»,
- «Производство пищевых продуктов, напитков и табака».

Реализованный в рамках Базы знаний инструментарий ДММ-КАМИН позволяет оценивать и сравнивать не только эффекты от инвестиционных импульсов, но и воздействие вариации инструментов макроэкономической политики и внешнеэкономических экстерналий на динамику макро- и отраслевых показателей национальной экономики и отдельных регионов.

5.5. Предварительный анализ крупных инфраструктурных инвестиционных проектов

Система ГИС-моделирования базы знаний позволяет не только помогать делать первичный анализ цепочек инвестиционных проектов с помощью анализа наличия инфраструктуры, спроса и ресурсного ограничения, как было описано в разделе 5.3, но и проводить первичный анализ инфраструктурных инвестиционных проектов.

Примером такого инфраструктурного инвестиционного проекта является строительство Северо-Сибирской магистрали, как один из ключевых проектов, анализируемых группой ученых ИЭОПП СО РАН во главе с Маловым В.Ю. и Тарасовой О.В. [Новый импульс..., 2022, с. 344–412]. Северосибирская магистраль – участок железной дороги, проходящий севернее рек Ангара, Кеть и Обь от города Усть-Илимска до города Нижневартовск (рис. 5.10). Анализ существующей железнодорожной сети показывает пересечение с Северо-Сибирской магистралью вблизи следующих городов и сел Усть-Илимск, Ангарский (Соленый), Лесосибирск, Степановка, Белоярское и Нижневартовск, что ха-

рактизует связывающий эффект планируемой железной дороги Азиатской России.

Важной частью первичного анализа такого проекта является оценка характеристик минерально-сырьевой базы в зоне влияния, которые становятся доступными для освоения при создании новой транспортной инфраструктуры. В Базе знаний полезные ископаемые делятся на твердые полезные ископаемые и углеводородное сырье. Рассмотрение твердых полезных ископаемых показывает наличие двух кластеров в зоне влияния магистрали: Ангаро-Енисейского и юга Ангаро-Анабарского кластера. В Ангаро-Енисейском кластере находятся в основном золоторудные месторождения, месторождения железа, свинца и цинка, а также потенциальные месторождения бурого угля. На юге Ангаро-Анабарского кластера преобладают железорудные месторождения и потенциальные алмазные рудные узлы (рис. 5.11).

Углеводородное сырье представлено Байкитско-Катангским, Юго-Восточным (Томским) и юго-востоком Западно-Сибирского кластеров (рис. 5.12).

С использованием Базы знаний можно оценить, какие объемы и каких полезных ископаемых станут доступными для освоения в результате создания магистрали:

- нефть – 312 млн т подготовленных объектов перспективно-го кластера и 893 млн т запасов действующих кластеров,
- газ – 1256 млрд куб. м запасов действующих кластеров,
- железо – 3713112 тыс. т руды с содержанием 1290804 тыс. т металла,
- золото – 1242 т золота,
- свинец – 77571 тыс. т руды с содержанием 5150 тыс. т металла.

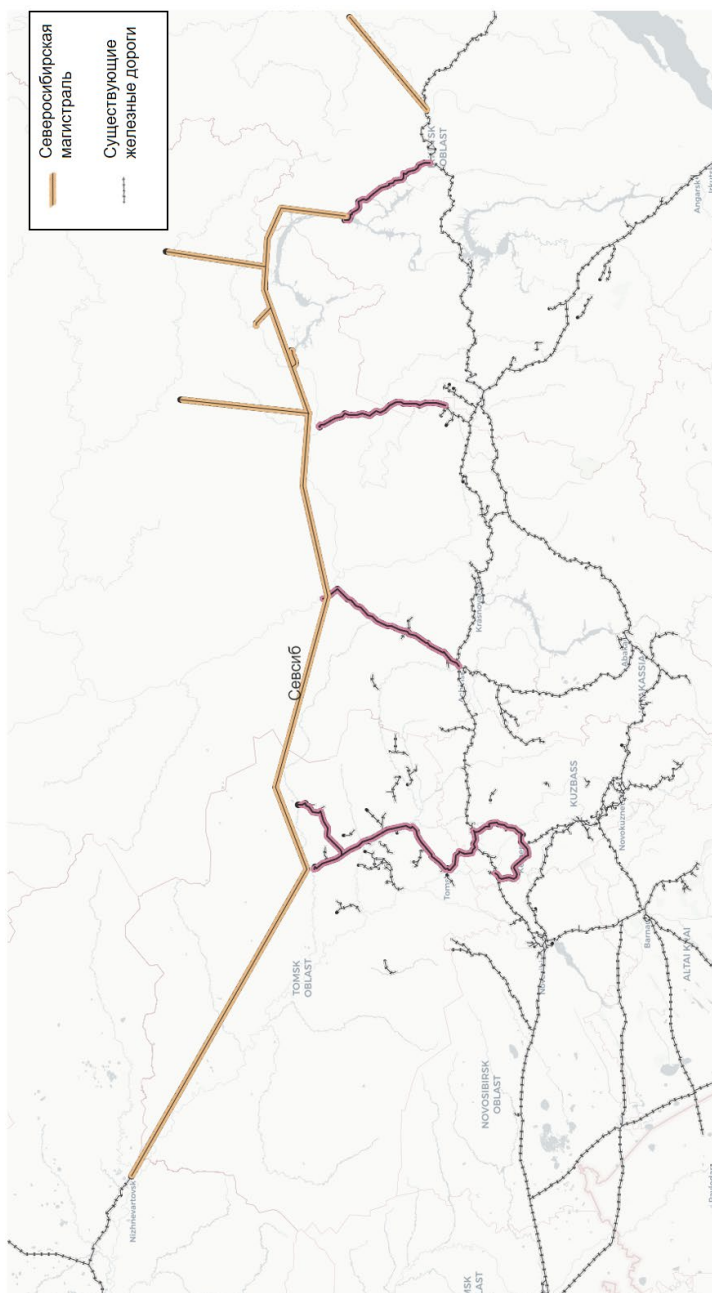


Рис. 5.10. План сопряжения Северо-Сибирской магистрали с сетью существующих железных дорог

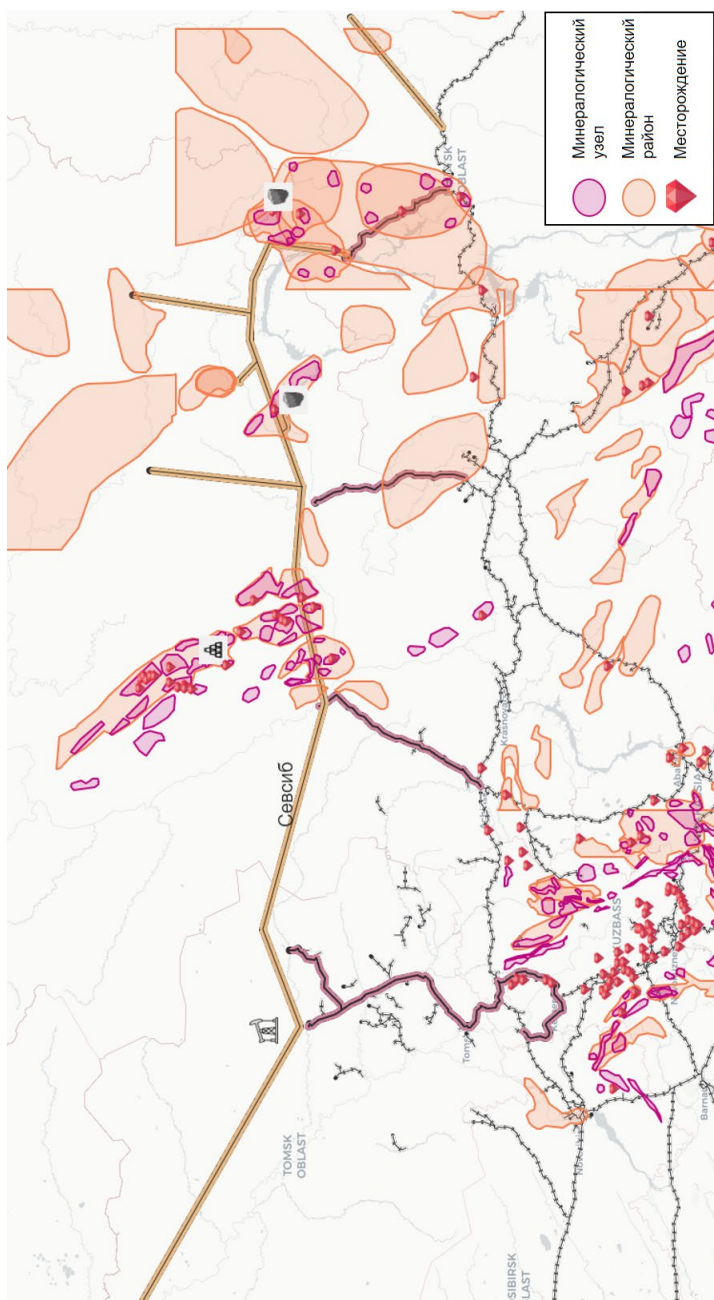


Рис. 5.11. Точечные месторождения, минералогические узлы и районы Северо-Сибирской магистрали

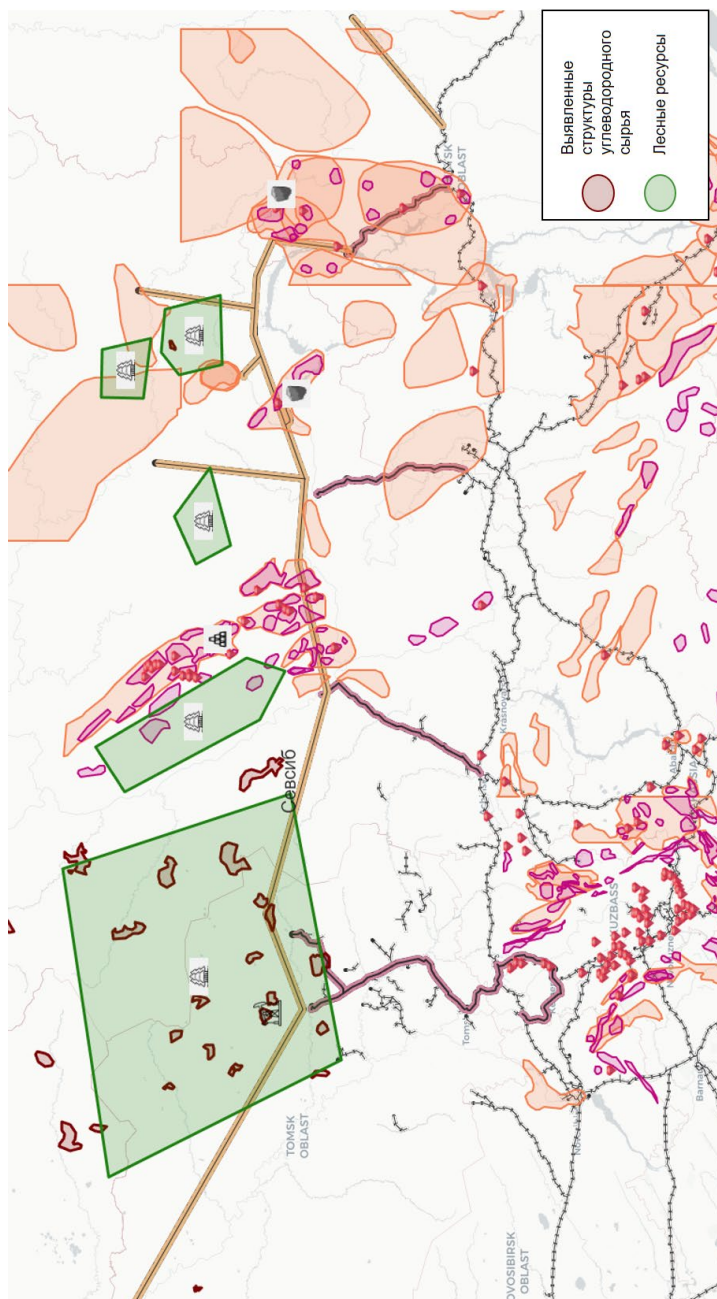


Рис. 5.12. Выявленные структуры углеводородного сырья и лесные ресурсы в зоне влияния Северо-Сибирской магистрали

Важной добавкой к потенциальному грузообороту Северо-Сибирской магистрали могут стать лесные ресурсы. Анализ данных Базы знаний позволил визуализировать точки их концентрации (см. рис. 5.12).

Привязка компаний и инвестиционных проектов к ГИС-координатам позволяет провести пространственный анализ финансовых показателей компаний и показателей инвестиционных проектов, которые уже действуют или реализуются на территории в зоне влияния магистрали (рис. 5.13 и 5.14). На территории реализуются 243 инвестиционных проекта с суммарным объемом инвестиций в 788 млрд руб. Основными отраслями, в которых реализуются инвестиционные проекты, являются обработка древесины и цветная металлургия. В цветной металлургии модернизируются и строятся горно-обогаительные комбинаты для разработки золоторудных месторождений и месторождений свинцово-цинковых руд возле Новоангарска. Инвестиционные проекты в лесном комплексе включают строительство целлюлозно-бумажных комбинатов, создание и модернизацию лесопильных комплексов, строительство заводов по производству картона, CLT-панелей, OSB-плит и др. Описанная конституция пула инвестиционных проектов подтверждает проведенный ранее анализ лесных ресурсов и твердых полезных ископаемых.

На территории в зоне влияния Северо-Сибирской магистрали на 2020 г. зарегистрировано 13853 действующих компаний с суммарной выручкой 1,52 трлн руб. и прибылью 478 млрд руб. То есть совокупный средний показатель ROS (рентабельность продаж) компаний составляет 31,4%. Более детальный анализ показал, что 440 из 478 млрд руб. суммарной прибыли принадлежит одной компании – АО «ПОЛЮС КРАСНОЯРСК». Крупнейшими компаниями, располагающимися на прилегающих к магистрали территориях, помимо АО «ПОЛЮС КРАСНОЯРСК», являются нефтегазовые: «Самотлорнефтегаз», «Томскнефть восточной нефтяной компании», «Славнефть-мегионнефтегаз», «ННК – Нижневартовское нефтегазодобывающее предприятие» и «Сибуртюменьгаз», с суммарной выручкой 800 млрд руб.

Предложенный подход можно использовать для первичного анализа крупных инфраструктурных проектов или совокупности инвестиционных проектов, формирующих цепочки добавленной ценности. При этом преимуществом созданной базы является возможность увидеть, как сам транспортный проект и связанные с ним производственные проекты, вписываются в планы и перспективы развития Азиатской России и страны в целом. Так, на рис. 5.15 видно, что Северо-Сибирская магистраль должна стать частью нового широтного транспортного коридора – Северо-Российской Евразийской магистрали (Ванино – Байкало-Амурская магистраль – Севсиб – Баренцкомур – Индига) – активно обсуждаемого в настоящее время на самом высоком уровне управления страны.

На текущий момент основными направлениями развития сформулированной задачи является накопление данных и знаний для формирования стратегии пространственного развития территорий Азиатской России на основе синергии транспортной доступности, системных знаний о природно-ресурсном потенциале, расширяющегося пространства межрегиональных взаимодействий.

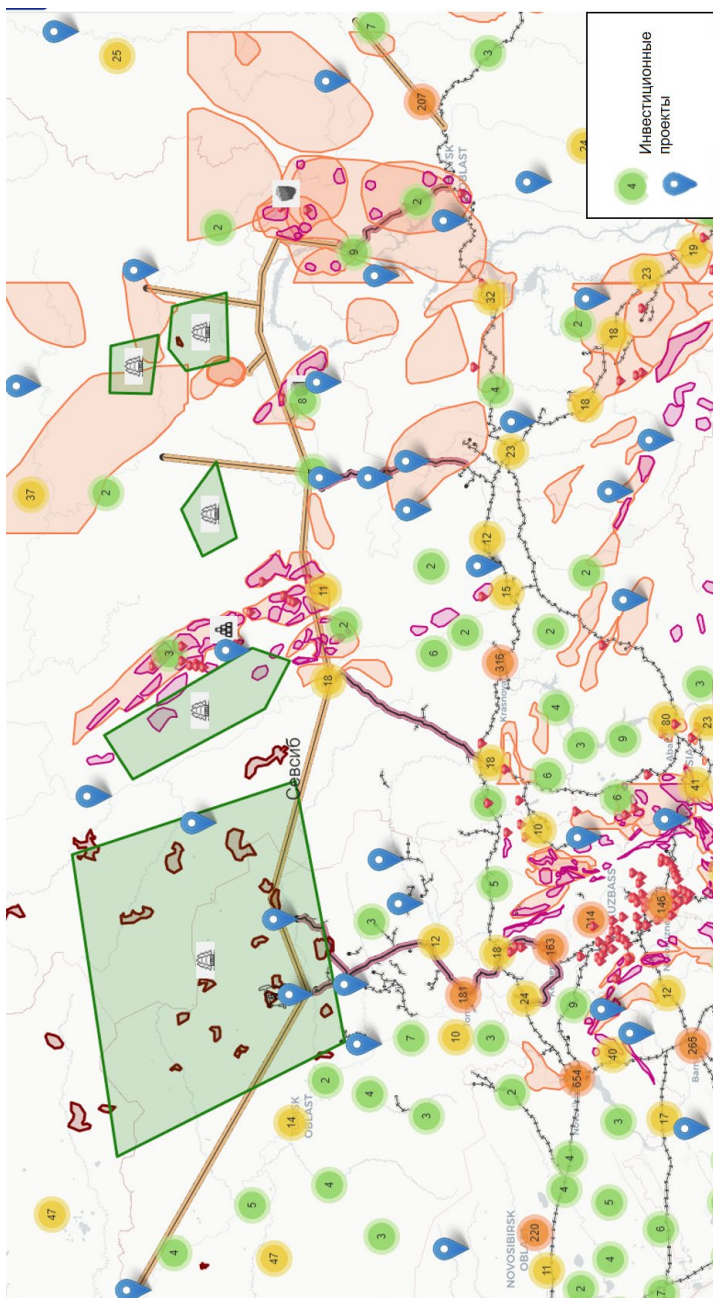


Рис. 5.13. Инвестиционные проекты, реализуемые в зоне влияния Северо-Сибирской магистрали

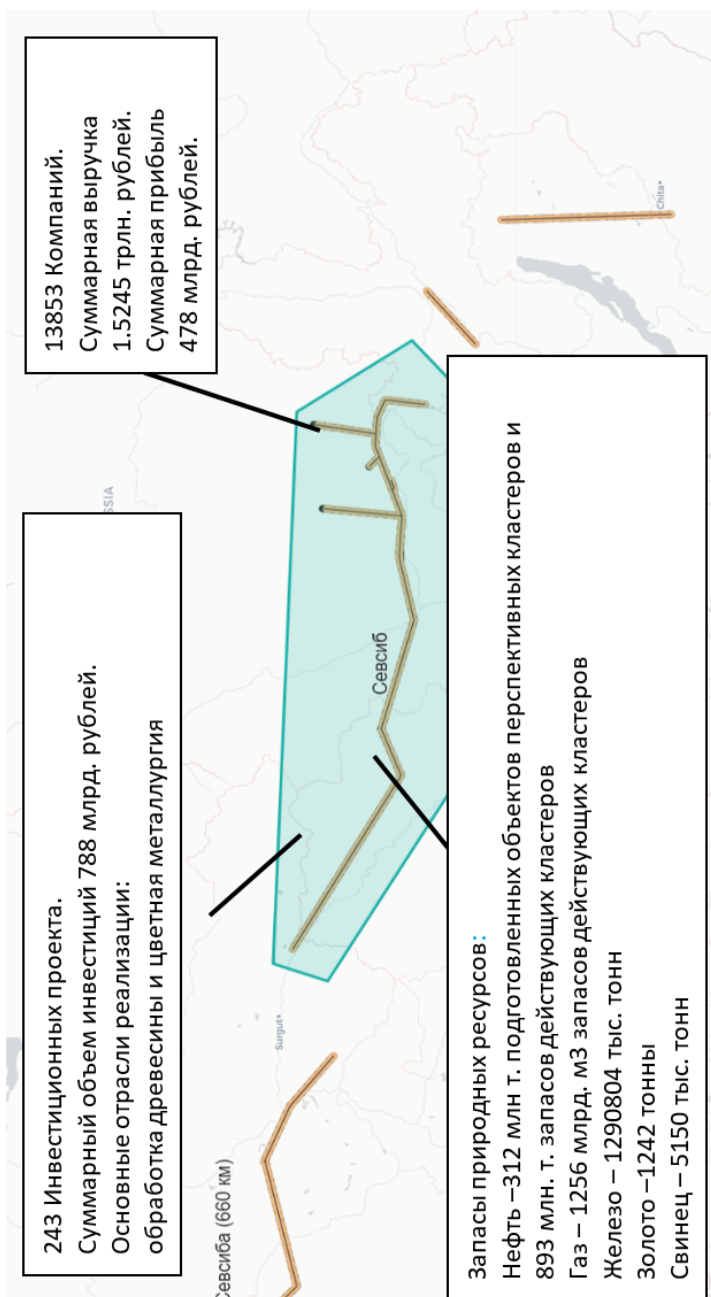
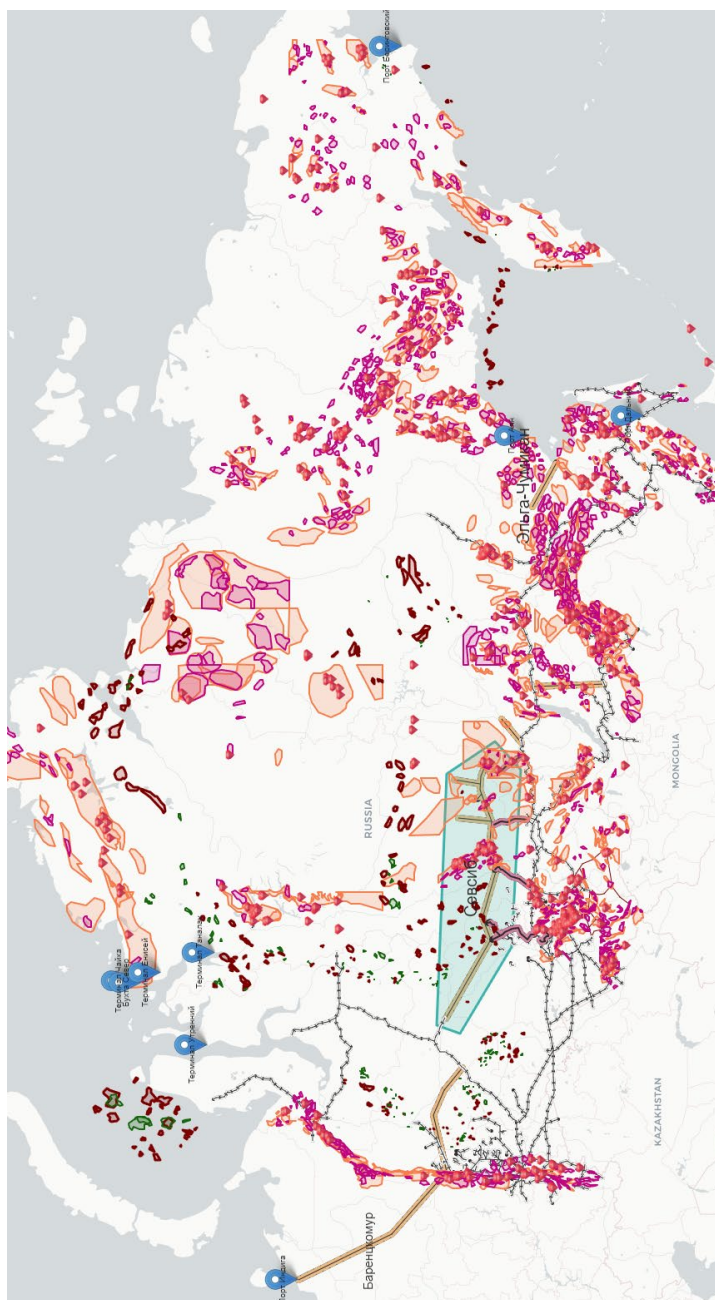


Рис. 5.14. Основные инвестиционные проекты в зоне влияния Северо-Сибирской магистрали и их потенциал



ЗАКЛЮЧЕНИЕ

В результате выполненных исследований получены следующие результаты:

1. Создана и апробирована информационная система, представляющая взаимосвязанную совокупность средств, методов и моделей, используемых для хранения, обработки и выдачи информации в интересах достижения поставленной цели – создания Базы знаний для разработки сценариев развития социально-экономических систем Азиатской России на основе взаимодействия природно-ресурсного, технологического и человеческого потенциалов, обеспечения институциональной, инвестиционной и инфраструктурной гармоний.

2. В рамках информационной системы разработана база данных, в которой, исходя из анализа потока данных, сформированы две ее разновидности: структурированная и неструктурированная БД. Созданы инструменты, позволяющие работать с базами данных, включающие сбор и хранение данных, поиск, фильтрацию, экспорт, первичную обработку и др.

3. Сформированы основные элементы развиваемой базы знаний, включающей:

- Блок информационного обеспечения (компании, секторы, комплексы, регионы, кластеры), содержащий информацию макро-, мезо- и микроуровня. Осуществлена привязка объектов к ГИС-пространству.

- Блок проектов – инвестиционных, программных, институциональных. Основное наполнение Блока проектов осуществлено на базе информации об инвестиционных проектах, представленных на цифровых платформах «Инвестиционные проекты» и «B2B Global», находящихся на разных стадиях реализации (планирование, предпроектные проработки, проектирование, подготовка к строительству, строительство, модернизация) в различных регионах России. Отраслевой и пространственный анализ проектов платформы позволяют оценить роль использования ресурсов в синергии инвестиций, выявить положительные и отрицательные их эффекты, сформировать наиболее эффек-

тивные цепочки взаимодействия проектов и направления инвестиционных импульсов.

– Блок по направлению работ на углеводородное сырьё (ресурсные кластеры, компании, участки недр, состояние геологоразведочных и добычных работ), Блок по отдельным видам твёрдых полезных ископаемых (ресурсные кластеры, компании, участки недр) и Блок по подземным водам (месторождения) были интегрированы в ГИС-проект и находятся в процессе преобразования для интеграции с ГИС-интерфейсом.

– Блок технологий, представляющий собой информацию о характеристиках технологических процессов нефтепереработки и переработки твердых полезных ископаемых.

4. В монографии приведены примеры исследований с использованием разработанного вычислительного комплекса, в котором с помощью когнитивного инструментария, динамической межотраслевой модели ДММ-КАМИН и аналитических блоков исследованы мультипликативные воздействия актуальных инвестиционных проектов, влияние цифровизации городской среды на развитие высокотехнологического бизнеса, и проведены первичные оценки инвестиционных проектов с использованием разнообразного инструментария.

Последующее расширение Базы знаний создает кумулятивный эффект и подразумевает многостороннее развитие:

– Разработка данных платформ, в том числе b2b, позволит осуществить формирование системы анализа цепочек добавленных стоимостей. С использованием сетевых, ГИС-моделей, данных о компаниях и месторождениях возможно создание системы, позволяющей получать более детальную оценку инвестиционных проектов с учетом ограничений спроса, логистических взаимосвязей, развития институциональных механизмов инвестирования, а также формирования инновационной среды для развития экономики.

– Развитие систем ГИС-моделирования с наработанными и новыми массивами данных позволит создать системы кластеризации и оценки издержек инвестиционных проектов.

– Построение системы непрерывного обновления данных из разных источников позволит увеличить надежность данных, в том числе и данных о компаниях.

– Анализ бухгалтерской отчетности компаний позволяет совершенствовать системы оценки эффективности компаний и их регионального социального эффекта.

– Развитие системы ГИС-моделирования позволяет осуществить создание индивидуального пользовательского пространства с возможностью самостоятельной визуализации ГИС-слоёв, их сохранение и редактирование.

– Проведенная обработка статистических данных (глава 3.3) позволяет существенно продвинуться в анализе и прогнозировании региональных и отраслевых взаимодействий путем применения и комбинации как известных методов, так и привлечения новых приемов анализа.

– Развитие аналитических информационных панелей позволит ускорить экспертный анализ данных.

– Комплексное развитие Базы знаний приведет к созданию дополнительного инструментария и знаний, способствующих повышению надёжности и качества социально-экономического планирования и прогнозирования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Арлазаров В.Л., Емельянов Н.Е. От баз данных к базам знаний (объекты, формы, содержание) // Труды Института системного анализа Российской академии наук. – 2006. – Т. 23. – С. 6–17.

Белан А.К., Шмат В.В. Анализ влияния ресурсных и нересурсных факторов на рост экономики Томской области с применением когнитивного подхода // Вестник Новосибирского государственного университета. Серия: Социально-экономические науки. – 2015. – Т. 15. – Вып. 1. – С. 78–93.

Варламов О.О. Эволюционные базы данных и знаний. Миварное информационное пространство // Известия Южного федерального университета. Технические науки. – 2007. – Т. 77. – № 2. – С. 77–81.

Гинис Л.А. Истоки современного когнитивного моделирования // Известия Южного федерального университета. Технические науки. – 2005. – Т. 50. – № 6. – С. 119–128.

Захарова Е.Н. О когнитивном моделировании устойчивого развития социально-экономических систем // Вестник Адыгейского государственного университета. Серия 1: Регионоведение: философия, история, социология, юриспруденция, политология, культурология. – 2007. – № 1. – С. 184–190.

Ивин А.А. и др. Философия: энциклопедический словарь. – 2004.

Костина Е.А., Костин А.В. Умный город как фактор развития высокотехнологичных компаний. – DOI: 10.15372/REG20230304 // Регион: экономика и социология. – 2023. – № 3 (119). – С. 84–110. RSCI, BAK.

Макарычев П.П., Гурьянов Л.В., Цукарев Э.В. Верификация и валидация данных в автоматизированной системе мониторинга и управления энергопотреблением // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2016. – № 4 (40). – С. 5–15.

Модели и методы прогнозирования: Азиатская Россия в экономике страны / под ред. А.О. Баранова, В.И. Суслова; Институт экономики и организации промышленного производства Сибирского отделения РАН. – Новосибирск: Изд-во ИЭОПП СО РАН, 2023. – 436 с. – URL: Электронный ресурс (Web) (дата обращения: 21.09.2023). – ISBN 978-5-89665-376-9; DOI: 10.36264/978-5-89665-376-9-2023-012-436.

Морозова М.Е., Шмат В.В. Среднесрочное прогнозирование российской экономики с использованием когнитивной модели // Проблемы прогнозирования. – 2017. – №. 3. – С. 19–25.

Новый импульс Азиатской России: источники и средства развития: в 2-х т. Т. 1 / под ред. В.А. Крюкова, Н.И. Суслова; Институт экономики и организации промышленного производства Сибирского отделения РАН. – Новосибирск: Изд-во ИЭОПП СО РАН, 2023. – 418 с. – ISBN 978-5-89665-375-2; DOI: 10.36264/978-5-89665-375-2-2023-011-418.

Новый импульс Азиатской России / под редакцией В.А. Крюкова, Н.И. Суслова; Сибирское отделение Российской академии наук, Институт экономики и организации промышленного производства. – Новосибирск: Изд-во СО РАН; Изд-во ИЭОПП СО РАН, 2022. – 571 с. – ISBN 978-5-6047824-9-1 (СО РАН), ISBN 978-5-89665-369-1 (ИЭОПП СО РАН); DOI: 10.53954/9785604782491.

Нонака И., Такеучи Х. Компания – создатель знания: зарождение и развитие инноваций в японских фирмах / И. Нонака, Х. Такеучи; Икуджиро Нонака, Хиротака Такеучи; [пер. с англ. А. Трактинского]. – Москва: ЗАО «Олимп-Бизнес», 2011. – 366 с. – ISBN 978-5-9693-0184-9. – EDN QURFIN.

Петров С.П. Технологические тенденции и направления развития черной металлургии Азиатской России // Черные металлы. – 2022. – № 8. – С. 76–82. DOI: 10.17580/chm.2022.08.13.

Петров С.П. Черная металлургия Азиатской России во втором и третьем десятилетиях XXI века. – Новосибирск: Изд-во ИЭОПП СО РАН, 2023. – 240 с. DOI 10.36264/978-5-89665-373-8-2023-009-240.

Полетаева Н.Г. Классификация систем машинного обучения // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. – 2020. – №. 1. – С. 5–22.

Робертс Ф.С. Дискретные математические модели с приложениями к социальным, биологическим и экологическим задачам. – М.: Наука, 1986.

Родионова М.Н. Моделирование баз знаний с использованием сетей Петри // Научный форум: Инновационная наука. – 2017. – С. 49–53.

Трофимова Л.А., Трофимов В.В. Управление знаниями: учебное пособие // СПб.: Изд-во СПбГУЭФ. – 2012. – С. 57.

Тузовский А.Ф. Разработка систем управления знаниями на основе единой онтологической базы знаний // Известия Томского политехнического университета. Инжиниринг георесурсов. – 2007. – Т. 310. – №. 2. – С. 182–185.

Тузовский А.Ф., Чириков С.В., Ямпольский В.З. Системы управления знаниями (методы и технологии). – Томск: Изд-во НТЛ, 2005. – 260 с.

Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных/ пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.

Черненко В.В., Пискорская С.Ю. Экспертные системы // Актуальные проблемы авиации и космонавтики. – 2012. – Т. 2. – №. 8. – С. 322–323.

Ягольницер М.А., Овсянникова М.А., Костин А.В. Синергия инвестиционных проектов: когнитивный подход // Мир экономики и управления. – 2022. – Т. 22. – №. 3. – С. 51–65.

Ahmed N. K. et al. An empirical comparison of machine learning models for time series forecasting // Econometric reviews. – 2010. – Т. 29. – No. 5–6. – С. 594–621.

Antony S., Batra D., Santhanam R. The use of a knowledge-based system in conceptual data modeling // Decision Support Systems. – 2005. – Т. 41. – No. 1. – С. 176–188.

Antunes F.N. et al. Accounting for Heteroscedasticity in Big Data // Proc. NSF Workshop on Big Data and Urban Informatics. – 2015. – C. 276.

Axelrod R. (ed.). Structure of decision: The cognitive maps of political elites. – Princeton university press, 2015.

Ben-Hur A., Elisseeff A., Guyon I. A stability based method for discovering structure in clustered data // Biocomputing 2002. – 2001. – C. 6–17.

Bishop C.M. Bayesian neural networks // Journal of the Brazilian Computer Society. – 1997. – T. 4. – C. 61–68.

Bishop C.M. Training with noise is equivalent to Tikhonov regularization // Neural computation. – 1995. – T. 7. – No. 1. – C. 108–116.

Breusch T.S., Pagan A.R. A simple test for heteroscedasticity and random coefficient variation // Econometrica: Journal of the econometric society. – 1979. – C. 1287–1294.

Burges C. et al. Learning to rank using gradient descent // Proceedings of the 22nd international conference on Machine learning. – 2005. – C. 89–96.

Felfernig A., Wotawa F. Intelligent engineering techniques for knowledge bases // AI Commun. – 2013. – T. 26. – No. 1. – C. 1–2.

Foray D., Lundvall B. The Knowledge-Based Economy: From the Economics of Knowledge. – 1996.

Goldfeld S.M., Quandt R.E. Some tests for homoscedasticity // Journal of the American statistical Association. – 1965. – T. 60. – No. 310. – C. 539–547.

Harenslak B.P., de Ruiter J. Data Pipelines with Apache Airflow. – Simon and Schuster, 2021.

Hart P.E., Stork D.G., Duda R.O. Pattern classification. – Hoboken: Wiley, 2000.

Hotelling H. Analysis of a complex of statistical variables into principal components // Journal of educational psychology. – 1933. – T. 24. – No. 6. – C. 417.

Hou X. et al. A semantic data-driven knowledge base construction method to assist designers in design inspiration based on traditional motifs // *Advanced Engineering Informatics*. – 2023. – T. 56. – C. 101987.

ISO/IEC/IEEE International Standard – Systems and software engineering – Vocabulary // *ISO/IEC/IEEE 24765:2017(E)*. – Vol. 2017 – No. IEEE – Pp.1–541. – 28 Aug. 2017, DOI: 10.1109/IEEESTD.2017.8016712.

Lundvall B.-A., Jonson B. The Learning Economy // *Journal of Industry Studies*. – 1994. – Vol. 1. – No. 2. – December. – Pp. 23–42.

Moody J., Darken C.J. Fast learning in networks of locally-tuned processing units // *Neural computation*. – 1989. – T. 1. – No. 2. – C. 281–294.

Pearson K. LIII. On lines and planes of closest fit to systems of points in space // *The London, Edinburgh, and Dublin philosophical magazine and journal of science*. – 1901. – T. 2. – No. 11. – C. 559–572.

Powell M.J.D. Radial basis functions for multivariable interpolation: A review // *Algorithms for Approximation* // J.C. Mason and M.G. Cox, ed. Clarendon Press. – Oxford, UK – 1987. – C. 143–167.

Samuel A.L. Some studies in machine learning using the game of checkers // *IBM J. Res. Dev.* – 1959. – T. 3. – No. 3. – P. 210–229.

Vassiliadis P. et al. A generic and customizable framework for the design of ETL scenarios // *Information Systems*. – 2005. – T. 30. – No. 7. – C. 492–525.

БЛАГОДАРНОСТИ

Автор выражает благодарность академику В.А. Крюкову, к.э.н. М.А. Ягольницеру за редактирование и помощь в написании раздела 5.1, д.э.н. Н.М. Ибрагимову за помощь в написании раздела 3.1.2, д.э.н. О.А. Баранову за помощь в написании раздела 3.1.1. и 5.4, Е.А. Костиной за помощь в написании раздела 5.2, С.П. Петрову за помощь в написании раздела 5.3, О.В. Тарасовой за помощь в написании раздела 5.5., Е.А. Федорову за помощь в написании раздела 1.1, Д.А. Родионовой за помощь в написании разделов 2.1.3 и 3.3, коллегам из ФИЦ ИВТ, с кем мы начинали проект, особенно А.А. Жирнову, д.э.н. Н.И. Суслову, к.э.н. Г.Д. Ковалевой, П.С. Браер и М.А. Овсянниковой за обсуждение монографии, ценные замечания и предложения по её улучшению.

План изданий ИЭОПП СО РАН, 2023 г.

Научное издание

Андрей Владимирович Костин

**БАЗА ЗНАНИЙ КАК ИНСТРУМЕНТАРИЙ
РЕШЕНИЯ ЗАДАЧ В ЭКОНОМИЧЕСКОЙ,
СОЦИАЛЬНОЙ И ПРОИЗВОДСТВЕННОЙ СФЕРАХ**

Оформление обложки
Компьютерная вёрстка

А.С. Кузнецова
С.А. Дучкова

Подписано к печати 05 февраля 2024 г. Формат бумаги 60×84¹/₁₆. Гарнитура «Таймс».
Объём п.л. 10. Уч.-изд.л. 9,7. Тираж 500 экз. Заказ № 8.

Издательство ИЭОПП СО РАН

Тел. (383) 330-17-95, e-mail: s.duchkova@icie.nsc.ru
Участок оперативной полиграфии ИЭОПП СО РАН,
630090, г. Новосибирск, проспект Академика Лаврентьева, 17.