МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение   
высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
(ФГБОУ ВО «ИГУ»)

|  |  |
| --- | --- |
| Институт математики  и информационных технологий | Кафедра теория вероятностей  и дискретной математики |

**ОТЧЕТ**

о прохождении преддипломной практики

10 мая 2021 года – 06 июня 2021 года

|  |  |
| --- | --- |
|  | Студента 4 курса группы 02441-ДБ  направления Математическое обеспечение и администрирование информационных систем  Шубина Дмитрия Александровича  К. т. н., доцент Парамонов В. В.  Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Иркутск – 2021

[Введение 3](#_Toc73945988)

[Глава 1 Теоретическая часть 6](#_Toc73945989)

[1.1 Методы по решению поставленной задачи 6](#_Toc73945990)

[1.1.1.1 Прогнозирование присутствия клещей с помощью картирования 7](#_Toc73945991)

[1.1.1.2 Сезонная синхронность: ключ к очагам клещевого энцефалита, выявленным по спутниковым данным 8](#_Toc73945992)

[1.1.1.3 Теоретические аспекты математического моделирования динамики развития популяции клещей 10](#_Toc73945993)

[1.1.2 Регрессионный анализ 11](#_Toc73945994)

[1.1.2.1 Оценка уравнения регрессии 13](#_Toc73945995)

[1.1.2.2 Матрица парных коэффициентов корреляции 14](#_Toc73945996)

[1.1.2.3 Коэффициент множественной корреляции 15](#_Toc73945997)

[1.1.2.4 Коэффициент детерминации 16](#_Toc73945998)

[1.1.2.5 Проверка общего качества уравнения множественной регрессии 16](#_Toc73945999)

[1.3 Методология выполнения практической части 17](#_Toc73946000)

[1.4 Аналоги, сравнение с другими похожими работами 18](#_Toc73946001)

[Глава 2 Практическая часть 21](#_Toc73946002)

[2.1 Какие подходы были использованы 21](#_Toc73946003)

[2.1.1 Структуризация данных 22](#_Toc73946004)

[2.1.2 Описательная аналитика 23](#_Toc73946005)

[2.1.3 Визуализация данных 25](#_Toc73946006)

[2.1.4 Предикативная аналитика 25](#_Toc73946007)

[2.2 Технологии используемые при работе 26](#_Toc73946008)

[2.2.1 Java и Apache Poi 26](#_Toc73946009)

[2.2.1.1 Чтение таблиц 26](#_Toc73946010)

[2.2.1.2 Разбор данных о состоянии погоды 28](#_Toc73946011)

[2.2.1.3 Запись и создание базы данных 29](#_Toc73946012)

[2.2.1.4 Моделирование поведения на основе полученных данных 29](#_Toc73946013)

[Заключение 31](#_Toc73946014)

[Список использованной литературы 32](#_Toc73946015)

[Приложения 33](#_Toc73946016)

# Введение

На Земле встречается более 50 тысяч видов клещей. Из них 713 видов иксодовых. Из них порядка 70 видов встречается в России. Все они вампиры, или «временные высокоспециализированные облигатные гематофаги»[[1]](https://lenta.ru/articles/2016/06/17/acari/). Все они являются переносчиками вирусов, риккетсий, бактерий, спирохет, трипаносом, филярий и пироплазмидов, способных вызвать у донора более 300 инфекционных и паразитарных заболеваний. Кроме того, в результате присасывания лесных и таежных иксодовых клещей часто развиваются смешанные инфекции, характерные более тяжелым течением заболевания.

Один из самых распространенных недугов, передающийся через укусы клещей, это боррелиоз, или болезнь Лайма. В половине случаев место укуса краснеет, пятно расширяется, достигая значительного размера в диаметре. В ряде случаев боррелиоз вначале протекает бессимптомно или маскируется под другие болезни, за что его называют невидимкой. На поздних стадиях заболевание приводит к поражению суставов, сердца и нервной системы. На домашних питомцев боррелиоз не распространяется.

Все клещевые инфекции лечатся на ранних стадиях, а вот единственный способ не заразиться боррелиозом — это не допускать присасывания клеща.

Больше всего в России сегодня страдает Сибирский Федеральный округ на него приходится почти 50% всем укусов в стране, несмотря на то что, что в СФО проживает всего лишь 12% от населения всей страны.

На середину апреля 2021 года в Российской Федерации уже насчитывается почти 2 500 случаев заболевания клещевым энцефалитом и 7 500 случаев боррелиоза. Из этих случаев около 90 случаев заболевания клещевым энцефалитом приходится на Иркутскую области, причем 76 из них приходится на прошедшею неделю. Согласно сайту irksib.ru[[2]](https://irksib.ru/allnews/12-social/23845-76-ukusov-kleshchej-zaregistrirovali-v-irkutskoj-oblasti-za-nedelyu#:~:text=%D0%90%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C%20%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9%20%D0%B2%D0%BE%D0%B7%D1%80%D0%B0%D1%81%D1%82%D0%B0%D0%B5%D1%82%20%D0%B2%20%D0%98%D1%80%D0%BA%D1%83%D1%82%D1%81%D0%BA%D0%BE%D0%B9,%D0%A8%D0%B5%D0%BB%D0%B5%D1%85%D0%BE%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%2C%20%D0%91%D0%B0%D1%8F%D0%BD%D0%B4%D0%B0%D0%B5%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%20%D0%B8%20%D0%9A%D0%B0%D1%87%D1%83%D0%B3%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%20%D1%80%D0%B0%D0%B9%D0%BE%D0%BD%D0%BE%D0%B2) , из 75 представленных для экспертизы клещей, 7 оказались заражены боррелиозом и 4 другими инфекциями.

Лечение данных заболеваний не является дешёвым, поэтому так актуально информирование население о том что клещи есть и они действительно представляют опасность нашему организму, профилактика укусов может стоит совсем небольших усилий и символической суммы, для прививания от укуса, но к сожалению они не гарантируют вашу неприкосновенность, но заметно снижают шанс заразить инфекцией, которую возможно этот клещ носил долгое время с собой. Также представляется актуальным создание информационного ресурса, демонстрирующего активности иксодовых клещей на территории региона и прогнозирование активности на будущие периоды на основе ретроспективных данных.

В связи с большой опасностью, которую несут клещи, необходимо всегда оценивать текущую статистику по случаям заражения, а также анализировать опыт прошлых лет, поэтому все обращения населения по поводу укусов фиксируется в медицинских учреждениях.

2021 год характеризуется ранним повышением среднесуточных температур, отчего сезон активности клещей начался примерно на месяц раньше планируемого. Поэтому представляет интерес понимание того, в каких местах следует находятся меньше, чтобы снизить, а в лучшем случае и вовсе исключить возможность заражения любой из переносимых клещами инфекций. Таким образом, создание математических и информационных моделей, позволяющих по ретроспективным данным проанализировать активность иксодовых клещей на территории Иркутской области является чрезвычайно актуальной задачей.

Целью выпускной квалификационной работы является разработка алгоритмического и программного обеспечения для моделирования активности иксодовых клещей.

Задачи поставленные для достижения цели дипломной работы:

* проанализировать данные об укусах клещами населения Иркутской области,
* провести корреляционный анализ,
* понять какие параметры больше всего влияют на повышения шанса укуса.
* А также с предоставленными данными предоставить карту, с указанием мест наиболее опасных для посещения, так как в них велик шанс укуса.

Объектом исследования являются статистические данные активности иксодовых клещей на территории Иркутской области.

Предмет исследования информационная модель активности иксодовых клещей.

# 

# Глава 1 Теоретическая часть

В документе опубликованном на сайте Роспотребнадзора – “Перечень административных территорий субъектов Российской Федерации, эндемичных по клещевому вирусному энцефалиту в 2020 году”[[7]](https://www.rospotrebnadzor.ru/documents/details.php?ELEMENT_ID=16919), в Иркутской области 30 из 36 административных регионов являются эндемичными, то есть территориями шанс в которых обнаружить клеща выше чем в других регионах. Ежегодно около 10 тысяч человек становятся жертвами этих паукообразных, большая часть которых конечно оказывается безвредными, то есть в момент укуса или обнаружения не переносили клещевые инфекции в своей крови, однако часть особей всё же являются переносчиками, самыми распространенными инфекциями являются клещевой энцефалит и боррелиоз. При своевременном обращении в медучреждение указанные выше заболевания не наносят никаких повреждений вашему здоровью, однако в случае нелечения, они могут развиваться до более тяжелой степени и привести вплоть до инвалидности.

Поэтому главной задачей поставленной в выпускной квалификационной работе являлась реализация алгоритмического и программного обеспечения с помощью которого можно будет построить модель возможного распределения клещей по Иркутской области.

## 1.1 Методы по решению поставленной задачи

Для правильного составления структуры практической части выпускной квалификационной работы были предприняты следующие методы

1.1.1 Изучение литературных источников

В ходе работы были исследованы литературные многие литературные источники для получения теоретических знаний на основе которых были выявлены некоторые закономерности и факторы на которые стоит обратить внимание при построении модели.

Несмотря что некоторые виды иксодовых клещей могут встречаться даже в субарктических и арктических регионах, в основном клещи выбирают места с комфортной температурной средой, поэтому в эндемичных субъектах Российской Федерации период активности иксодовых приходится на март-октябрь. В марте клещи просыпаются от спячки, становятся агрессивными и начинают чаще кусать. В мае-июне-июле наблюдается миграция людей на природу, так как этот период средняя температура воздуха поднимается до 15-20 градусов тепла и начинаются отпуска и каникулы. Поэтому в этот период наблюдается наибольшее количество обращений населения в медучреждения по поводу обнаружения клещей. В сентябре-октябре иксодовые клещи снова становятся агрессивными так как необходимо запастись энергий на длительный период спячки, однако количество зарегистрированных случаев укусов клещей падает так как, миграция населения направлена в учебные и рабочие организации[[8]](https://moscow-777.ru/prochee/kogda-snizhaetsya-aktivnost-kleshhej.html).

#### 1.1.1.1 Прогнозирование присутствия клещей с помощью картирования

Оригинальная статья написана Центром инфекционных заболеваний, RIVM (Национальный Институт общественного здравоохранения и окружающей среды).

В данной статье рассматривается моделирование распространения клещей путем сбора информации о содержании клещей на единице территории а также спутниковых снимках исследуемой территории. Плотность клещей на единицу расстояния измерялась путем волочения одеяла по территории и позже подсчёт количества собранных образцов. Спутниковые снимки использовались, так как предполагается что растительность, которая в свою очередь является естественным убежищем для данных паукообразных, поэтому была необходимость оценить взаимосвязь распространения клещей в зависимости от растительности, преобладающей на данной территории. Исследования производились в период активности – с апреля по сентябрь в течение нескольких лет, чтобы из полученных данных получить наиболее репрезентативную выборку.

Так как построение модели планировалось для всей территории Нидерландов, а данные были собраны лишь с 0.02% её территории, полученные данные нужно было экстраполировать на оставшиеся территорию и полученным данным получилось, что 20 698 квадратных километров из 30 001 пригодны для проживания клещей. Таким образом, по оценкам, 54% поверхности суши соответствует условиям для поддержания жизненного цикла клещей.

#### 1.1.1.2 Сезонная синхронность: ключ к очагам клещевого энцефалита, выявленным по спутниковым данным

Оригинальная статья написана S. E. RANDOLPH, R. M. GREEN, M. F. PEACEY and D. J. ROGERS, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

В статье поднимается вопрос о зависимости сезонности и случаев клещевого энцефалита. Данные о наличие клещей были собраны путём волочения одеяла и подсчёта собранных особой, а также личинок.

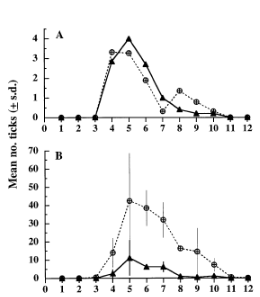


рис. 1 Сезонные колебания численности незрелых стадий Ixodes ricinus в Словакии. (А) Требисов (восточный), (B) Дунайская степь (западный)

Как можно заметить наибольшее количество личинок клещей наблюдается в период с апреля по июнь. В апреле клещи просыпаются из спячки и начинает активно плодиться, увеличивая свою популяцию. К маю-июню количество клещей достигает своего апогея и активность кормления снижается. Эти данные можно хорошо коррелировать с данными о средней температуры поверхности и скоростью осеннего охлаждения.

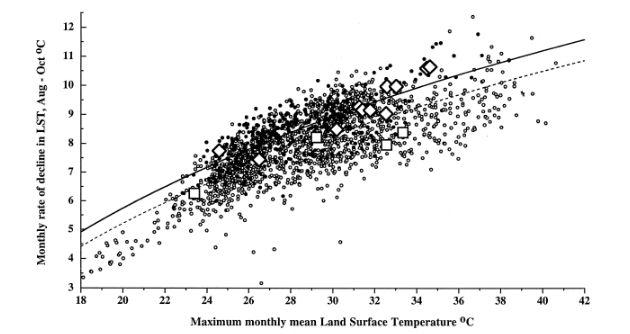


рис 2 Взаимосвязь между максимальной среднемесячной температурой поверхности Суши (°C) и месячной скоростью осеннего охлаждения с августа по октябрь

#### 1.1.1.3 Теоретические аспекты математического моделирования динамики развития популяции клещей

Оригинальная статья написана Т.М. Ковалёвой

Автор рассматривает возможность моделирования динамики развития популяции клещей в Алтайском крае.

В процессе исследования особенностей жизненного цикла клеща выявлены четыре фазы развития: от появления кладки яиц до развития взрослой особи – имаго. После изучения особенностей развития проведена формализация каждой фазы. В итоге получилась следующая схема развития.

I фаза развития – рождение клеща. Начинается после того, как самка клеща откладывает яйца в верхнюю почвенную прослойку. В зависимости от природных условий через несколько недель, иногда через семь месяцев, появляются шестиногие личинки длиной около 1 мм.

II фаза развития – появление личинок. Начинается, когда из яиц появляются личинки. За это время под воздействием совокупности факторов погибают некоторое количество яиц.

III фаза развития – перерождение личинок в нимф. Следуя жизненному циклу, нимфы ищут хозяина-донора.

IV фаза развития – появление имаго. После третьей фазы развития в природе появляются взрослые клещи, способные к укусу человека

В число факторов, действующих на кладку яиц в первой фазе развития, входят в основном только абиотические и доля случайной гибели яиц от внешних воздействий.

В число факторов, влияющих на жизнеспособность личинок, нимф и взрослых особей, входит гораздо большее число разнообразных факторов. При трансмиссивном пути заражения в очаге клещевых инфекций происходит взаимодействие факторов, которые можно структурно разделить на абиотические, биотические, географические, антропогенные и социально-экономические.

К абиотическим факторам относятся теплообеспеченность, увлажнение, сумма среднемесячных температур осенью предыдущего года (июль-август), сумма среднемесячных температур (ноябрь–февраль); сумма осадков в июне.

К биотическим факторам относятся: соблюдение жизненного цикла развития клеща, численность и активность потомства, обеспечение наличия позвоночного хозяина-донора на каждом этапе развития, плотность распространения и инфицированность клещей, численность переносчиков-распространителей, соблюдение условий переноса и распространения инфекции и т. д.

В числе географических факторов рассматриваются особенности ландшафта исследуемой территории и возможности проживания на ней определенного вида клещей.

К числу антропогенных факторов отнесем влияние человека на изменение ландшафтной структуры территории. Это проявляется в распаханности земель, вырубке лесов, наличии сенокосных угодий, пастбищ. Также необходимо учесть выгоревшие территории; наличие населенных пунктов, площадь земель, которые они занимают.

В число социально-экономических факторов входит интенсивность связей населения с природными очагами, в туристических зонах с необработанными от инфекций территориями, сельских местностях, богатых ресурсами, природных зонах интенсивность связей населения многократно увеличивается, отсюда идет рост риска заражения клещевыми инфекциями.

### 1.1.2 Регрессионный анализ

Регрессионный анализ – это набор статистических методов оценки отношений между переменными. Его можно использовать для оценки степени взаимосвязи между переменными и для моделирования будущей зависимости.» [9]. То есть, другими словами, регрессионный анализ позволяет понять, как будет себя вести изучаемая величина при изменении факторов, влияющих на нее.

В регрессионном анализе выделяют 2 вида регрессий:

* Линейная регрессия – метод, при котором выявляется зависимости одной (объясняемой, зависимой) переменной Y от другой или нескольких других переменных (факторов, независимых переменных) X с линейной функцией зависимости. Данный метод позволяет предсказывать значения зависимой переменной Y по значениям независимой переменной X. Линейная регрессия имеет вид: Y = BX, где B – это набор коэффициентов для фактора(ов) Х и так же коэффициент b0 – значение, которое примет Y в случае, если все факторы Х будут равны нулю.
* Нелинейная регрессия – аналогично линейной регрессии, но зависимость нельзя представить в виде линейной функции. В нелинейных регрессиях выделяют:   
  степенную – выражается по формуле: Y = AX^B,   
  показательную – выражается по формуле: Y = AB^X и экспоненциальную, выражаемая по формуле: Y = e^(AX + B) .

Рассмотрим алгоритм выполнения регрессионного анализа для линейной модели включающая в себя m факторов.

Алгоритм построения регрессионного анализа.

1. Оценка уравнения регрессии
2. Матрица парных коэффициентов корреляции
3. Коэффициент множественной корреляции
4. Коэффициент детерминации
5. Проверка общего качества уравнения множественной регрессии

Разберем каждый пункт поподробнее.

#### 1.1.2.1 Оценка уравнения регрессии

Для составления уравнения множественной линейной регрессии, согласно методу МНК – метод наименьших квадратов. Чтобы получить коэффициенты B, в общем уравнении регрессии Y = BX, необходимо решить систему линейных уравнений s = (X^T \* X) \* (X^T \* Y), где Х^T – транспонированная матрица Х.

Чтобы в ходе решения получить коэффициент b0, к исходной матрице Х, составленная из значений факторов Х1, Х2, ... Хm, нужно добавить первым столбцом единичный. Наша матрица Х будет иметь следующий вид:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Х11 | Х12 | ... | Х1m |
| 1 | Х21 | Х22 | ... | Х2m |
| 1 | ... | ... | ... | ... |
| 1 | Хn1 | Xn2 | ... | Xnm |

Далее необходимо транспонировать данную матрицу, получим:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | ... | 1 |
| X11 | X21 | ... | Xn1 |
| X12 | X22 | ... | Xm2 |
| ... | ... | ... | ... |
| Xm4 | Xm4 | ... | Xmn |

Проведя оставшиеся арифметические операции, мы получим набор значений Y(X), где первым будет является коэффициент b0, оставшиеся соответствуют Хi.

|  |
| --- |
| b0 |
| b1 |
| b2 |
| ... |
| bm |

#### 1.1.2.2 Матрица парных коэффициентов корреляции

Коэффициент корреляции говорит о том, насколько связаны между собой изучаемая величина или фактор с другим фактором. Матрица парных коэффициентов представляет собой матрицу связи всех факторов и изучаемой величины между собой, соответственно её размерность (n+1) x(n+1), где n – число факторов участвующих в исследовании.

Коэффициент рассчитывается по формуле:

= ,

После вычислений мы получим матрицу:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | rx1x2 | ... | rx1xn |
| rx2x1 | 1 | ... | rx2xn |
| ... | ... | ... | ... |
| rxnx1 | rxnx2 | .... | 1 |

#### 1.1.2.3 Коэффициент множественной корреляции

Тесноту совместного влияния факторов на результат оценивает индекс множественной корреляции.

В отличии от парного коэффициента корреляции, который может принимать отрицательные значения, он принимает значения от 0 до 1.

Поэтому R не может быть использован для интерпретации направления связи. Чем плотнее фактические значения yi располагаются относительно линии регрессии, тем меньше остаточная дисперсия и, следовательно, больше величина Ry(x1,…,xm).

Таким образом, при значении R близком к 1, уравнение регрессии лучше описывает фактические данные и факторы сильнее влияют на результат. При значении R близком к 0 уравнение регрессии плохо описывает фактические данные и факторы оказывают слабое воздействие на результат.

Существует несколько формул по которым можно рассчитать данный коэффициент, мы будем использовать вычисление через матрицы парных коэффициентов корреляции и его минора.

– формула для вычисления.

Мы используем уже ранее рассчитанную матрицу коэффициентов парной корреляции

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | rx1x2 | ... | rx1xn |  |  |  |  |
| rx2x1 | 1 | ... | rx2xn |  | 1 | ... | rx2xn |
| ... | ... | ... | ... |  | ... | .. | ... |
| rxnx1 | rxnx2 | ... | 1 |  | rxnx2 | ... | 1 |

Далее необходимо найти определители эти матриц и вставить в расчетную формулу для нахождения коэффициента множественной корреляции.

#### 1.1.2.4 Коэффициент детерминации

Коэффициент множественной корреляции показывает связь между изучаемой величиной Y и факторами Хi . Однако наибольшую точность связи описывает скорректированный коэффициент детерминации.

, где R – коэффициент множественной корреляции, n – количество наблюдений(экспериментов), m – количество факторов, участвующих в исследуемой модели.

Число v = n - m - 1 называется числом степеней свободы. Считается, что при оценивании множественной линейной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений, по крайней мере, в 3 раза превосходило число оцениваемых параметров.

Интерпретация коэффициента детерминации аналогично значению множественной корреляции, чем ближе к 1, тем изменение факторов сильнее объясняют поведение изучаемой величины Y, чем ближе к 0, величина Y менее зависит от исследуемых факторов.

Добавление в модель новых объясняющих переменных осуществляется до тех пор, пока растет скорректированный коэффициент детерминации. Таким образом можно оценить добавление или удаление тех или иных факторов влияет положительно или отрицательно на качество модели.

#### 1.1.2.5 Проверка общего качества уравнения множественной регрессии

Проверка выполняется на основе t-статистики, для этого необходимо значение критерия Стьюдента с параметрами равными числу степеней свободы и 0.025, тем самым получим значение .

По формуле: получим значение которое останется сравнить с , в случае если окажется меньше, чем , то статистическая надежность коэффициента регрессии не подтверждается. Значение находятся из уравнения регрессии в естественном масштабе, а дисперсия параметра определяется корнем квадратным из значения ковариационной матрицы.

Для уточнения важности коэффициента следует составить доверительный интервал для выбранного и если полученный интервал будет содержать в себе 0, то коэффициент признается незначимым.

Определим доверительные интервалы с надежностью 95% по формуле:

.

## 1.3 Методология выполнения практической части

Методология – учение об организации процесса изучения или исследования какого-либо объекта.

При следовании заранее построенному структурированному плану гораздо проще следовать, тогда шанс упустить какие-то важные моменты стремится к нулю.

Многие современные методологии нашли своё применение в области Data Mining, однако подходы использующие данные методологии допустимы к использованию в любой исследовательской области.

В ходе выпускной квалификационной работы была выбрана методология “SEMMA” разработанная в среде SAS Data Mining Solution, название методологии является аббревиатурой от: Sample, Explore, Modify, Model, Assess.

Каждое слово является определенной частью на пути изучения поставленной задачи.

* Sample – Выборка данных. На данном этапе подготавливаются наборы данных для последующего изучения, чем больший объём данный получится собрать, тем более репрезентативной модель может получиться, однако в тоже время данные не должны быть слишком перегружены.
* Explore – Изучение. Фаза, в которой выявляются закономерности, а также возможные отклонения
* Modify – Изменение. Этап, во время которого, данные подвергаются изменению, форматированию, подготовкой к построению модели
* Model – Моделирование. Применяются различные методы для построения предиктивной модели
* Assess – Оценка. Необходимо оценить качество составленной модели, понять на сколько она надежна и на сколько точно способна предсказать значения

## 1.4 Аналоги, сравнение с другими похожими работами

Современные технологии информационно-аналитической оценки активности и прогноза распространения иксодовых клещей на примере города Иркутска

При поддержке руководителей из института динамики систем и теории управления СО РАН, а также научного центра проблем здоровья семьи и репродукции человека СО РАМН.

Основной идеей которой вдохновляла авторов статьи это создание единой ИАС, которая объединяет все необходимые данные и средства их анализа.

Система должна базироваться на информационных и геоинформационных технологиях и обеспечивать открытый доступ к просмотру данных через сеть Интернет и включать следующие базовые компоненты: – единое хранилище данных и средства загрузки/выгрузки данных; – методы пространственного и временного анализа; – отображение данных и результатов анализа. Кроме того, эта технология должна позволять включать новые методы анализа данных.

Для реализации данного проекта необходимо придерживаться некоторых аспектов таких как:

Единое хранилище данных. В различных организациях и ведомствах имеются БД, содержащие информацию о заболеваниях, переносчиками которых являются иксодовые клещи, противоклещевых обработках территорий города, метеорологических условиях и пр. Все БД разрознены и не связаны между собой, что затрудняет их использование для определения опасности территорий и прогноза активности иксодовых клещей в зависимости от различных факторов. Таким образом, в настоящее время возникла необходимость создания единого хранилища данных, позволяющего аккумулировать всю тематическую информацию, полученную из разнородных источников, с последующим предоставлением доступа к ней с помощью WEB-технологий всех заинтересованных лиц.

Методы пространственного и временного анализа. Геокодирование данных о местах укусов иксодовыми клещами в совокупности с информацией о дате укусов станет основой динамических тематических карт. Сопоставление полученной информации с местностями и датами противоклещевых обработок обеспечит наглядное представление о распространении иксодовых клещей на территории города Иркутска и его окрестностей в разные периоды теплого времени года, а также о наличии клещей и степени опасности этих территорий в отношении риска заражения клещевыми инфекциями.

Отображение данных и результатов анализа. Для визуализации пространственных данных нами разработан оригинальный сервис, называемый SMDServer, который реализует поддержку стандарта OGC Web Map Server (WMS). Поддержка WMS позволяет использовать любой клиент визуализации пространственных данных, например, широко распространенный web-клиент OpenLayers или настольная ГИС Map Info. Основным достоинством SMDServer яв ляется возможность работы с векторными данными большого объема на высокой скорости.

Таким образом, оригинальность предлагаемого сервиса заключается в комплексном подходе к формированию единой БД, развитию технологий удаленного доступа к информации в БД и геоинформационных технологий. Реализация идеи связана с развитием OLAP-, WEB- и геоинформационных технологий. Полученные прогнозы и модели станут основой обеспечения экологической безопасности населения и принятия компетентных управленческих решений по профилактике инфекционных заболеваний [10].

# Глава 2 Практическая часть

Целью практической работы являлось реализовать поставленные ранее задачи, такие как:

Модель поведения клещей в зависимости от погодных условиев и прочих факторов. Понимание того в каких ситуациях активность клещей достигает своего апогея является одной из важнейших в процессе изучения данной проблемы, так как понимание того какие места окажутся наиболее предпочтительными для появления клещей напрямую могут повлиять на количество последующих случаев обращения в медучреждения. Цель моделирования заключается в том, чтобы постараться снизить это количество до минимального возможного значения. (тут как-то водянисто получилось)

Карта. Человек воспринимает до 80% информации через зрение, поэтому наиболее удобный способ представления информации является визуальным. Была поставлена задача составления карты случаев укусов клещей на территории Иркутской области, для выявления мест, в которых шанс встретить клеща наиболее вероятен, таким образом при информировании населения будет возможность снизить количество случаев укуса клещей, что впоследствии приведет к меньшим количествам обращения людей с подозрением на заболевания, которые клещи могут переносить в себе и распространять при укусе.

## 2.1 Какие подходы были использованы

Сегодня в двадцать первом веке, в мире, который полностью пропитан информационными технологиями, где основная валюта это бит информации, мы ежедневно обрабатываем огромные массивы данных. И в 2008 году Клиффордом Линчем, редактором журнала Nature было предложен термин “Big Data” что дословно в переводе с английского и означает “большие данные”.

Big Data – это структурированные или неструктурированные массивы данных большого объёма. Их обрабатывают при помощи специальных автоматизированных инструментов либо по заранее подготовленным алгоритмам, чтобы использовать для статистики, анализа, прогноза и принятия решений в той или иной области.

### 2.1.1 Структуризация данных

Структуризацию перед началом обработки данных можно назвать “нулевым” методом, так как эффективность работы с хаотичными данным может в десятки раз снижать эффективность работы. Поэтому перед началом работы с большим набором данных их следует структурировать: взять нужное и убрать всё ненужное.

Именно этот метод использовался при обработки электронной таблицы, содержащей в себе данные о погодных условиях, так как изначально файл содержал в себе излишнее количество информации, гораздо большее чем необходимо для последующего анализа.

Исходный набор данных содержит в себе информацию о погодных условиях с 2008 года до 2013 года. Первым шагов было выделение из этого массива только данные в промежутке с 2010 года по 2012.

Для каждой даты предусмотрено большое количество параметров, в нашем случае не все они являются нужными, так как влияние на модель является минимальным, а потому можно исключить.

На рис.3 можно видеть исходную структуру данного файла, из всего этого нам необходимо взять:

* дату
* температура
* влажность

Далее список данных выявленный выше берётся только для тех дат, которые указаны в файлах со статистическими данными о случаях укуса клещами населения на территории Иркутской Области.

После проведенных необходимых манипуляций все необходимые данные получены и структурированы, теперь можно переходить к следующим методам, которые непосредственно работают с данными.

### 2.1.2 Описательная аналитика

Данный вид аналитики самый распространенный в решении вопроса с Big Data. Она отвечает на вопрос «Что произошло?», анализирует данные, поступающие в реальном времени, и исторические данные. Главная цель – выяснить причины и закономерности успехов или неудач в той или иной сфере, чтобы использовать эти данные для наиболее эффективных моделей.

Описательная или дескриптивная аналитика – это область статистики, методы которой ориентированы на сбор, систематизацию и обобщение «сырых» данных из различных источников с целью обнаружения в них зависимостей и закономерностей, а также последующей интерпретации происходящего.

Как правило, описательная аналитика опирается на данные уже случившихся событий, предоставляя возможность понять смысл, который необходим для понимания происходящих процессов. Именно поэтому её часто называют «аналитикой случившегося», поскольку она позволяет получить более глубокий взгляд на данные, чтобы объяснить причины тех или иных событий, поведения процессов и т.д[6].

В нашем случае примером задачи описательной аналитики является разделение данных на периоды времени, внутри которых делаются выводы о дальнейшем поведении и активности клещей на территории Иркутской области.

Аналитические модели «захватывают» зависимости и закономерности в данных и экстраполируют их в будущее.

Описательную аналитику можно рассматривать как этап обработки данных для моделирования дальнейшего поведения, однако результаты её работы также имеют определенную ценность. Например, структурированные данные в описательной аналитике позволяют перейти к этапу визуализации и наглядно отобразить те зависимости которые были обнаружены в ходе изучения предоставленных данных.

Для примера возьмем уже структурированные данные по случаям укусов на территории Иркутской Области за 2010 год попробуем выяснить какие месяцы были наиболее благоприятными для деятельности клещей. Проведя сортировку данных по месяцам мы получили: что на июнь приходится 3 422 из 7 883 случаев обращения жителей области в медучреждения (см рис. 6), что составляет 43% от общего числа и делает месяц июнь наиболее активным для деятельности клещей, однако почему так происходит?

Рассмотрев распределение обращений по дням в июне 2010 года, можно заметить график имеет некоторую цикличность, расстояние между 2 соседними вершинами составляет 6–7 дней. Дни с наибольшим количеством обращений получились: 1, 7, 15, 21, 28. Найдя календарь за 2010 год, можно выяснить что эти даты приходятся на понедельники и вторники. Что является странным ведь это рабочие дни, вероятнее всего люди обращались после выходных, так как возвращаясь поздно после отдыха или будучи в ненадлежащем состоянии не могли обратиться в медучреждение в день происшествия. Проверим теорию на достоверность сравнив с графиком, в которой данные отсортированы по предполагаемым датам укуса клеща или его обнаружения. Видно, что вершины графика содержат по две точки и использовав тот же календарь, мы видим, что вершины приходятся на субботу и воскресенье.

Отсюда следует что предположения о том, что обращения происходят через несколько дней после случая укуса, но чаще всего в первый же день после выходных подтвердилась.

Таким образом, мы получили какой месяц и в какие дни наиболее велик шанс быть укушенным клещом, для более точного анализа нужно использовать также географические данные, тем самым мы бы получили время и место, где лучше не стоит находится, если не хотите поближе познакомиться с клещами.

### 2.1.3 Визуализация данных

Нашим главным источником информации является наше зрение. С его помощью мы получаем до 80% информации, поэтому то, что мы видим мы понимаем лучше чем, то что слышим или чувствуем тактильно. Визуализации занимает очень важное место в аналитике ведь с его помощью голые цифры и десятки гигабайт данных можно представить в виде пары информативных графиков. И главное что и “сырой”, и что вид через график отображает одни и те же данные, но данные в графиках человек лучше воспринимает и их проще анализировать на выявление зависимостей или обращения внимания на самые важные моменты, компьютеры в свою очередь лучше работают с сырыми данными, для них тысячи строк таблицы это источник данных, а для человека набор букв и цифр.

В нашем случае визуализация используется для представления информации о распространении на территории Иркутской области клещей, чтобы эту информацию можно было донести до более широкого пользователя и наглядно увидеть в какие места следует ограничить количество посещений и в какие периоды.

### 2.1.4 Предикативная аналитика

Прогнозная или предикативная – следующий уровень аналитики. Так как построить предикативную модель можно только обработав ранее полученные данные и выбрав для нее только необходимое. Поэтому данный этап в обработке данных выполняется после структуризации и описательной аналитике. Эффективность полученной модели зависит от того насколько правильные, то есть сильнее других влияющих на поведение модели, переменные были выбраны для ее построения.

## 2.2 Технологии используемые при работе

### 2.2.1 Java и Apache Poi

На сегодняшний день язык программирования Java занимает третье место по доле на рынке с результатом в 11.87 %[5]. Поэтому разработка приложений на Java всё ещё актуальна хоть и постепенно падает. Язык был выбран по большей части, так как имеется гораздо больший опыт работы с ним чем другими.

Для работы с электронными таблицами была выбрана библиотека Apache Poi, так как является наиболее часто используемой и имеющая достаточно читаемую документацию, а также в сети интернета содержится значительное количество статей посвященных Poi, так что даже такому новичку как мне, удалось разобраться с основами и вследствие чего научился читать/записывать и создавать электронные таблицы программно.

#### 2.2.1.1 Чтение таблиц

При чтении Excel–файла следует учитывать его расширение, так как будут использоваться разные классы.

В ходе работы мне было предоставлено четыре файла, из которых необходимо взять и обработать нужную информацию. Однако первые два файла, были реализованы в версии Excel 97–2003, в которых файл создаётся по умолчанию с расширением \*.xls, двое других выполнены на базе более новой версии, использующая расширение \*.xlsx, применяющееся по умолчанию ко всем электронным таблицам созданных в Excel 2007 и более поздних версиях.

Поэтому алгоритм для чтения данных состоит всего из двух пунктов:  
 1) Определить расширение

2Используя нужные классы, прочитать нужные наборы данных

Определим расширение через деление на подстроки по символу “точка” и так как в названии могут быть символы “точки” то в строковую переменную extension отправится только последняя подстрока, которая и будет в себе содержать искомое расширение. Далее запустим **excelParser** класса **Parser**, c параметрами: полное название файла и его расширение.

public static void readExcelFileByName(String fileName) throws IOException {

String[] splittedString = fileName.split("\\.");

String extension = splittedString[splittedString.length - 1];

Parser.*excelParser*(fileName, extension);

}

На (Листинге 1) представлена работа по чтению самого файла на примере классов HSSF\*, которые используются для чтения файлов с расширением \*.xls, для распишерения \*.xlsx используется аналогичный алгоритм, только в основе используется набор классов XSSF\*. В основе этого алгоритма лежит перебор электронной таблицы по строкам, а внутри каждой строки читается значение каждой ячейки используя метод getDateCellValue() или getStringCellValue() для соответствующего типа данных.

В случае их отсутствия ячейка считается пустой и будет содержать в себе либо пустую строку либо “нулевую дату” (исправить потом на создание просто пустой ячейки).

Три электронных таблицы содержащие в себе информацию по обращениям населения на случаи укусов клещей, имеют общую структуру, что позволило создать классовую модель, и структурировать данные, позже имея возможность прочитать любые данные из неё,, использую переменные объявленные в классе, содержащие в себе соответствующее значения из Excel-файла.

После чтения таблицы необходимо закрыть IO(Input-Output) поток, для экономии ресурсов java-машины.

#### 2.2.1.2 Разбор данных о состоянии погоды

В файле “данные\_по\_погоде\_2008\_2013” содержится информация о погодных условиях с 2008 года по 2013, из этого файла были взяты данные за 2010-2012 год, так как имеется данные по случаям обращения населения только за эти года.

Для построения модели наиболее приближенной к реальному распространению клещей, необходимо выяснить какие параметры влияют на их поведение сильнее остальных.

Несмотря на повсеместность клещей, те в свою очередь встречаются даже в Арктике и Антарктиде, один из наиболее явных параметров, сильно влияющих на их распространение в широтах, на которых расположилась Иркутская область является температура. Проведя небольшой анализ на данные полученные из статистических данных можно увидеть, что период активности клещей приходится на на промежуток от марта до октября, что соответствует весьма благоприятной температуре, а пиковое значение обращений населения приходится на июнь, которую в свою очередь у многих является месяцем отдыха, каникул и отпусков, а также и температура воздуха в среднем составляющая около 20 градусов являются благоприятными для проведения свободного времени на свежем воздухе, чем насекомые охотно пользуются.

К уже ранее полученным данным, содержащим информацию о времени и месту укуса, а также дополнительной информации клеще, с которым не посчастливилось столкнуться человеку, обратившемуся в медучреждение. Необходимо добавить среднее значение температуры, так как точное время укуса не указывается. Этим и занимается следующий алгоритм разделенный на несколько частей:  
 1) Рассчитать среднюю температура воздуха за конкретный день

2) Составить классовую модель содержащую в себе такие поля как: дата, температура, и прочие данные \*не только температура\*

#### 2.2.1.3 Запись и создание базы данных

Создаётся база данных PostgreSQL, первым делом необходимо открыть подключение с сервером Postgres. После создания таблицы в цикле выполняется sql запрос INSERT INTO (список колонок) название созданной таблицы. Так же в строку передаются параметрами сами значения, которые и будут записаны в базу данных.

#### 2.2.1.4 Моделирование поведения на основе полученных данных

На основе данных за 3 года сбора статистических данных об укусах клещей на территории Иркутской области в период с 2010 по 2012 год, была составлена модель показывающая предполагаемое место и количество обращений.

Описание алгоритма:

1) Время.

Берутся данные из файла, полученного путем объединения нескольких файлов, для большего комфорта. В котором содержится вся необходимая информация, такая как количество обращений, дата обращения человека в медицинское учреждение, а также предполагаемая дата совершения укуса или обнаружения клеща в непосредственной близости с человеком. Из этих данных выбирается наиболее опасный месяц, тот в котором произошло самое большое количество обращений за 3 года наблюдений. Используя результаты полученные на этапе исследования данных, выясняется, что наиболее опасными днями недели оказались выходные это суббота и воскресенье, что в целом неудивительно так как многие желают провести день на открытом воздухе либо в дали от городской суеты в местном парке или лесу. Также в июне активно начинают работу детские лагеря, хоть и на таких территориях активно проводится дезинфекция от опасных насекомых, случаи заражения всё же возможны. Поэтому несмотря на то что в выходные вероятность обнаружить клеща наибольшая опасность быть укушенным остаётся на протяжении всего месяца.

# 

# Заключение

В ходе проделанной квалификационной выпускной работы, был выполнен анализ литературных источников, из которых удалось подчерпнуть множество полезной информации, такой как на что следует обратить внимание при построении модели, о том, что наличие базы данных является важной частью для сбора и хранения данных.

С помощью регрессионного анализа была построена модель, которая на 59.7% описывает представленные данные и примерно с такой же точностью способна предсказывать значения в будущем. Такое качество оценивания сказывает нехваткой данных для построения модели, а также не принятием во внимание некоторых факторов, замерить которых не предстоит возможности, но которые могут заметно влиять на поведение и распространение клещей соответственно.

Однако всё же было выяснено что наибольший вес на количество случаев обращения влияет месяц и день недели, так как выходные дни и начало летних месяцев, когда происходит потепление до комфортных 15-20 градусов тепла и началом периода отпусков и каникул.

Была реализована база данных, в которой были агрегированы необходимые сведенья для последующего анализа. Теперь возможность чтения данных имеется в любой момент времени, необходимо лишь подключится к PostgreSQL используя логин и пароль.

# Список использованной литературы

1. <https://lenta.ru/articles/2016/06/17/acari/>
2. [https://irksib.ru/allnews/12-social/23845-76-ukusov-kleshchej-zaregistrirovali-v-irkutskoj-oblasti-za-nedelyu#:~:text=Активность%20клещей%20возрастает%20в%20Иркутской,Шелеховского%2C%20Баяндаевского%20и%20Качугского%20районов](https://irksib.ru/allnews/12-social/23845-76-ukusov-kleshchej-zaregistrirovali-v-irkutskoj-oblasti-za-nedelyu#:~:text=%D0%90%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C%20%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9%20%D0%B2%D0%BE%D0%B7%D1%80%D0%B0%D1%81%D1%82%D0%B0%D0%B5%D1%82%20%D0%B2%20%D0%98%D1%80%D0%BA%D1%83%D1%82%D1%81%D0%BA%D0%BE%D0%B9,%D0%A8%D0%B5%D0%BB%D0%B5%D1%85%D0%BE%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%2C%20%D0%91%D0%B0%D1%8F%D0%BD%D0%B4%D0%B0%D0%B5%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%20%D0%B8%20%D0%9A%D0%B0%D1%87%D1%83%D0%B3%D1%81%D0%BA%D0%BE%D0%B3%D0%BE%20%D1%80%D0%B0%D0%B9%D0%BE%D0%BD%D0%BE%D0%B2)
3. [https://glavses.ru/уничтожение-клещей/активность-клещей](https://glavses.ru/%D1%83%D0%BD%D0%B8%D1%87%D1%82%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9/%D0%B0%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9)
4. <https://ria.ru/20210401/kleschi-1603726822.html>
5. <https://tiobe.com/tiobe-index/>
6. <https://wiki.loginom.ru/articles/descriptive-analytics.html>
7. <https://www.rospotrebnadzor.ru/documents/details.php?ELEMENT_ID=16919>
8. <https://moscow-777.ru/prochee/kogda-snizhaetsya-aktivnost-kleshhej.html>
9. [https://vc.ru/u/425321-sf-education/224225-chto-takoe-regressionnyy-analiz#:~:text=Регрессионный%20анализ%20—%20это%20набор,можно%20зафиксировать%20изменение%20«зависимой%20переменной](https://vc.ru/u/425321-sf-education/224225-chto-takoe-regressionnyy-analiz#:~:text=%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9%20%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%20%E2%80%94%20%D1%8D%D1%82%D0%BE%20%D0%BD%D0%B0%D0%B1%D0%BE%D1%80,%D0%BC%D0%BE%D0%B6%D0%BD%D0%BE%20%D0%B7%D0%B0%D1%84%D0%B8%D0%BA%D1%81%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D1%82%D1%8C%20%D0%B8%D0%B7%D0%BC%D0%B5%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5%20%C2%AB%D0%B7%D0%B0%D0%B2%D0%B8%D1%81%D0%B8%D0%BC%D0%BE%D0%B9%20%D0%BF%D0%B5%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D0%BE%D0%B9)»
10. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ ОЦЕНКИ АКТИВНОСТИ И ПРОГНОЗА РАСПРОСТРАНЕНИЯ ИКСОДОВЫХ КЛЕЩЕЙ НА ПРИМЕРЕ ГОРОДА ИРКУТСКА - статья

# 

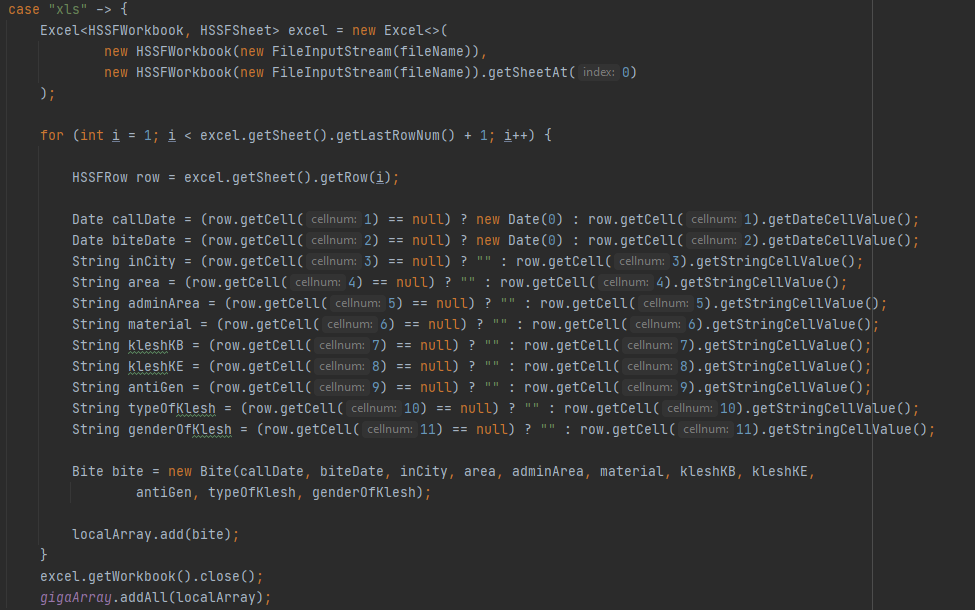
# Приложения



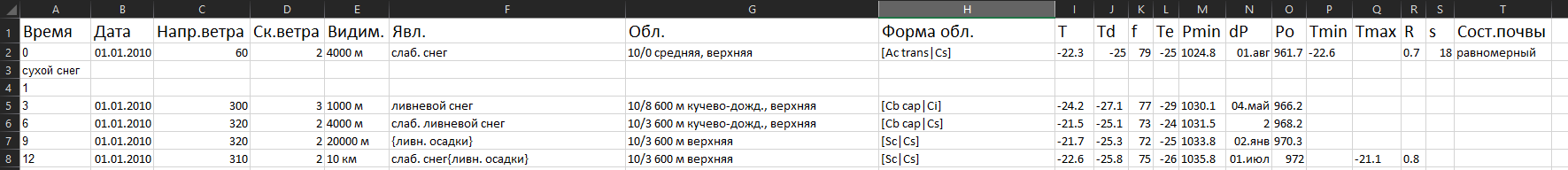
Рис. 1 Количество заболевших энцефалитом и боррелиозом в регионах Российской Федерации за 2021 год [[3]](https://glavses.ru/%D1%83%D0%BD%D0%B8%D1%87%D1%82%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9/%D0%B0%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9)



Рис. 2 Эндемичность регионов Российской Федерации с шансом заболевания клещевым энцефалитом [[3]](https://glavses.ru/%D1%83%D0%BD%D0%B8%D1%87%D1%82%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9/%D0%B0%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C-%D0%BA%D0%BB%D0%B5%D1%89%D0%B5%D0%B9)



Листинг 1 Код, читающий данные электронной таблицы, заносит их в массив



(Рис 3. Структура Excel-таблицы данные\_по\_погоде\_2008\_2013.xlsx)