

Исследования данных арбитражных судов Российской Федерации: методики и первые результаты

Денис Савельев,
научный сотрудник Института проблем правоприменения
при Европейском университете в Санкт-Петербурге, канд.
юрид. наук dsaveliev@eu.spb.ru, www.enforce.spb.ru

Санкт-Петербург, Президентская библиотека 04.12.2020

Эмпирические правовые исследования

- *Выводы, основанные на данных, а не интуиции*
- *Проверка теории доказательными методами*
- *Уметь находить и получать данные*
- *Уметь ставить вопросы к данным*
- *Объединение данных дает новое качество*
- *Использовать действительно современные методы анализа данных в сотрудничестве с профессионалами*
- *Уметь интерпретировать результаты и соотносить их с реальностью*

Как мы это делаем

The screenshot displays the PyCharm IDE interface. The main editor window shows two Python files. The left file, `dsaveliev_05_add_okved.py`, contains code for processing a DataFrame `okved_2018` and generating a heatmap. The right file, `dsaveliev_01_collect_document_types.py`, contains code for collecting document types from a CSV file. The Python Console at the bottom shows the output of the `okved_2018.info()` command, displaying the DataFrame's structure and memory usage.

```
okved_2018.set_index('okved', inplace = True)

n_parties_max = pd.DataFrame(index = okved_2018.index, column
n_parties_max.fillna(0, inplace = True)
n_parties_max = n_parties_max.add(okved_2018.N, axis='columns
n_parties_max = n_parties_max.add(okved_2018.N, axis='index')

result5 = result3.div(n_parties_max)
sns.set()
ax = sns.heatmap(result5, cmap='Blues')

result6 = result5 * 100
sns.set()
ax = sns.heatmap(result6, cmap='OrRd')

writer = pd.ExcelWriter(WD+'crosstab_who_sues_whom.xlsx')
result6.to_excel(writer)
writer.save()
```

```
# This script returns names of ContentTypes as well as its q
# to interim measures or not

import pandas as pd
from datetime import datetime

DOCUMENTS_DATA = 'D:\\commercial_courts_project\\data\\overvi
CHUNKSIZE = 10**4
SEARCH_FIELD = 'ContentTypes'

GET_FIELDS = ['ContentTypes']
WD = 'D:\\dsaveliev\\tmp\\'
OUTFILENAME = WD + 'documenttypes'

def chunk_handler(chunk):
    """
    :param chunk: slice of data csv
    :return: all of contenttype strings found in a chunk
    """
```

Python Console:

```
...: ax = sns.heatmap(result6, cmap='OrRd')
...:
In[31]: okved_2018.info()
<class 'pandas.core.frame.DataFrame'>
Index: 91 entries, 60 to 99
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ---
0   year    91 non-null           int64
1   N       91 non-null           int64
dtypes: int64(2)
memory usage: 4.6+ KB

In[32]:
```

Special Variables:

- WD = (str) 'D:\\dsaveliev\\tmp\\'
- ax = (AxesSubplot) AxesSubplot(0.125,0.11;0.62x0.77)
- cases_all = (DataFrame) caseid ... plaintiff_o...View as DataFrame
- n_parties_max = (DataFrame) okved 60 96 41 43 ... 12...View as DataFrame
- okved_2018 = (DataFrame) year N\\nokved \\n60 2C...View as DataFrame
- okved_stats = (DataFrame) year okved N\\n0 2014 60 676...View as DataFrame
- plaintiff_okveds = (Index) Index(['', '01', '01.10', '01.11', '01.11.10', '01.11.11', '01.11... View
- result3 = (DataFrame) defendant_okved_first 01 02 03 05 (...View as DataFrame
- result5 = (DataFrame) 01 02 03 ... 96 97 98 99...View as DataFrame
- result6 = (DataFrame) 01 02 03 ... 96 97 98 99...View as DataFrame
- writer = (_OpenpyxlWriter) <pandas.io.excel._openpyxl._OpenpyxlWriter object at 0x000...

Таблицы данных и их объединение

- *Данные карточки арбитражного дела (17+ млн)*
- *Данные каждого документа по делу (230+ млн.)*
- *Данные ЕГРЮЛ/ЕГРИП/ФО (11 млн.ИНН, 3.5млн. действующих)*
- *Классификаторы, напр.,ОКВЭД*
- *Тексты решений и постановлений судов (9+ млн. в доступном датасете)*
- *Данные об арбитражных делах предоставлены компанией Право.ру для научных целей, иные данные получены из соответствующих официальных источников*

Аналитическая записка – форс-мажор



– 3,5 млн решений арбитражных судов первой инстанции за 2012–2018 гг., по гражданским делам

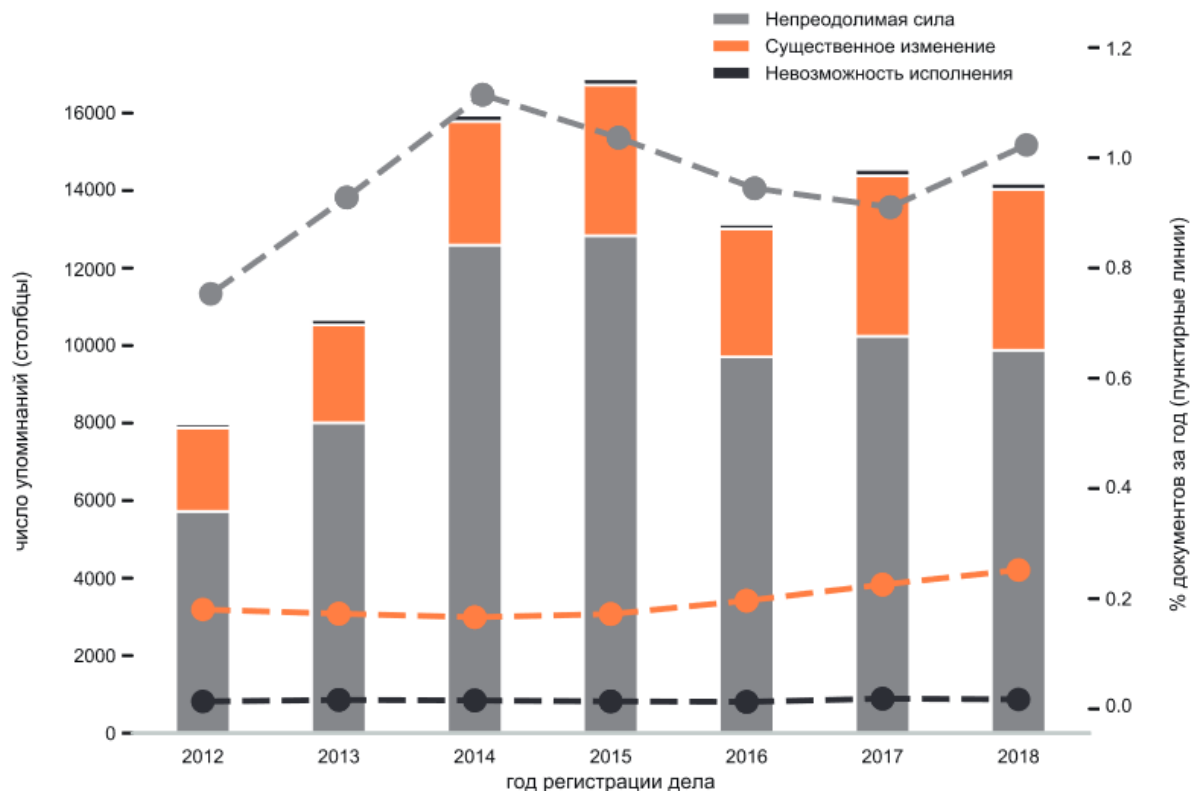
– Трактовка судебной системой ответственности расходится с пониманием заявителей.

Савельев Д. А. Решения арбитражных судов в сфере применения норм о невозможности исполнения обязательств: аналитическая записка / под ред. К. Д. Титаева, Д. А. Скугаревского. — СПб. : Институт проблем правоприменения при Европейском университете в Санкт-Петербурге, 2020.; вып. 8(2020)). — 12 с.

https://enforce.spb.ru/images/analit_zapiski/force_majeure_online.pdf

Аналитическая записка – форс-мажор

Рисунок 1. Частота упоминаний терминов, связанных с невозможностью исполнения обязательств



На столбчатой диаграмме показано абсолютное число упоминаний указанных терминов в год в документах, в которых таких упоминаний 2 и более. Это число отложено по оси Y слева. Подсчитывалось каждое упоминание в документе. Пунктирными линиями отложена доля (в процентах) числа документов, содержащих два и более таких упоминания, среди всех дел за указанный год в используемой выборке. Доля отложена по оси Y справа.

Аналитическая записка – форс-мажор

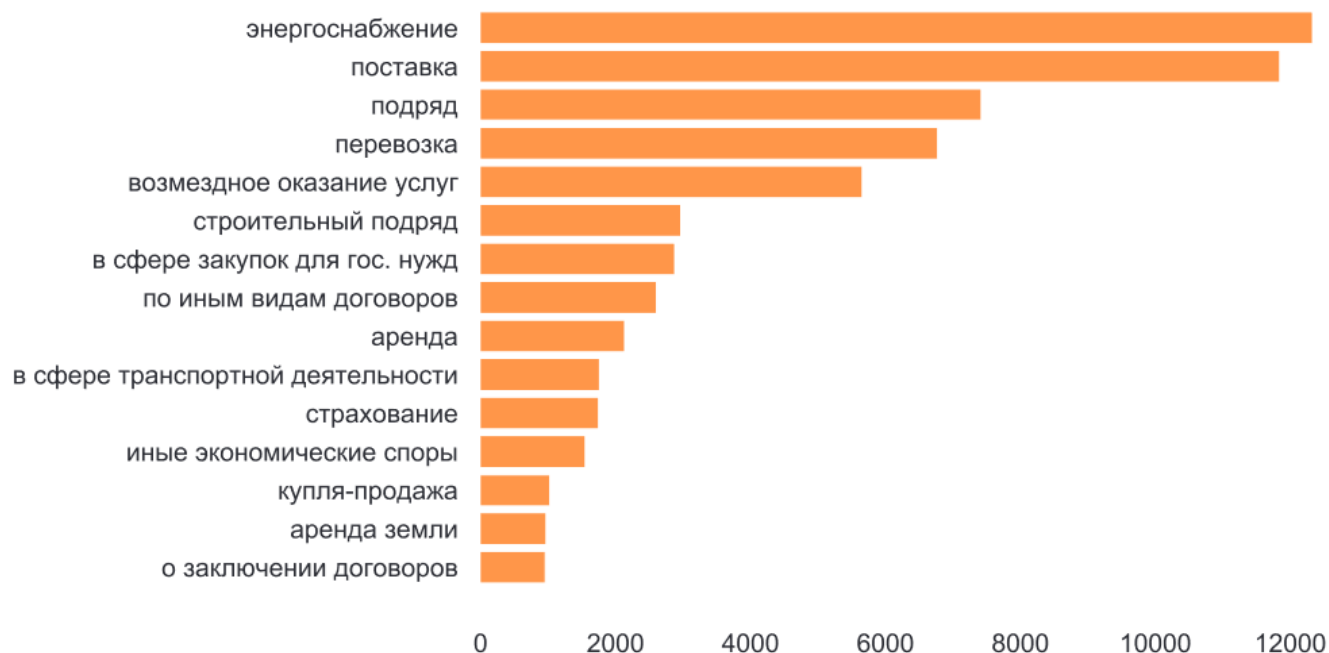
Таблица 1. Распределение числа документов по количеству упоминаний непреодолимой силы и соответствующей статьи ГК РФ

<i>Упоминаний в документе</i>	<i>Документов</i>
20 и более	5
15–19	14
10–14	97
5–9	1308
4	1626
3	4976
2	18 936
1	70 136

При чтении текстов решений выясняется, что суды часто допускают цитирование положений текста ГК РФ о форс-мажоре, тогда как из решения однозначно невозможно установить, что наличие форс-мажора заявлялось сторонами или было предметом рассмотрения. Такое цитирование можно назвать избыточным. В связи с этим далее мы учитывали документы, в которых число упоминаний было 2 и более. Это существенно снижает число найденных документов (как видно из табл. 1), однако позволяет увидеть документы, в которых упоминание термина, скорее всего, не является случайным.

Аналитическая записка – форс-мажор

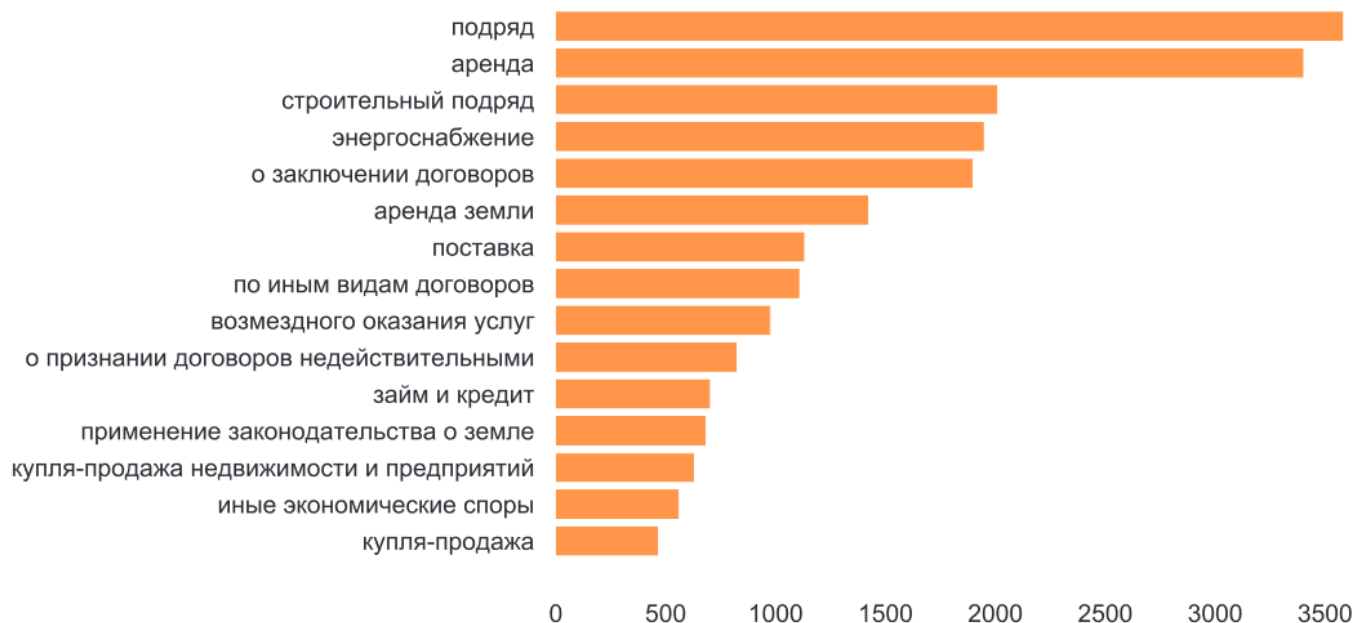
Рисунок 2. Категории дел с наибольшим числом упоминаний непреодолимой силы



На рисунке представлено число решений, которые содержат 2 и более упоминания указанных терминов, по категориям дел. В названиях категорий дел удалено выражение «о неисполнении или ненадлежащем исполнении договоров в сфере...». Представлено абсолютное значение числа, а не доля среди решений по категории, поскольку число дел, относящихся к той или иной категории, может различаться в десятки тысяч раз, и при этом доля будет высока у дел из категорий, где малое общее число дел. Разница в числе упоминаний по категориям имеет свое значение для оценки независимо от доли в документах одной категории.

Аналитическая записка – форс-мажор

Рисунок 3. Категории дел с наибольшим числом упоминаний существенного изменения обстоятельств



На рисунке представлено число решений, которые содержат 2 и более упоминания указанных терминов, по категориям дел. В названиях категорий дел удалено выражение «о неисполнении или ненадлежащем исполнении договоров в сфере...». Представлено абсолютное значение числа, а не доля среди решений по категории, поскольку число дел, относящихся к той или иной категории, может различаться в десятки тысяч раз, и при этом доля будет высока у дел из категорий, где малое общее число дел. Разница в числе упоминаний по категориям имеет свое значение для оценки независимо от доли в документах одной категории.

Типичные ошибки

- **Изначальная неполнота данных**
- **Плохое заполнение, неправильное ведение классификаторов**
- **Незнание данных и неумение их чистить**
- **Неучет ненаблюдаемых значений переменных**
- **Неиспользование нормирования**
- **Неиспользование анализа «при прочих равных»**
- **Ошибочные представления о распределениях**
- **Ковариация – еще не каузальность**

Аналитическая записка – читаемость



– 884 842 текста решений арбитражных судов России по делам 2013–2017 гг., в которых участвовал крупный бизнес и банки

— решения по административным делам написаны значительно сложнее решений по гражданским делам

Савельев Д. А., Кучаков Р. К. Решения арбитражных судов субъектов Российской Федерации: лексическое и синтаксическое качество текстов: (Аналитические записки по проблемам правоприменения; вып. 3 (2019)). — 20 с

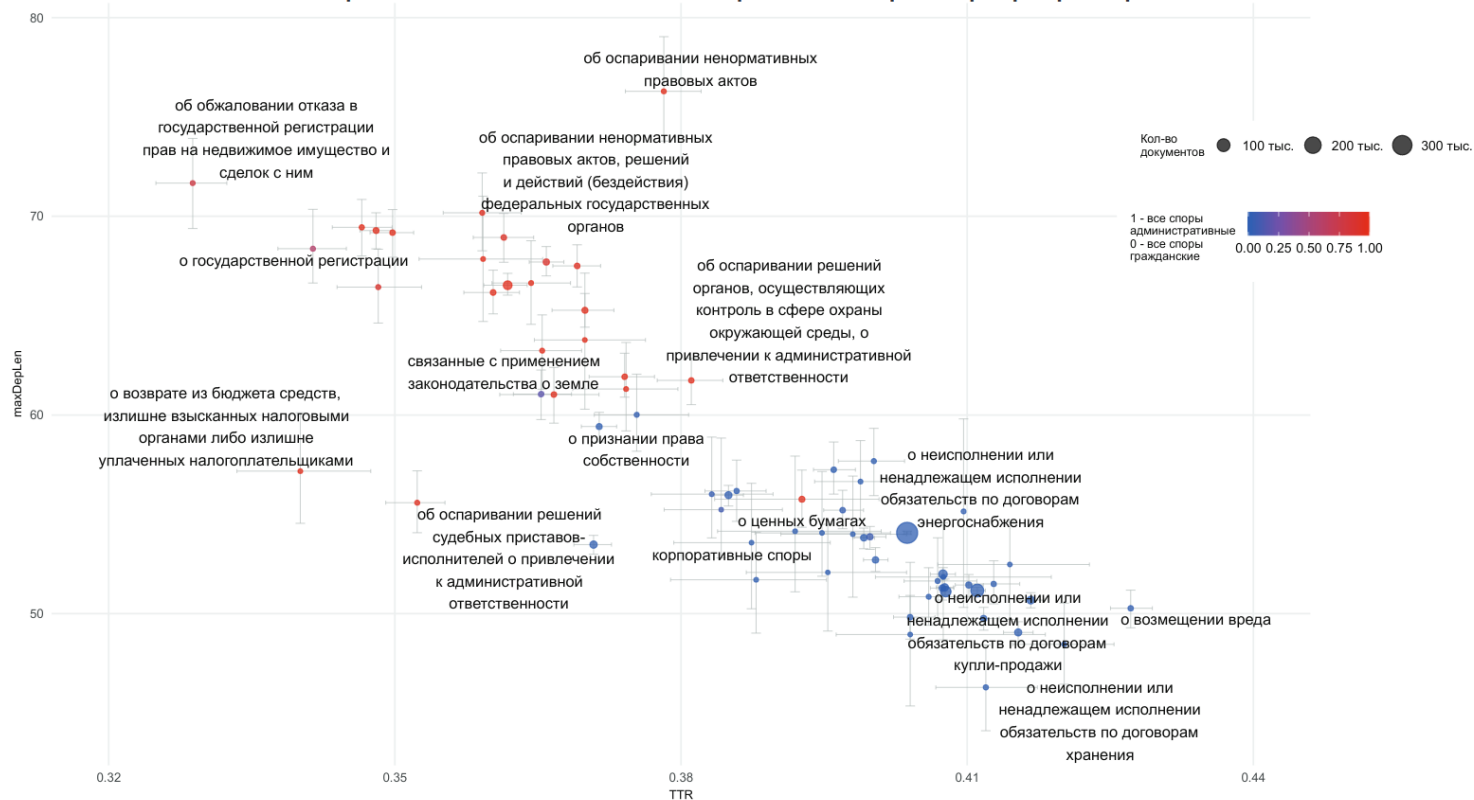
https://enforce.spb.ru/images/analit_zapiski/readability_commercial_courts.pdf

Морфосинтаксическая разметка

www.eu.spb.ru

Аналитическая записка – читаемость

Рис. 4. Предсказанные значения TTR и maxDepLen по категории спора при прочих равных



На осях X и Y отложены предсказанные значения TTR и maxDepLen. Точки отражают предсказанные значения соответствующих метрик из регрессий значения данной метрики в судебном решении на год вынесения решения, результат рассмотрения дела, категорию спора, сумму исковых требований, суд, присутствие в том же регионе апелляционного или кассационного суда, количество сторон в деле, количество документов, упрощенное производство, количество токенов в тексте и квадрат от данной величины. Используются тексты решений таких категорий споров, которые содержат 50 и более решений. Из выборки удалены тексты, длина которых превышает 99,9 процентиль. Серые пересечения в виде крестов отражают доверительные интервалы, построенные с помощью бутстрэпа с 1000 репликаций и покрывающие 95% значений переменной. Диаметр точки соответствует количеству решений. Цвет точки выступает в качестве градиента, отражающего отношения административных и гражданских дел в указанной категории спора. Для облегчения читаемости рисунка использованы только некоторые названия категорий.

Спасибо за внимание!

Автор выражает огромную признательность всем специалистам, работавшим и работающим в арбитражной системе, которые сделали возможными сбор этих данных, компании Право.Ру, предоставившей их для научных исследований, и коллегам в Институте проблем правоприменения, Кириллу Титаеву, Дмитрию Скугаревскому, Руслану Кучакову, Арине Дмитриевой, и всем, кто участвовал и помогал в разработке исследований

Отдельное огромное спасибо Президентской библиотеке за возможность выступления

Денис Савельев