Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Ярославский государственный университет им.
П. Г. Демидова»

Кафедра информационных и сетевых технологий

Сдано на кафедру
« 15 » июня 2020 г.
Заведующий кафедрой,
к. ф.-м. н., декан

_____ Д. Ю. Чалый

Выпускная квалификационная работа

Разработка алгоритма иерархической кластеризации для снижения размерности при составлении оптимального инвестиционного портфеля

по направлению
09.04.03 Прикладная информатика

Научный руководитель
д. э. н., доцент

_____ Е. М. Спиридонова
« 15 » июня 2020 г.

Студент группы ПИЭ-21МО

_____ А. Ю. Полетаев
« 15 » июня 2020 г.

Ярославль, 2020

Реферат

Объем 56 с., 3 гл., 6 рис., 14 табл., 21 источников, 2 прил.

Ключевые слова: **алгоритмы, инвестиционные портфели, кластеризация, прикладная информатика.**

В данной работе предлагается алгоритм сокращения размерности при составлении оптимального инвестиционного портфеля, позволяющий ускорить его работу. Исследуются возможности применения кластеризации для снижения размерности и история работ по данному вопросу; производится и обосновывается выбор способа вычисления расстояния между ценными бумагами. Приводится разработанный алгоритм и изучаются результаты его работы на реальных наборах данных.

Содержание

Введение

Составление оптимального портфеля ценных бумаг является важным и частым случаем решения задачи оптимизации. Практическое применение существующих методов составления оптимального портфеля часто затруднено или экономически не обосновано из-за большого числа доступных для инвестирования ценных бумаг (и, как следствие, большой размерности исходных данных). Следовательно, необходимо исследовать возможность сокращения размерности данных о ценных бумагах за счёт кластеризации — объединения ценных бумаг в группы — кластеры.

В первой части данной работы приводится обзор предпринимавшихся ранее попыток использовать кластеризацию для составления инвестиционных портфелей и даётся краткая справка по алгоритмам кластеризации.

Во второй части данной работы предлагается алгоритм снижения размерности исходных данных при составлении инвестиционного портфеля, основанный на иерархической кластеризации доступных для инвестирования ценных бумаг. В качестве меры близости ценных бумаг для иерархической кластеризации используется мера расстояния, основанная на коэффициенте парной корреляции Пирсона.

В третьей части работы описывается программная реализация предложенного алгоритма на языке программирования *Python*; далее исследуется влияние предложенного метода на качество получаемого оптимального решения на нескольких примерах составления оптимального портфеля ценных бумаг по модели Марковица. Также исследуется влияние параметров иерархической кластеризации (метрики межкластерного расстояния и порогового значения кластеризации) на изменение качества получаемого оптимального решения. Исследуется зависимость между целевой доходностью портфеля и возможностью снижения размерности с помощью предложенного метода. Для каждого рассмотренного примера приводятся графики и таблицы с основными полученными результатами применения метода — понижением размерности и падением доходности (снижением качества оптимального решения) у портфеля, построенного с применением предложенного метода по сравнению с портфелем, построенным без применения предложенного метода.

1. Задача снижения размерности при составлении оптимального инвестиционного портфеля

1.1. Постановка задачи

Составление оптимального портфеля ценных бумаг является важным и частым случаем решения задачи оптимизации. Согласно портфельной теории, впервые сформулированной Гарри Марковицем в 1952 г., при составлении инвестиционного портфеля нужно ориентироваться:

во-первых, на то, насколько высокий доход он способен принести своему владельцу (точнее говоря, насколько сильно приумножить инвестированные в портфель средства)

во-вторых, на то, насколько рисковым является портфель, т.е. на то, насколько велика вероятность того, что реальная выгода владельца портфеля будет отличаться от прогнозируемой [1].

Следовательно, для составления оптимального портфеля из n ценных бумаг необходимо оценить лишь два показателя (см. формулы (1)–(2)).

1. ожидаемую доходность
$$R = \sum_{i=1}^{n} R_i X_i \tag{1}$$
2. меру риска (изменчивости)
$$V = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{i,j} X_i X_j \tag{2}$$

Здесь R_i — ожидаемая доходность i-ой ценной бумаги; X_i — доля средств, инвестированных в неё (следовательно, $\sum_{i=1}^{n} X_i = 1$); $\sigma_{i,j}$ — ковариация доходностей ценных бумаг i и j.

Портфель может быть оптимизирован по заданной ожидаемой доходности (для минимизации ожидаемого риска), по заданному ожидаемому риску (в этом случае будет максимизироваться ожидаемая доходность). Также возможна оптимизация, например, по методу RAPOC (risk-adjusted return), предложенному Д. Борге — тогда максимизируется целевая функция $R - \gamma V$, где γ — некоторый

коэффициент [2]. Тем не менее, при любом из подходов результатом оптимизации является вектор $X = [X_1, \ldots, X_n]$, где X_i — доля i-ой ценной бумаги в оптимальном портфеле.

В настоящее время разработано достаточно много математических методов оптимизации портфеля по Марковицу [3, 4], однако, их общим недостатком является достаточно высокая вычислительная сложность. Объём биржевых данных, как правило, велик (например, в 2015 году только на Нью-Йоркской фондовой бирже торговались акции более 3000 компаний). Кроме того, на практике инвестор практически никогда не составляет оптимальный портфель из ценных бумаг, торгуемых только на одной бирже (кроме акций, это и производные финансовые инструменты, в т.ч. привязанные к товарам, например, фьючерсы на ресурсы), что ещё сильнее увеличит объём используемых для составления оптимального портфеля данных. Поскольку оптимизация портфеля на динамичном рынке может требоваться достаточно часто, особенно при торговле на динамичном рынке, кроме того, необходимо учитывать, что иногда при торговле необходимо «ловить» очень краткосрочные колебания спроса и предложения на ценные бумаги, что требует достаточно быстрой реакции — как человека, так и ассистирующей ему компьютерной программы. Также важным фактором является то, что практически вся работа с инвестиционными портфелями сводится к работе с дробными числами, то, следовательно, практически любое полученное оптимальное решение будет, на самом деле, вычислено с некоторой предельной точностью. Ускорение оптимизации позволит, с одной стороны, за то же время вычислять решение с большей точностью, а, с другой — вычислять не одно, а несколько различных решений, что может помочь, например, когда нужно рассмотреть несколько возможных вариантов, или просчитать возможные действия для нескольких спрогнозированных исходов некоторого события. Исходя из вышесказанного, можно сделать вывод о том, что нужно искать пути ускорения составления оптимального инвестиционного портфеля.

Во-первых, это позволит выполнять более глубокий анализ рыночной ситуации за то же время, что и ранее.

Во-вторых, это позволит трейдерам с не слишком мощными компьютерами лучше конкурировать с теми, у кого есть ресурсы для покупки передовых комплектующих. Этот аспект требует некоторых дополнительных замечаний. В настоящее время для того, чтобы начать активно торговать на бирже, в принципе требуется достаточно существенная сумма денег (о точной существуют разные мнения, но практически все авторы сходятся в том, что иметь нужно не менее, чем 50.000 рублей [6]). К этой сумме необходимо добавить цену достаточно современного компьютера, на котором можно одновременно и выполнять математические расчёты, и наблюдать в реальном времени за динамикой цен и новостями, могущими оказать на них влияние (причём делать всё это достаточно быстро, во

всяком случае, так, чтобы наибольшие задержки рождала пропускная способность интернет-канала и природная способность человека воспринимать информацию). Автор знаком с мнением специалистов, согласно которому стоимость нового достаточно мощного компьютера $\sim$ \$1000, что, даже если обойтись без некоторых ненужных для торговли на бирже функций или взять часть комплектующих б/у, всё равно приводит к тому, что стартовые вложения для начинающего трейдера сейчас обойдутся не менее, чем в 100.000 рублей. Следовательно, возможность вместо покупки нового компьютера обойтись использованием обычного бытового может достаточно сильно снизить финансовый «порог вхождения» в торговлю, что (оставив в стороне морально-этическую сторону вопроса) сделает торги на бирже более конкурентными, и, как считается, эффективнее [5].

В-третьих, полученное решение в будущем можно потенциально обобщить с задачи оптимизации инвестиционных портфелей на задачу оптимизации «в принципе», что само по себе будет важным научным достижением.

1.2. Понижение размерности при составлении инвестиционного портфеля

Поскольку сама по себе задача оптимизации является, в первую очередь, относящейся к областям математики и алгоритмов, конечно же, можно ускорять оптимизацию инвестиционного портфеля за счёт составления новых, более эффективных, алгоритмов. В этом направлении постоянно ведутся научные изыскания, постоянно приносящие новые результаты (и появление новых алгоритмов, и «доводка» и определение эффективности уже существующих).

Также, поскольку само составление оптимальных портфелей в настоящее время выполняется с помощью вычислительной техники, то, конечно, появление новых, более мощных, комплектующих для компьютеров также способно ускорить процесс составления оптимального портфеля. Однако, этот путь, во-первых, слишком зависит от технического прогресса (который вполне может остановиться), а, во-вторых, увеличивает сумму стартового капитала, который необходим для того, чтобы начать заниматься торговлей на бирже.

Существует другой достаточно интересный путь (в хорошем смысле «третий путь») ускорения составления оптимального портфеля — не повышать эффективность алгоритма при его работе на некотором наборе, например, из n доступных для инвестирования ценных бумаг, и не повышать скорость работы того же алгоритма на том же наборе за счёт использования бóльших вычислительных мощностей, а свести набор из n доступных для инвестирования ценных бумаг к набору из k «ценных бумаг», каждая из которых несёт в себе информацию о некоторой группе ценных бумаг из n. То есть, фактически, снизить размерность задачи оптимизации. Конечно, для того, чтобы было ускорить процесс составления оптимального портфеля, должно выполняться $k \ll n$. Полученный в результате портфель будет хуже по своим качественным показателям, чем тот, который был бы получен в результате «честных» расчётов, однако, поскольку сама природа ценных бумаг тяготеет к тому, что в них существуют группы, выделяемые по какому-либо признаку (например, акции фирм одной отрасли или страны, фьючерсные контракты на разные марки нефти и т.д.), можно предполагать, что это снижение качества портфеля окажется не слишком сильным.

Важно отметить, что применение снижения размерности ни в коем случае не ограничивает применение двух других путей ускорения процесса составления оптимального портфеля — использования более мощной компьютерной техники и разработки лучших алгоритмов оптимизации, а скорее наоборот, дополняет их.

Данная идея (оперировать при составлении инвестиционного портфеля не отдельными ценными бумагами, а их группами) достаточно логична, поскольку берёт начало в собственно природе человеческого мышления — стремлении всё вокруг группировать и обобщать. Тем не менее, поскольку инвестирование на-

прямую связано с деньгами, и деньгами большими, то для принятия конкретных решений требуется достаточно хорошо проработанный и обоснованный формальный математический аппарат. Таким аппаратом в данном случае — когда необходимо разделить n объектов на k групп (причём даже примерные границы групп заранее неизвестны) — является кластеризация [7].

Кластеризация (также часто называемая кластерным анализом), говоря формально, — процесс разбиения выборки $X = \{x_1, \ldots, x_l\}$ на несколько кластеров (подмножеств, «сгустков»), образующих множество Y, на основе некоторой меры расстояния между объектами $\rho(x_a, x_b)$ так, чтобы каждый кластер состоял из объектов, близких по ρ, а объекты из разных кластеров различались [9]. Классический подход предполагает, что каждому $x_i \in X$ ставится в соответствие метка кластера $y_i \in Y$, т.е. полученные кластеры не пересекаются, однако, существуют и алгоритмы нечёткой кластеризации, разделяющие X на несколько пересекающихся подмножеств. Кластеризация часто относится к методам машинного обучения без учителя (unsupervised), поскольку метки а множество X в таком случае, соответственно, является обучающей выборкой. Тем не менее, поскольку о множестве Y априори, как правило, ничего не известно, это рождает достаточно сильные проблемы с постановкой задачи и проведением анализа, отдавая очень многое на откуп исследователя. Также кластеризация часто относится к методам статистического анализа, однако, например, в классической работе [8] указывается, что метод кластерного анализа остаётся за общепринятыми рамками дисциплины «математическая статистика», как не опирающийся на вероятностную природу обрабатываемых данных.

Среди алгоритмов кластеризации выделяют две основные группы:

1. Иерархические — строящие систему вложенных одно в другое разбиений выборки на непересекающиеся кластеры, т.е. каждое следующее разбиение как бы «уточняет» предыдущее. Можно говорить о том, что в результате выполнения алгоритма кластеризации получается дерево кластеров, корнем которого является всё X (один большой кластер), а листьями — кластеры меньшего размера. Чаще всего такое дерево строится на основе матрицы парных расстояний (также называемой матрицей близости) — и это является очень интересным в рамках изучаемой темы, т.к. очень схожая ковариационная матрица используется для составления оптимального портфеля по модели Марковица. Чтобы из такого дерева получить непосредственно искомое разбиение, необходимо задать какую-то «границу», например, искомое число кластеров, или расстояние, которое является максимально возможным для того, чтобы объекты могли быть отнесены к одному кластеру. Алгоритмы иерархической кластеризации также разделяются на более агломеративные (или восходящие), вначале помещающие каждый объект в свой кластер, а затем постепенно объединяющие их, и дивизивные

(нисходящие), вначале помещающие все объекты в один большой кластер, а затем постепенно разбивающие его. В настоящий момент более распространёнными являются агломеративные, тем не менее, принципиальная разница между ними невелика.

2. Плоские — строят одно разбиение объектов на кластеры, но, как правило, за несколько итераций, перед первой из которых случайным образом инициализируются центры кластеров, а затем происходит определение того, какие объекты к какому кластеру относятся, и, на основании этой информации, пересчёт центров кластеров. Итерации прекращаются, когда достигнут некий критерий остановки (чаще всего используется некоторое минимальное изменение среднеквадратической ошибки). Наиболее популярным алгоритмом плоской кластеризации в настоящий момент является метод k средних (k-means), также можно выделить метод нечёткой кластеризации с средних (c-means), являющийся модификацией метода k средних.

Таким образом, нужно понимать, что кластеризация является очень мощным, интересным и динамично развивающимся инструментом (очень большую роль в её активном развитии сыграло широкое распространение компьютеров), однако, для успешного применения требующего грамотной «настройки», и, в идеале, контроля своей работы с помощью классических методов математической статистики.

1.3. Существующие подходы к применению кластеризации в составлении оптимальных портфелей

Поскольку, как уже упоминалось ранее, идея сокращения размерности при составлении инвестиционного портфеля не нова (можно даже сказать, что она «витает в воздухе»), к настоящему моменту уже предпринято несколько попыток применения кластеризации при составлении оптимального портфеля.

Из исследований на русском языке можно выделить, в первую очередь, вышедшую в 2010 году статью Е.М. Бронштейна и И. Н. Ишманова [10]. Несмотря на кажущееся странным относительно темы снижения размерности название, данная работа, скорее всего, является первым русскоязычным исследованием возможности применить кластеризацию в задаче оптимизации портфеля по модели Марковица. В то же время, проведённое авторами статьи исследование не лишено некоторых достаточно существенных недостатков. Во-первых, для кластеризации используется метод k средних, который для своей работы требует заранее задать число выделяемых кластеров, что делает его слабо применимым на практике для задачи разделения на кластеры ценных бумаг на бирже (т.к. число их групп, во-первых, практически всегда заранее неизвестно, во-вторых, достаточно велико, что делает практически невозможным его динамическое определение).

Во-вторых, кластеризация проводится на основе четырёх параметров для каждой ценной бумаги — математического ожидания, дисперсии, асимметрии и эксцесса (поскольку асимметрия и эксцесс являются безразмерными величинами, авторы использовали отношение математического ожидания к дисперсии). Безотносительно оценок «удачности» или «неудачности» выбора конкретных характеристик для проведения кластеризации, необходимо упомянуть то, что они, как и было задумано исследователями, позволили выделить высокодоходные и нерисковые ценные бумаги, однако, никак не разделили между собой ценные бумаги с противоположной изменчивостью, более того — они могли попасть в один кластер и, поскольку по авторской методике из кластера для инвестирования выделяется только одна ценная бумага, то в итоговый портфель попадала бы только одна из них. Такой подход очевидно снижает диверсифицированность портфеля, а, кроме того, приводит к тому, что из всего множества возможностей инвестирования используется только часть. В-третьих, объём проведённых экспериментов — 15 опытов на 90 акциях с информацией о ценах на них за 20 периодов, скорее, нельзя считать достаточным для определения реальной эффективности проводимой кластеризации, сила которой более полно раскрывается на большем числе объектов (к тому же, как было упомянуто ранее, на биржах торгуются тысячи ценных бумаг).

Также из отечественных исследований можно выделить работу Е.М. Тюховой и Д.С. Сизых [11]. Данная работа является, безусловно, очень качественно проработанной с точки зрения экономики, однако, в ней, как и в [10], для группировки ценных бумаг используется метод k средних. В то же время, в работе используется сильная сторона метода k средних — то, что он способен определять принадлежность наблюдения к кластеру с некоторой вероятностью, и предпринимается попытка преодолеть необходимость задания числа кластеров, определяя число кластеров по количеству обобщённых факторов, полученных в результате факторного анализа. Но, поскольку даже на самых обширных данных получить много факторов получится с трудом, а на реальных биржевых данных кластеров может существовать очень большое количество, то данный метод нельзя считать общеприменимым. Кроме того, выбранные авторами для проведения кластерного анализа факторы: котировки стоимости акций и динамические показатели их изменения и основные рыночные мультипликаторы деятельности их компаний эмитентов делают негарантированным непопадание в один кластер акций со схожей изменчивостью (и, как следствие, попадание в разные кластеры акций с различной изменчивостью). Также стоит отметить, что использование в качестве фактора кластеризации информации о компании-эмитенте (т.е. не относящейся напрямую к динамике цен на акции), с одной стороны, безусловно повышает качество решения с экономической точки зрения (например, может позволить «выбросить» из рассмотрения некоторые совсем «мусорные» акции), тем не менее, не позволяет впоследствии использовать полученную методику

для сокращения размерности в не-экономических задачах, что, как следствие, снижает универсальность выработанного решения.

Говоря об иностранных работах, нужно упомянуть, по-видимому, первое исследование на схожую тематику — статью итальяно-американской группы учёных под руководством V. Tola [12]. В ней предлагается производить разбиение акций на кластеры с помощью иерархической кластеризации. Однако, в упомянутом исследовании не производится расчёта ковариационной матрицы доходностей кластеров, а для оптимизации портфеля по модели Марковица в качестве матрицы σ используется полученная в ходе иерархической кластеризации матрица межкластерных корреляций. Такой подход, с одной стороны, ускоряет проведение расчётов, но с другой — авторами [12] признаётся риск того, что полученная матрица окажется отрицательно определённой (что сделает невозможным дальнейшие вычисления), однако, данный риск игнорируется, поскольку он ни разу не реализовался в ходе экспериментов.

Идея о применении иерархической кластеризации высказывается и в недавнем исследовании [13], выполненном в рамках проекта по исследованию оптимизации портфелей, основанной на кластеризации (в настоящий момент работа по данному проекту, скорее всего, продолжается, поэтому можно ждать новых работ той же группы учёных). Однако, из-за слишком сильной ориентированности на практику (например, использование для оценки качества полученного оптимального решения метода Шарпа, специфичного для инвестиционных портфелей), результаты [13] сложно использовать для решения других задач оптимизации, кроме оптимизации инвестиционных портфелей.

Крайне интересная идея показана в работе Marcos López de Prado [14] — согласно ей, предварительное применение иерархической кластеризации над ковариационной матрицей позволяет сделать полученный в результате оптимальный портфель более устойчивым. Причём, что особенно важно, данный эффект теоретически применим не только к оптимизации инвестиционных портфелей, но и к задаче оптимизации «в целом», что особенно ценно на фоне остальных исследований. Работу [14] хочется особенно отметить также за приведённый полностью программный код, с помощью которого были получены результаты.

Исходя из вышесказанного, можно сделать следующие выводы:

1. задача снижения размерности при составлении оптимального инвестиционного портфеля крайне актуальна
2. есть постоянный прогресс в создании новых техник снижения размерности за счёт применения кластеризации, причём, в первую очередь, среди зарубежных исследований
3. в то же время, «идеальной» до сих пор не появилось, а самой частой проблемой существующих является слишком сильная ориентация на экономические реалии, из-за чего снижается возможность применения полученного

решения для других задач.

4. и у иерархической кластеризации, и у метода k средних для снижения размерности есть свои преимущества
5. разделение на кластеры должно быть таким, чтобы в одном кластере оказывались ценные бумаги с максимально схожей изменчивостью (а, соответственно, в разных кластерах — с противоположной изменчивостью)
6. кроме непосредственного снижения размерности, применение иерархической кластеризации имеет и другие положительные эффекты

Следовательно, задача состоит в том, чтобы предложить алгоритм снижения размерности, который будет использовать, предпочтительнее всего, иерархическую кластеризацию, и с её помощью на основе взаимной изменчивости разделять n ценных бумаг на k кластеров для того, чтобы снизить размерность в задаче оптимизации инвестиционного портфеля.

2. Предлагаемый алгоритм для снижения размерности

2.1. Основные этапы алгоритма

Опишем основные этапы алгоритма, требуемого для того, чтобы снизить размерность решаемой задачи составления оптимального инвестиционного портфеля.

Например, необходимо составить оптимальный портфель (задаваемый вектором весов X) с минимальным риском для заданной доходности R. Для инвестирования доступно n ценных бумаг (множество N), для которых есть данные об изменениях цен на за m периодов, формирующие матрицу P, где P_{ij} — данные о стоимости ценной бумаги i в момент j. Поскольку стоимости ценных бумаг являются динамичными и разноразмерными, тогда как при составлении реального портфеля практически нет разницы, изменилась ли, например, цена акции с \$10 до \$11, или с \$100 до \$110 — важно то, что она увеличилась на 10%, что увеличило инвестированные в покупку этих акций средства на 10%, в дальнейшем работа будет происходить не с абсолютным изменением стоимости ценных бумаг, а с относительным, за базу отсчёта принята цена в начале исследуемого периода. Таким образом, для $\forall x \in N : P_{x1} = 1{,}0$, а вся матрица будет иметь вид:

$$P = \begin{pmatrix} 1 & p_{12} & \dots & p_{1m} \\ 1 & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & p_{n2} & \dots & p_{nm} \end{pmatrix}$$

Также дан вектор доходностей акций Y. Для упрощения можно за доходность принять отношение стоимости ценной бумаги в последнем наблюдаемом периоде к стоимости в первом наблюдаемом периоде: $r_i = x_{im}/x_{i1}$ (т.е. доход, который будет получен, если купить ценную бумагу в начале наблюдаемого периода и продать в конце), однако, при реальном применении можно без труда заменить такую «заглушку» на значения ожидаемой доходности, полученные в результате применения какого-либо метода прогнозирования.

Тогда у алгоритма можно выделить следующие этапы:

1. Разделение n ценных бумаг на k кластеров ($k \ll n$), что может быть выражено функцией $N \to K$, сопоставляющей каждой ценной бумаге $n_i \in N$ её метку кластера $k_i \in K$. Результатом данного этапа будет являться вектор

меток кластеров

$$marks = \begin{bmatrix} mark_1 & \dots & mark_n \end{bmatrix}$$

где $mark_i = k \in K$ — метка, характеризующая принадлежность i-ой ценной бумаги к кластеру k.

2. Пересчёт цен для получения динамики «цен» каждого кластера в среднем. Данный этап также можно выразить функцией $K \to P^c$, ставящей каждому кластеру ценных бумаг k_i в соответствие информацию об изменении средней стоимости акций, его формирующих

$$P_i^c = \begin{bmatrix} 1 & p_{i2}^c & \dots & p_{im}^c \end{bmatrix}$$

Результатом данного этапа будет матрица динамики «цен» кластеров:

$$P^c = \begin{bmatrix} 1 & p_{12}^c & \dots & p_{1m}^c \\ 1 & p_{22}^c & \dots & p_{2m}^c \\ \vdots & \vdots & \ddots & \vdots \\ 1 & p_{k2}^c & \dots & p_{km}^c \end{bmatrix}$$

где p_{ij}^c — «цена» i-ого кластера в момент j.

3. Расчёт «доходности» каждого кластера как средней доходностей входящих в него ценных бумаг, также описываемый в виде функции как $K \to Y^c$. Практически аналогично предыдущему пункту (особенно в связи с принятыми нами допущениями) каждому кластеру k_i ставится в соответствие его «доходность» Y_i^c. Результатом данного этапа будет являться вектор «доходностей» кластеров

$$Y^c = \begin{bmatrix} y_1^c & \dots & y_k^c \end{bmatrix}$$

где y_i^c — «доходность» i-ого кластера.

4. Оптимизация портфеля на основе «доходностей» кластеров и динамики их «цен» в заданных ограничениях. В рассматриваемом примере таким ограничением будет являться минимальная доходность $R_{\min}$, однако, возможно и ограничение максимального риска $V_{\min}$, и оптимизация на основе уже упоминавшегося выше *RAPOC*. Также введём ограничение $\forall x_i \in X : x_i > 0$, т.е., фактически, запретим короткие продажи. Это необходимо, во-первых, для упрощения самой задачи (чтобы не «множить сущности»), а, во-вторых, для повышения универсальности алгоритма. Результатом данного этапа будет являться вектор весов кластеров в оптимальном портфеле

$$W = \begin{bmatrix} w_1 & \dots & w_k \end{bmatrix}$$

где w_i — доля i-ого кластера в оптимальном портфеле.

5. Расчёт на основе полученного на предыдущем этапе вектора весов кластеров в портфеле W вектора X индивидуальных весов для каждой ценной бумаги. Результатом данного этапа будет искомый вектор индивидуальных весов ценных бумаг

$$X = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$$

где x_i — доля средств, затраченных на создание инвестиционного портфеля, инвестированных в i-ую ценную бумагу.

2.2. Разделение ценных бумаг на кластеры

Как уже было упомянуто ранее, существует множество методов, позволяющих решить задачу разделения n наблюдений на k кластеров. Кроме того, конечно, для решения поставленной задачи можно изобрести новый алгоритм, если ни один из уже существующих не будет признан подходящим.

Для того, чтобы принять решение о том, какой именно метод кластеризации будет использоваться в описываемом алгоритме, необходимо сформулировать, какими свойствами он должен обладать:

Во-первых, ценные бумаги должны объединяться в кластеры на основе схожести динамики их цен. Заметим, что для обеспечения универсальности решения лучше всего будет объединять их только на основе динамики цен, не привлекая к анализу, например, данные об эмитентах.

Во-вторых, объединение должно быть устойчивым (в смысле неизменности основной структуры разделения на кластеры при добавлении новой ценной бумаги, а также незначительных изменений числа выделяемых кластеров, максимального расстояния, для которого возможно объединение, или других параметров алгоритма).

В-третьих, алгоритм не должен требовать вручную задавать число выделяемых кластеров (во всяком случае, должен иметь возможность проводить кластеризацию без этого задания).

Если первое свойство логически следует из самой постановки задачи и не нуждается в дополнительном комментарии, то второе и третье разумно будет пояснить дополнительно.

Устойчивость объединения позволит, с одной стороны, грамотно подстроить параметры алгоритма под конкретный фондовый рынок (поскольку он, как и любая крупная и сложившаяся система, обладает внутренней устойчивостью, позволяющей на основе информации о прошлом строить предположения о будущем), а, с другой стороны, искать параметры алгоритма, общие для всех фондовых рынков в принципе.

Необходимость возможности работы без предварительного задания числа выделяемых кластеров k позволит алгоритму предсказуемо работать на «неизвестных» данных (говоря языком машинного обучения, на тестовой выборке). Представляется гораздо более перспективным задавать число кластеров не напрямую, а косвенно, посредством задания минимальной схожести динамики цен, для которой возможно включить ценные бумаги в один кластер.

Конечно, возможен и обратный подход — представляя, насколько сильно требуется снизить размерность задачи оптимизации, задать соответствующее такому понижению требуемое число кластеров (например, имея $n = 1500$ и

необходимость снизить размерность в 3 раза, можно задать $k = 500$). Такое решение, безусловно, является интересным, однако, может привести к нежелательным результатам — не слишком критичным в случае, если, например, вместо 500 кластеров их на рынке более-менее обоснованно можно выделить только 499 (в этом случае, например, один из кластеров окажется ошибочно разделённым ещё на два), или 501 (в этом случае будет просто не выделен ещё один кластер, что снизит возможности диверсификации портфеля). Однако, в случае, когда алгоритм применяется в «боевых» условиях, и, в результате какого-либо рыночного сдвига ценные бумаги перестали группироваться так же хорошо, как раньше (и, например, каждая акция стала представлять свой собственный кластер), такой подход приведёт к тому, что будут выделены не существующие группы, в которые будут ложно объединены не слишком похожие по изменчивости ценные бумаги, что может очень сильно ухудшить оптимальный портфель. В то же время, алгоритм, основанный на задании минимальной схожести динамики цен, в такой ситуации отработает более корректно.

Для обеспечения первого свойства необходимо выбрать подходящую меру расстояния для определения схожести в изменчивости ценных бумаг.

Наиболее логичным (и достаточно часто использующимся) подходом к определению схожести изменчивости временных рядов (которыми, по сути, являются данные об изменениях цен на акции) является использование, в той или иной форме, коэффициента парной корреляции между ними (формула (3)):

$$r_{xy} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}} \tag{3}$$

где:

n — число наблюдений;

x_i, y_i — значения x и y соответственно в момент i;

$\overline{x}$, $\overline{y}$ — средние величины x и y соответственно

Значения коэффициента корреляции изменяются в диапазоне $[-1; 1]$ и трактуются следующим образом: $|r_{xy}| = 1$ означает, что x и y функционально зависимы, $|r_{xy}| = 0$ — что x и y изменяются полностью независимо друг от друга. Характер связи описывается знаком r_{xy}: если $r_{xy} > 0$, то связь прямая, если $r_{xy} < 0$ — обратная.

Сам по себе коэффициент парной корреляции в качестве показателя схожести изменчивости очень хорош, особенно с учётом того, что он безразмерен. Однако, как можно видеть, главной проблемой использования «чистого» коэффициента парной корреляции в качестве меры схожести изменчивости ценных

бумаг является то, что он, во-первых, может принимать отрицательные значения, а, во-вторых, что он равен 0, когда ценные бумаги x и y независимы между собой, и 1, когда они изменяются практически одинаково, т.е. «расстояние» меньше для таких ценных бумаг, которые мы бы хотели видеть в разных кластерах, и выше для таких ценных бумаг, которые мы бы хотели видеть в одном кластере.

Выходом является использование не «чистого» коэффициента корреляции, а основанного на нём расстояния. Хорошим примером такого расстояния является, отлично подходящим для нашей цели, является расстояние (4):

$$\rho_{xy} = 1 - r_{xy} \tag{4}$$

Такое расстояние будет равно 0, когда x и y изменяются одинаково, 1, когда они изменяются независимо, и 2, когда они изменяются противоположным образом, т.е. на его основе можно проводить кластеризацию ценных бумаг именно так, как необходимо.

Единственной проблемой такого расстояния является то, что оно не является метрикой (в частности, для него не будет выполняться неравенство треугольника), а, следовательно, его использование в гладких алгоритмах кластеризации затруднено (иерархическая кластеризация к этому нетребовательна). В работе [15] приводится построенная на основе коэффициента парной корреляции не просто расстояние, а расстояние, являющееся метрикой (5).

$$d_{xy} = \sqrt{1 - (r_{xy})^2} \tag{5}$$

Тем не менее, для цели кластеризации ценных бумаг такая метрика оказывается хуже, чем ρ_{xy} (4) (в первую очередь — из-за того, что перестаёт отделять прямую зависимость от обратной). Рассмотрим пример (абстрактный, но вполне возможный при реальном применении алгоритма):

$$P = \begin{bmatrix} 1{,}00 & 1{,}20 & 1{,}30 & 2{,}00 \\ 1{,}00 & 1{,}07 & 1{,}30 & 1{,}10 \\ 1{,}00 & 1{,}20 & 0{,}70 & 0{,}90 \end{bmatrix}$$

тогда

$$\rho = \begin{bmatrix} 0{,}00 & 0{,}83 & 1{,}28 \\ 0{,}83 & 0{,}00 & 1{,}78 \\ 1{,}28 & 1{,}78 & 0{,}00 \end{bmatrix}; \quad d = \begin{bmatrix} 0{,}00 & 0{,}99 & 0{,}96 \\ 0{,}99 & 0{,}00 & 0{,}63 \\ 0{,}96 & 0{,}63 & 0{,}00 \end{bmatrix}$$

т.е. d_{xy}, в отличие от ρ_{xy}, предложило объединить в один кластер 1 и 3 ценные бумаги (как имеющие минимальное расстояние), тогда как в реальности можно видеть, что динамика цен на них практически противоположна.

Также необходимо выбрать метод кластеризации. Поскольку все плоские

алгоритмы кластеризации для своей работы требуют предварительно задать число кластеров k, то, несмотря на все преимущества, которые могла бы дать нечёткая кластеризация (поскольку ценная бумага может относиться к нескольким группам, данная тема требует широкого отдельного изучения), в настоящей работе будет использоваться иерархическая агломеративная кластеризация, для алгоритмов которой есть возможность задать максимальное значение расстояния (t — от англ. *thresold*), для которого возможно объединение объектов в кластеры. Для работы алгоритмов иерархической кластеризации необходимо также выбрать метод вычисления межкластерных расстояний. Чаще всего используются методы:

1. Ближнего соседа (одиночной связи, ближайшей точки, *single*)— за межкластерное расстояние принимается минимальное расстояние между объектами, принадлежащими кластерам (6).

$$d(u, v) = \min_{ij}(dist(u[i], v[j])) \tag{6}$$

   Данный метод хорошо работает, если требуется «быстрое» объединение объектов, при этом не имеет слишком большого значения, что в одном кластере могут оказаться достаточно непохожие объекты.
2. Дальнего соседа (полной связи, дальней точки, *complete*) — за межкластерное расстояние принимается максимальное расстояние между объектами, принадлежащими кластерам (7).

$$d(u, v) = \max_{ij}(dist(u[i], v[j])) \tag{7}$$

   Данный метод хорошо работает в случае, когда объединение требуется производить постепенно, «осторожно», а в один кластер не должны попасть непохожие объекты.
3. Средней связи (*average*) — за межкластерное расстояние принимается максимальное расстояние между объектами, принадлежащими кластерам (8).

$$d(u, v) = \sum_{ij} \frac{dist(u[i], v[j])}{|u| * |v|} \tag{8}$$

   где $|u|$, $|v|$ — число наблюдений в кластерах u и v соответственно. Результаты применения данного метода обычно представляют собой нечто среднее между результатами применения методов ближнего соседа и дальнего соседа.

Поскольку «на берегу» оценить, какой именно из методов окажется лучше для кластеризации ценных бумаг, представляется затруднительным, в дальнейшем проведём эксперименты и сделаем вывод на их основе. Тем не менее, можно

заранее отметить, что при использовании методов ближнего и дальнего соседа кластеризация будет проводиться быстрее, чем при использовании метода одиночной связи, за счёт возможности не пересчитывать на каждом шаге все межкластерные расстояния, что может оказаться полезным при ускорении составления оптимального портфеля.

Приведём получившийся алгоритм разделения ценных бумаг на кластеры (1).

Алгоритм 1. Разделение ценных бумаг на кластеры

Вход : матрица данных о стоимости ценных бумаг P (размерность $n \times m$), пороговое значение кластеризации t. Также задана функция вычисления межкластерного расстояния `InterclustDist`

Выход : массив меток кластеров marks (размерность n)

```
1  Function FormClusters(P, t)
2      clusters = {{1}, ..., {n}}
3      while True do
           /* найти два ближайших кластера                    */
4          U, V ← arg min_{U≠V} InterclustDist(U, V)
5          if CopheneticDistance(U, V) ⩽ t then
6              W ← U ∪ V
7              clusters = clusters ∪ W \ {U, V}
8          else
9              break
10         end
11     end
12     marks ← SetMarks(clusters)
13     return marks
14 end
15 Function CopheneticDistance(U, V)
16     return max_{u∈U, v∈V} InterclustDist(U, V)
17 end
18 Function SetMarks(clusters)
19     marks ← Z_n                          // массив из нулей размера n
20     foreach cluster_number in clusters do
21         foreach stock_number in clusters[cluster_number] do
22             marks[stock_number] ← cluster_number
23         end
24     end
25     return marks
26 end
```

Данный алгоритм позволяет разделять ценные бумаги на кластеры на основе данных об изменчивости их доходностей (причём, что важно, на основании

этих данных), не требует задавать требуемое число кластеров, а позволяет задавать некоторую предельную границу объединения, чтобы контролировать, насколько разные ценные бумаги могут оказаться в одном кластере. Таким образом, алгоритм удовлетворяет всем сформулированным ранее требованиям, необходимым для его использования для сокращения размерности в задаче составления оптимального портфеля ценных бумаг. В то же время, он, при минимальной модификации, может быть использован и в целом для снижения размерности в других задачах оптимизации.

2.3. Пересчёт цен и доходностей

На данном этапе алгоритма (строго говоря, объединяющем в себе сразу два в силу их сильной схожести) необходимо, во-первых, на основе матрицы P и распределении n ценных бумаг на k кластеров (вектор $marks$ из n целых чисел, каждое в диапазоне от 1 до k) получить массив P^c данных о динамике «цен» кластеров в среднем (как среднего динамики цен ценных бумаг, формирующих кластер), а, во-вторых, провести аналогичный пересчёт для «доходностей» кластеров (на основе Y и $marks$ получить вектор Y^c).

В качестве способа вычисления среднего значения будем использовать геометрическое среднее ((9)), так как и данные об изменении стоимости ценных бумаг, и данные о доходности представляют собой индексы, а именно среднее геометрическое позволяет получить усреднённую информацию по нескольким индексам.

$$G(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 * x_2 * \cdots * x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \tag{9}$$

Таким образом, значением «цены» кластера i в момент t будет (10)

$$p_{it}^c = \left(\prod_{j:\, marks[j]=i} p_{jt} \right)^{\frac{1}{|i|}} \tag{10}$$

где p_{jt} — цена j-ой ценной бумаги из кластера i в момент t, $|i|$ — число ценных бумаг в кластере i.

Аналогично, значение «доходности» кластера i можно вычислить по следующей формуле (11)

$$y_i^c = \left(\prod_{j:\, marks[j]=i} y_j \right)^{\frac{1}{|i|}} \tag{11}$$

где y_i^c — цена i-ой ценной бумаги из кластера i в момент t, $|i|$ — число ценных бумаг в кластере i.

Данная задача по сравнению с рассмотренной ранее кластеризацией является сугубо вычислительной, тем не менее, необходимо решать её безошибочно и достаточно быстро.

Используем для вычисления матрицы «цен» кластеров алгоритм 2, аналогично ему запишем алгоритм 3 для расчёта матрицы «доходностей» кластеров.

Как можно видеть, у алгоритмов 2 и 3 достаточно много общих операций, поэтому, при необходимости, их можно объединить в один 4.

Также с целью оптимизации можно не рассчитывать размеры кластеров

Алгоритм 2. Пересчёт динамики средней цены ценных бумаг для кластеров

Вход : массив данных о стоимости ценных бумаг P (размерность $n \times m$),
массив меток кластеров marks (размерность n),
число кластеров k

Выход : массив данных о «стоимости» кластеров PC (размерность $k \times m$)

```
1  Function RecalculateClustersPrices(P, marks)
2      PC ← 𝕁_{k,m}                    // массив из единиц размера k × m
3      sizes ← ℤ_k                      // массив из нулей размера k
4      for i = 1, ..., n do
5          sizes[marks[i]] ← sizes[marks[i]] + 1
6          for j = 1, ..., m do
7              PC[marks[i], j] ← PC[marks[i], j] · P[i, j]
8          end
9      end
10     for i = 1, ..., k do
11         for j = 1, ..., m do
12             PC ← PC[i, j]^{1/sizes[i]}
13         end
14     end
15     return PC
16 end
```

Алгоритм 3. Пересчёт доходности ценных бумаг для кластеров

Вход : массив данных о доходности ценных бумаг Y (размерность n),
массив меток кластеров marks (размерность n),
число кластеров k

Выход : массив данных о «доходности» кластеров YC (размерность k)

```
1  Function RecalculateClustersReturns(Y, marks)
2      YC ← 𝕁_k                        // массив из единиц размера k
3      sizes ← ℤ_k                      // массив из нулей размера k
4      for i = 1, ..., n do
5          sizes[marks[i]] ← sizes[marks[i]] + 1
6          YC[marks[i]] ← YC[marks[i]] · Y[i]
7      end
8      for i = 1, ..., k do
9          YC ← YC[i]^{1/sizes[i]}
10     end
11     return YC
12 end
```

Алгоритм 4. Пересчёт всех данных ценных бумаг для кластеров

Вход : массив данных о стоимости ценных бумаг P (размерность $n \times m$),
массив данных о доходности ценных бумаг Y (размерность n),
массив меток кластеров marks (размерность n),
число кластеров k

Выход : массив данных о «стоимости» кластеров PC (размерность $k \times m$),
массив данных о «доходности» кластеров YC (размерность k)

```
1  Function RecalculateClustersData(P, marks)
2      PC ← 𝕁_{k,m}
3      YC ← 𝕁_k
4      sizes ← ℤ_k
5      for i = 1, ..., n do
6          sizes[marks[i]] ← sizes[marks[i]] + 1
7          YC[marks[i]] ← YC[marks[i]] · Y[i]
8          for j = 1, ..., m do
9              PC[marks[i], j] ← PC[marks[i], j] · P[i, j]
10         end
11     end
12     for i = 1, ..., k do
13         YC ← YC[i]^{1/sizes[i]}
14         for j = 1, ..., m do
15             PC ← PC[i, j]^{1/sizes[i]}
16         end
17     end
18     return PC, YC
19 end
```

$sizes$, а использовать вместо них полученное в алгоритме 1 C, также содержащее информацию о том, какие акции входят в каждый из кластеров (и, следовательно, их размеры). Однако, такая модификация достаточно сильно снизит читаемость алгоритмов, поэтому ей стоит пользоваться с крайней осторожностью.

2.4. Получение долей кластеров

На данном этапе необходимо на основе данных о динамике цен P^c и доходностях Y^c получить такой вектор весов

$$W = [w_1 \dots w_k]$$

$$\sum_{i=1}^{k} w_i = 1$$

где w_i — доля i-ого кластера,
чтобы (в рассматриваемом примере с оптимальным портфелем заданной доходности), с одной стороны, выполнялось условие достижения минимальной доходности (12), а с другой — условие минимизации риска (13).

$$R = \sum_{i=1}^{k} Y_i^c W_i \leqslant R_{\min} \tag{12}$$

где Y_i^c — доходность i-ого кластера, W_i — его доля в составленном портфеле.

$$V = \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{i,j} W_i W_j \to \min \tag{13}$$

где $\sigma_{i,j}$ — ковариация «доходностей» кластеров i и j.

Для решения этой задачи существует множество алгоритмов, и появляются новые, поэтому будет правильнее с методологической точки зрения не «привязываться» к конкретному алгоритму, а, как и в большинстве работ на схожую тему, оставить этот этап «чёрным ящиком», о котором известны только его входные и выходные данные.

2.5. Расчёт итоговых долей ценных бумаг

На этом, последнем этапе алгоритма, необходимо на основе полученного вектора долей кластеров в оптимальном портфеле $W = [w_1 \dots w_k]$ и меток кластеров $marks = [m_1 \dots m_n]$ получить вектор

$$X = [x_1 \dots x_n]$$

$$\sum_{i=1}^{n} x_i = 1$$

где x_i — доля i-ой ценной бумаги в полученном оптимальном портфеле.

Соответственно, для каждого кластера i есть, с одной стороны, её доля в портфеле w_i, а с другой — s ценных бумаг, формирующих данный кластер, и необходимо «распределить» эту долю на эти s ценных бумаг. Безусловно, эту задачу можно решать с использованием какой-либо метаинформации о ценных бумагах (например, распределять долю инвестированных средств в зависимости от уровня капитализации компаний-эмитентов ценных бумаг), однако, это снизит возможность применять алгоритм для других задач оптимизации, так как для них придётся выбирать аналогичные правила, которые, к тому же, могут в принципе отсутствовать. Также нерациональным выглядит решение, применяющееся в некоторых работах — вкладывать всю долю средств, приходящихся на кластер, в одну, «наилучшую», ценную бумагу. Кроме того, что такой подход слишком сильно завязан на определение той самой «лучшей» ценной бумаги (а ведь её выбор не является тривиальным — его можно, например, осуществлять как на основе наивысшей доходности, так и на основе наименьшей вариации доходности), он ещё и сильно снижает саму по себе диверсифицированность портфеля.

Поэтому наилучшим решением представляется определять долю каждой ценной бумаги по формуле (14):

$$x_i = \frac{W_j}{S_j} \tag{14}$$

где
j — кластер, в который входит ценная бумага i,
S_j — число ценных бумаг, формирующих кластер j.

Данная задача также является сугубо вычислительной, алгоритм для её решения 5 во многом схож с алгоритмами для пересчёта цен и доходностей для кластеров (что объясняется схожей природой этих задач, так как, по сути, они являются обратными друг другу).

Приведённый алгоритм также, как и алгоритмы для пересчёта данных для кластеров, может быть оптимизирован за счёт использования ранее рассчитанных

Алгоритм 5. Расчёт индивидуальных весов кластеров

Вход : вектор долей кластеров в портфеле W (размерность k),
массив меток кластеров marks (размерность n),
число кластеров k

Выход : вектор индивидуальных долей каждой акции в портфеле X
(размерность n)

```
1  Function CalculateIndividualWeights(W, marks)
2      sizes ← ℤ_k                          // массив из нулей размера k
3      X ← ℤ_n                              // массив из нулей размера n
4      for i = 1, ..., n do
5          sizes[marks[i]] ← sizes[marks[i]] + 1
6      end
7      for i = 1, ..., n do
8          X ← W[marks[i]]/sizes[i]
9      end
10     return X
11 end
```

объёмов кластеров $sizes$ или полученных в алгоритме (1) C, однако, поскольку эти модификации снизят читаемость алгоритма, здесь они не приводятся. Стоит также отметить, что, несмотря на то, что в данном алгоритме дважды выполняется проход по $i = 1, \ldots, n$, объединить их в один невозможно, так как только после завершения первого прохода будут известные точные размеры каждого кластера и будет возможным выполнять расчёт доли каждой ценной бумаги (однако, упомянутая выше оптимизация с использованием рассчитанного ранее $sizes$ полностью снимает эту проблему).

Необходимо упомянуть одну важную особенность описанного выше алгоритма. Как можно получить из (14), рассчитанное на шаге непосредственной оптимизации портфеля значение доходности R останется неизменным и будучи рассчитанным для индивидуальных весов (стоит отметить, что если бы R изменялось, то это ставило бы под сомнение всю целесообразность снижения размерности описываемым способом). В то же время, точное значение меры риска V изменяется (что является причиной упомянутого в первой главе ухудшения качества портфеля при снижении размерности). Не углубляясь в рамках данной, во многом разведочной работы, в этот вопрос, тем не менее, отметим, что он требует дополнительного изучения. Отметим только, что, чем на меньшее число кластеров будут в итоге разделены n ценных бумаг, тем более сильным будет снижение размерности (иначе говоря, чем выше будет t), тем сильнее ухудшится качество полученного портфеля. Данное утверждение имеет следующее экономическое обоснование: чем с более обобщённым, «грубым» делением будет формироваться портфель, тем меньше у него будет возможностей для диверсификации (которая, в первую очередь, и помогает снизить риск портфеля).

Как следствие, при применении описанного алгоритма для понижения размерности следует с осторожностью и предварительной проверкой оптимизировать портфель по критериям, связанным с риском V. Стоит также отметить, что определённые «проблемы» с оптимизацией портфеля по заданному риску известны достаточно давно, и они привели, например, к появлению критерия оптимизации $RAPOC$.

3. Применение предлагаемого алгоритма

Для проверки полезности предложенного метода необходимо проверить его работу на практике. Следовательно, необходимо:

во-первых, реализовать описанный выше алгоритм

во-вторых, изучить, как влияет применение алгоритма на составление инвестиционного портфеля

3.1. Программная реализация предлагаемого алгоритма

Для реализации выбран язык программирования *Python* так как он, во-первых, достаточно популярен для задач обработки данных (как следствие, для большинства задач обработки данных существуют готовые, многократно проверенные решения для *Python*), а, во-вторых, используется во многих работах схожей тематики. Недостатком *Python* для реализации предложенного алгоритма является его интерпретируемость, что приводит к, иногда, достаточно сильному снижению скорости работы по сравнению с компилируемыми языками. Особенно проблемной и замедляющей является работа с циклами, особенно вложенными.

Для решения этой проблемы можно, с одной стороны, использовать вместо де-факто «стандартного» интерпретатора *CPython*, например, *PyPy*, поддерживающий JIT-компиляцию, за счёт чего достигается достаточно сильное ускорение работы программы. Проблемой такого пути будет то, что большинство сторонних библиотек для *Python* ориентированы, в первую очередь, на работу с *CPython*, и, как следствие, не имеют полной совместимости с *PyPy* (равно как и с другими сторонними интерпретаторами). Например, периодически возникают проблемы совместимости *PyPy* с крайне популярной библиотекой для работы с массивами *NumPy*. Конечно, и в случае несовместимости можно найти выход, однако, при нацеленности реализации в первую очередь на проверку научной гипотезы (и, как следствие, заинтересованности в том, чтобы максимально широко использовать готовые решения) использование *PyPy* вместо *CPython* следует признать нецелесообразным. Вместе с тем, при запуске алгоритма в промышленное использование (даже как части более крупной системы) стоит задуматься о том, чтобы переписать его полностью на *PyPy*, или, в самом радикальном варианте, на компилируемом языке программирования, например, *C* (тем более, что модуль, написанный на *C*, можно после небольшого «наведения мостов» использовать вместе с *CPython*).

Другим решением будет использовать интерпретатор *CPython* и максимально широко задействовать библиотеки для него (такие, как *NumPy*), содержащие в себе оптимизированный для ускорения работы код (чаще всего — написанный на

компилируемом языке, например, уже упоминавшемся выше *C*). В то же время, в случае, если для какой-либо из требуемых трудоёмких задач не найдётся уже готового достаточно быстро работающего решения, то создадим его сами с использованием *Cython* — средства, позволяющего транслировать код на *Python*-подобном языке высокого уровня (отличающемся от *Python*, в первую очередь, статической типизацией) в код модуля для *Python* на *C*. Такой подход очень популярен среди разработчиков библиотек для *CPython*, например, им активно пользуются авторы популярной библиотеки *scipy*. В итоге получится решение, в котором всю трудоёмкую работу выполняют скомпилированные модули, а интерпретируемый код *Python* используется, в первую очередь, для обмена информацией между этими подключаемыми модулями. Такая схема получила название «*Python* как клей» (англ. «*Python* as glue»), и является достаточно популярной, особенно для научной и околонаучной разработки, так как позволяет максимально широко использовать реализованные ранее (и, как правило, многократно проверенные и оптимизированные) алгоритмы, причём реализация могла быть получена ещё десятилетия назад, например, на *Fortran*, не теряя при этом в читаемости и выразительности кода.

Именно это решение и использовано при реализации предложенного алгоритма, однако, для сравнения примерной производительности те элементы, для реализации которых использовался *Cython*, были также реализованы без него в ещё двух вариантах — с использованием связки *Python* + *NumPy* и с использованием сторонней библиотеки *numpy-indexed*, написанной на «чистом» *Python* и добавляющей некоторые полезные функции для работы с массивами.

Кроме *Cython*, также использовались следующие сторонние библиотеки:

1. *NumPy* — библиотека, предназначенная для быстрой и эффективной работы с массивами (в том числе многомерными) и различными способами их обработки, предоставляет гораздо более широкие возможности, чем модуль стандартной библиотеки *Python* `array`. «Ядро» `NumPy` написано на *C*, что делает его очень эффективным и стало причиной того, что практически любая другая библиотека, работающая с данными, будет использовать *NumPy*.
2. *SciPy* — библиотека, предоставляющаяя широкий функционал для научных и инженерных расчётов, в том числе модуль `scipy.cluster.hierarchy`, позволяющий осуществлять очень широкий спектр расчётов для проведения кластеризации [19]. *SciPy* использует массивы *NumPy* для обработки данных, а многие из её модулей написаны на *Cython*, что позволяет получать достаточно высокую производительность. Кроме *SciPy*, для задач кластеризации (как и машинного обучения в целом) также очень популярна библиотека *scikit-learn*, однако, её модуль `AgglomerativeClustering` представляет собой, в первую очередь, «обёртку» для `scipy.cluster.hierarchy`. Таким образом, использование *scikit-learn* представляется оправданным в

программах, работающих с широким спектром методов машинного обучения (практически все они представлены в *scikit-learn*), однако, для реализации алгоритма, практически полностью связанного с кластеризацией, и только с ней, использование `scipy` предпочтительнее, так как, во-первых, позволит, при необходимости, проводить достаточно тонкую настройку кластеризации, не ограничивая возможности API посредника, а, во-вторых, избавит проект от лишних зависимостей.

3. *CVXPY* — библиотека, встраивающая в *Python* язык для определения оптимизационных задач [17, 18]. В соответствии с упомянутой ранее концепцией непосредственно оптимизации как «чёрного ящика» упомянем только, что *CVXPY* поддерживает достаточно широкий спектр оптимизаторов (которые, чаще всего, представляют собой скомпилированные модули на *C*). В *CVXPY* и входные, и выходные данные в задаче оптимизации представляются в виде массивов *NumPy*. Как следствие того, что оптимизация в предлагаемом алгоритме является «чёрным ящиком», *CVXPY* можно безболезненно заменить на другую библиотеку, предоставляющую функционал для решения задачи оптимизации.

Полный код реализации предлагаемого решения приведён в приложении А. Кроме самого описанного алгоритма, в нём реализованы также некоторые другие связанные с портфельной оптимизацией функции.

Функции с одинаковыми задачами — `cythonized.recalculate_data()`, `python_numpy.recalculate_data()` и `python_npi.recalculate_data()`, а также `cythonized.recalculate_returns()`, `python_numpy.recalculate_returns()` и `python_npi.recalculate_returns()` приведены для иллюстрации того, как одна и та же задача пересчёта данных для кластеров может быть решена средствами «чистого» *Python* (а используемый модуль *numpy-indexed*, несмотря на название, и то, что работает с *NumPy*-массивами, реализован средствами «чистого» *Python*)и с использованием *Cython*.

Полученная реализация, несмотря на стремление сделать её максимально эффективной, конечно же, не может считаться полностью оптимальной, в частности, в ней для сохранения чистоты и читаемости кода не внедрены оптимизации, упомянутые в описаниях алгоритмов во второй главе; также, строго говоря, можно вообще убрать все вставки-«мостики» на *Python* и реализовать модуль полностью на *C* или *Cython*, однако, данная реализация уже достаточно близка к оптимальной.

Собранный на основе данной реализации пакет доступен для установки через *PyPI*:
`https://pypi.org/project/clustering-optimization-speedup/`.

3.2. Влияние предлагаемого алгоритма на составление инвестиционного портфеля

3.2.1. Методология изучения влияния алгоритма на составление инвестиционного портфеля

Для того, чтобы исследовать то, как предложенный метод сокращения размерности повлияет на составление оптимального портфеля, необходимо ещё раз вернуться к вопросу о том, влияние какого характера от него ожидается и разобраться с тем, как его можно оценить.

1. Сокращение размерности позволит ускорить процесс составления инвестиционного портфеля (за счёт того, что оптимизационная задача будет решаться для меньшего числа исходных данных). Главной оценкой этого ускорения будет, конечно же, непосредственное сравнение времени работы программы для «обычного» составления оптимального инвестиционного портфеля и времени работы программы, использующей алгоритм сокращения размерности. Однако, такая оценка будет необъективной. С одной стороны, причина лежит в особенностях программной реализации, которая, хоть и, как упоминалось ранее, близка к оптимальной, но, скорее всего, не является такой. С другой — из-за того, что любое точное сравнение времени будет зависеть от непосредственного оптимизатора (который специально оставлен «чёрным ящиком»), и нет гарантии, что при замене *CVXPY* на *scipy.optimize* (которая может произойти по совершенно «природным» в смысле их неизбежности причинам, например, прекращению поддержки одной из библиотек её авторами) это значение останется прежним. Вносят свою лепту и конкретные используемые алгоритмы оптимизации, а то, как поведёт себя это соотношение при реализации решения на другом языке программирования (что, теоретически, может потребоваться для практического применения), предсказать вообще практически невозможно. Третьим важным фактором (пожалуй, наиболее важным в краткосрочной перспективе) является то, что на разных компьютерах соотношения времени работы будут разными — и эффект от применения метода при работе на достаточно новом настольном компьютере может сильно отличаться от эффекта для не слишком нового ноутбука, тогда как постановка задачи с экономической точки зрения требует проверить полезность применения метода на устройствах разных ценовых категорий (и, следовательно, мощностей).
   Следовательно, необходимо, кроме непосредственного исследования времени работы (причём, желательно, на отличающихся компьютерах) оценить также некоторый объективный показатель, который покажет, как сильно это время, в принципе, может сократиться при применении предложенного

метода сокращения размерности. В качестве такого показателя предложим коэффициент снижения размерности E (от англ. *economy*):

$$E = \frac{n}{k}$$

где
n — число ценных бумаг, которые доступны для инвестирования (иначе говоря, размерность исходной задачи),
k — число выделенных кластеров (иначе говоря, размерность задачи, которую пришлось решать фактически).
E показывает, насколько сильно снизилась размерность задачи оптимизации. $0 < E <= 1$, и чем ближе E к 0, тем сильнее снизилась размерность, и, следовательно, сильнее ускорилось составление оптимального портфеля. Логично, что E будет зависеть от выбранных метода кластеризации (вычисления межкластерных расстояний) и порогового значения кластеризации t

2. Сокращение размерности приведёт к тому, что полученный портфель будет по своим характеристикам хуже, чем портфель, составленный без сокращения размерности (более подробное обоснование приведено во 2 главе данной работы), в частности, при составлении портфелей для одинаковой ожидаемой доходности R мера риска V_c «кластеризованного» портфеля будет выше, чем мера риска V_u «некластеризованного» портфеля.
Поскольку непосредственное сравнение V_c и V_u не даст необходимой информации (в первую очередь — по причине безразмерности V), для оценки экономического эффекта используем следующую технику: рассчитаем для заданной доходности R_{ic} «кластеризованный» оптимальный портфель, для него получим меру риска V_{ic}. Затем для того же набора ценных бумаг рассчитаем «некластеризованный» оптимальный портфель по заданному риску V_{ic}, для него получим доходность R_{uc}. Введём величину L (от англ. *loss*), характеризующую снижение доходности:

$$L = R_{iu} - R_{ic}$$

L зависит от R_{ic}, метода кластеризации и t, так как, с одной стороны, чем выше требуемая доходность, тем сложнее составить инвестиционный портфель, а, с другой, чем сильнее (и, в общем-то, «грубее») проведена группировка, тем хуже окажется «кластеризованный» портфель по сравнению с «некластеризованным».
Экономически можно трактовать L как недополученную прибыль инвестора от применения алгоритма снижения размерности, например, если при $R_{ic} = 1{,}2$ $L = 0{,}03$, то можно считать, что без применения метода

инвестор получил бы прибыль на вложенные средства 23%, а в результате применения метода снижения размерности — только 20%.
Часть полученных в данном разделе результатов была ранее опубликована автором в [16].

Таким образом, для проверки влияния предложенного алгоритма на составление оптимального портфеля необходимо проанализировать E, L и непосредственные скорости работы программы для различных значений t, R и различных методов кластеризации (ближнего соседа, дальнего соседа, средней связи). Логично, что снижение размерности будет тем полезнее, чем выше E и ниже L, однако, поскольку они обратно зависимы, то необходимо изучить их соотношение. Также будет необходимо изучить связь E и реального изменения скорости составления «кластеризованного» портфеля по сравнению с «некластеризованным». Примем пока что, на уровне рабочей гипотезы, что допустимым считается $L <= 0{,}05$.

Эксперименты, для повышения объективности, проводились с использованием двух компьютеров:

1. «Современный» — настольный компьютер 2019 года, со следующими основными характеристиками:
   - Процессор Intel Core i5-9400, 6x2,90 ГГц
   - 15,6 ГГб ОЗУ
   - ОС linux, ядро версии 5.6.15-1-MANJARO
   - Python 3.8.3 [GCC 10.1.0]
2. «Устаревший» — ноутбук 2014 года HP Pavillion 15, основные характеристики которого также приведены ниже:
   - Процессор AMD A4-5000, 4x1,50 ГГц
   - 3,4 ГГб ОЗУ
   - ОС linux, ядро версии 5.6.15-arch1-1
   - Python 3.8.3 [GCC 10.1.0]

Также, для обеспечения снижения влияния случайных факторов на результаты экспериментов, все тестовые случаи запускались несколько раз, и бралось среднее значение для всех этих запусков.

3.2.2. Набор данных для проведения экспериментов

Поскольку, как было показано ранее, реальная полезность метода зависит от E и L, которые, в свою очередь, зависят не только от t и метода кластеризации, но и от особенностей исходных данных — в биржевой ситуации, при которой ценные бумаги можно достаточно хорошо разделить на группы, метод будет эффективнее, чем в биржевой ситуации, при которой доходности всех ценных бумаг изменяются приблизительно одинаково (например, если рынок вошёл в период рецессии).

Следовательно, нужно провести эксперименты на нескольких наборах данных, желательно, отличающихся по своей структуре.

В качестве таких наборов данных используем:

1. Акции компаний из рейтинга Standard & Poor's 500 за 2014–2016 г.г. (данные об акциях 480, 486 и 494 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 1 (все данные нормированы).

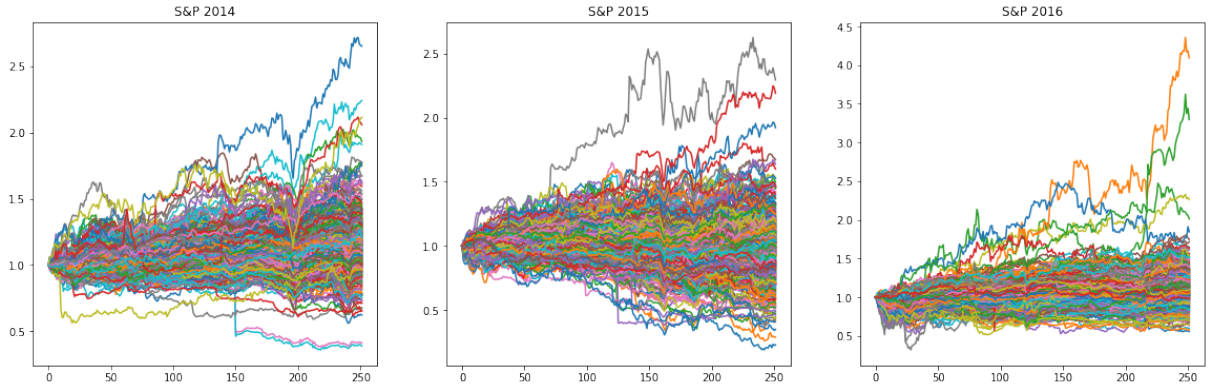

Рис. 1 — Изменение цен на акции компаний из рейтинга S&P 500 в 2014–2016 годах

Как можно видеть, набор данных выглядит достаточно сбалансированным, в нём нет каких-либо резких, шоковых колебаний рынка или компаний, цены на акции которых вели бы себя странно — в этом и проявляется особенность работы с акциями компаний из рейтинга S&P 500 — все они достаточно респектабельны и относительно стабильны, но, как следствие этого, на акциях из этого набора невозможна и какая-либо сверхприбыль. В то же время, стоит отметить, что выбранный период большинство исследователей называют временем достаточно стабильной макроэкономической ситуации, что важно для проверки предложенного метода, так как в случае, когда макроэкономическая ситуация нестабильна, как известно, прогнозирование в принципе всегда затрудняется. Также на рисунке 2 приведена зависимость числа выделяемых кластеров от метода и порога кластеризации.

Кластеры выделяются достаточно хорошо — при использовании метода одиночной связи быстро (возможно, даже слишком быстро), при использовании методов полной связи и средней связи медленнее, при этом «наилучшим» предварительно выглядит метод средней связи, так как при его использовании число выделяемых кластеров будет проще всего контролировать (а, следовательно, контролировать E и L). Тем не менее, то, что скорости выделения кластеров примерно соответствуют ожидаемым, является хорошим признаком того, что способ вычисления расстояния между наблюдениями выбран верно.

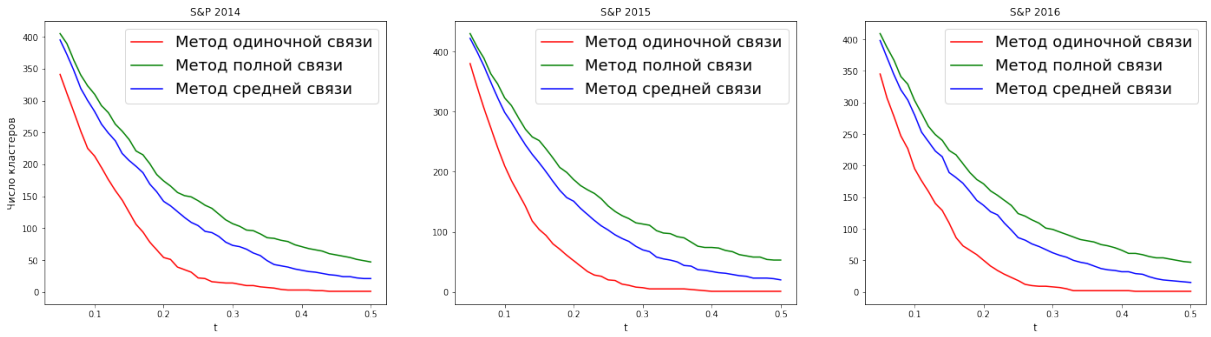

Рис. 2 — Число кластеров в зависимости от порога кластеризации t и метода кластеризации для акций компаний из рейтинга S&P 500 в 2014–2016 годах.

2. Акции компаний, торгуемые на Нью-Йоркской фондовой бирже (NYSE) в 2004–2006 г.г. (данные об акциях 1285, 1354 и 1421 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 3 (все данные нормированы, даны в логарифмической шкале) Этот набор данных

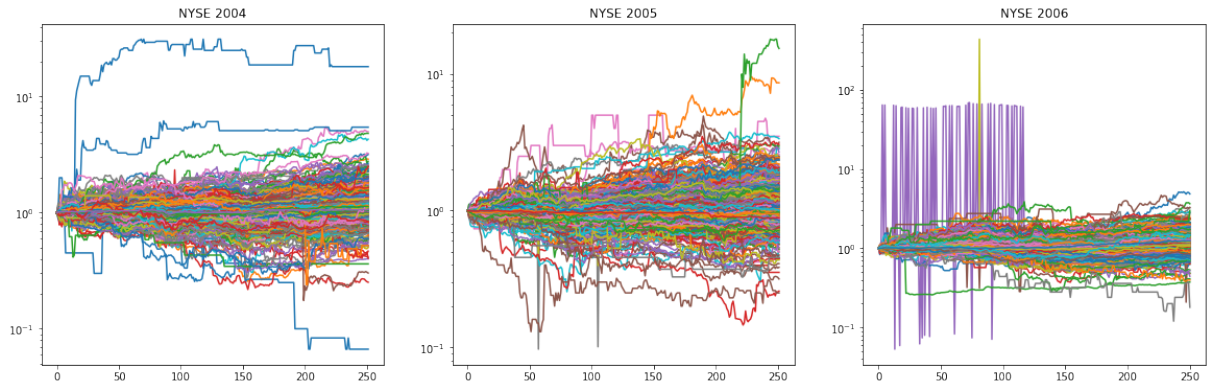

Рис. 3 — Изменение цен на акции компаний, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах.

является гораздо более разнообразным, чем предыдущий — в нём есть и акции, цены на которые совершали скачки на порядок, и акции, цены на которые порой вели себя совершенно нестабильно, и «обыкновенные», достаточно стабильные акции. Такие сложные изменения привели к тому, что график потребовалось приводить с логарифмической шкалой для индексов цен, так как обычная не позволила бы продемонстрировать все варианты изменчивости. В то же время, можно предполагать, что на таких интересных данных и метод понижения размерности с помощью кластеризации покажет интересные результаты — если бы работу по группировке схожих акций выполнял человек, то она бы точно пошла на пользу (например, через отделение стабильных акций от нестабильных). Аналогично первому набору данных, на рисунке 4 приведена зависимость числа выделяемых кластеров от метода и порога кластеризации.

Ситуация с зависимостью числа выделенных кластеров от параметров кла-

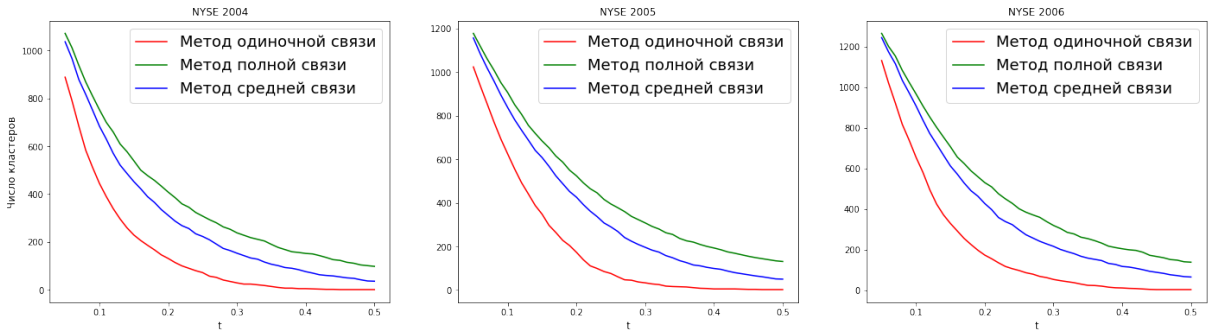

Рис. 4 — Число кластеров в зависимости от порога кластеризации t и метода кластеризации для акций компаний, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах.

стеризации принципиально не отличается от такой же для набора данных с ценами акций компаний из рейтинга S&P, что, опять же, является хорошим признаком с точки зрения правильности кластеризации.

3. Акции компаний, торгуемых на бирже NASDAQ в 2004–2006 г.г. (данные об акциях 1143, 1207 и 1280 компаний соответственно). Данные об изменениях цен на акции приведены на рисунке 5 (все данные нормированы, даны в логарифмической шкале). Для этого набора данных ситуация выглядит

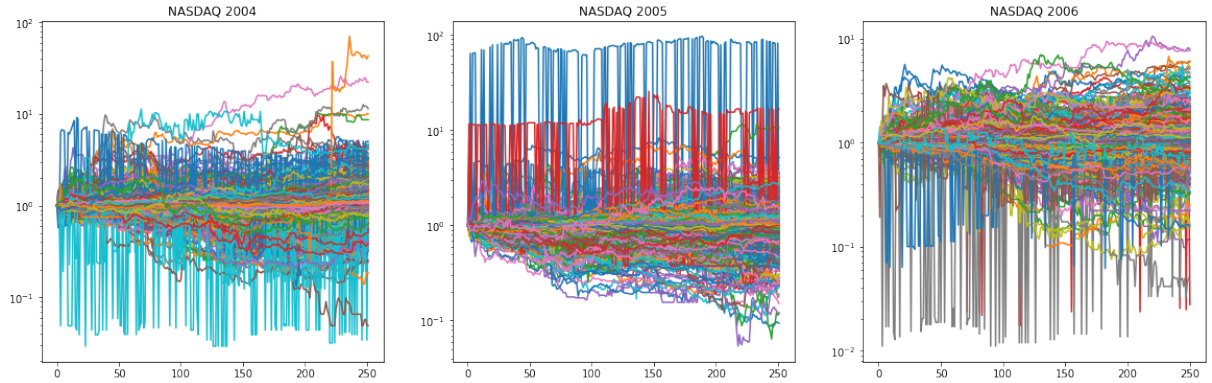

Рис. 5 — Изменение цен на акции компаний, торгуемых на бирже NASDAQ в 2004–2006 годах.

ещё интереснее, чем для предыдущего — в первую очередь, в силу самой специфики биржи NASDAQ, которая считается «высокотехнологичной». Как следствие, очень сильные колебания цен на многие акции, торгуемые на ней, обуславливаются, с одной стороны, динамичным характером самой отрасли, а с другой — высокой долей спекуляций на них. В то же время, предположительно, на этом наборе данных метод может показать результат хуже, чем на предыдущих, так как большинство компаний, чьи акции торгуются на NASDAQ, принадлежат к отрасли высоких технологий, а значит, и спрос на их ценные бумаги будет уже сам по себе связан, что обусловит их схожую изменчивость и затруднит качественное выделение кластеров. Аналогично двум предыдущим наборам данных, на рисунке 6 приведена за-

висимость числа выделяемых кластеров от метода и порога кластеризации.

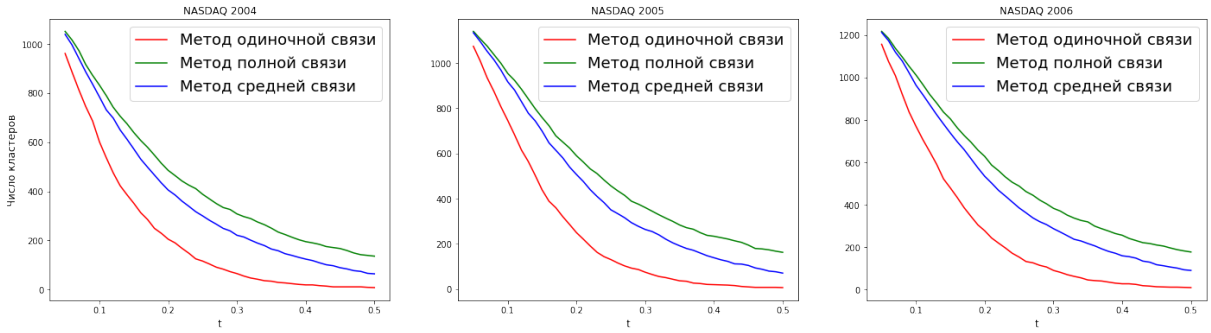

Рис. 6 — Число кластеров в зависимости от порога кластеризации t и метода кластеризации для акций компаний, торгуемых на бирже NASDAQ в 2004–2006 годах.

Как и предполагалось теоретически, для данного набора выделение кластеров происходит немного медленнее, чем для «разносторонних» первого и второго наборов. Однако, это может иметь положительный эффект за счёт упрощения контроля числа выделяемых кластеров. Тем не менее, кроме числа выделенных кластеров важно также их качество и итоговое влияние кластеризации на составление оптимального портфеля, и этот вопрос, к рассмотрению которого мы переходим, требует гораздо более пристального рассмотрения.

3.2.3. Влияние предлагаемого алгоритма при кластеризации по методу одиночной связи

При кластеризации по методу одиночного соседа кластеры во всех трех наборах данных выделяются примерно одинаково, быстро и «гладко». В акциях компаний из рейтинга S&P кластеры выделяются при меньших пороговых значениях, чем в акциях компаний Нью-Йоркской фондовой биржи и акциях компаний биржи NASDAQ, что обусловлено, во-первых, меньшим объёмом данных, во-вторых, тем, что в рейтинг S&P попадают, в первую очередь, крупные компании, цены на акции которых достаточно стабильны и не показывают как значительного, «взрывного» роста, так и сильного снижения.

Средние результаты применения предлагаемого метода приведены в приложении Б («–» означает, что построить кластеризованный оптимальный портфель с заданным параметром доходности не удалось).

Первый же вывод, который можно сделать на основе полученных результатов — предложенный алгоритм понижения размерности действительно помогает ускорить составление оптимального портфеля. Вторым важным выводом является то, что реализация алгоритма с использованием *Cython* является наилучшей среди всех представленных, и именно её следует рассматривать как основную.

При использовании других реализаций выгода от использования алгоритма полностью или частично «теряется». В то же время, совершенно верной оказалась исходная гипотеза о необходимости использовать для оценки влияния некие дополнительные показатели, кроме времени работы программы. Зависимость между временем работы и исходными параметрами, безусловно, есть, но она настолько сильно зависит от каких-то не поддающихся прямому анализу (во всяком случае, в рамках данной работы) факторов, что совершенно необходимо использовать обобщающий показатель E.

Также, как можно видеть, кроме очевидных зависимостей E от t и времени работы от t, существует ещё несколько зависимостей:

- L возрастает с ростом E при постоянном R_c.
- L, в целом, возрастает с ростом R_c при постоянном E.

Кроме того, можно отметить, что при меньшем объёме данных (акции компаний из рейтинга S&P 500) E растёт с ростом t быстрее, чем при большем объёме данных, и при некоторых значениях t и R_{ic} оптимальный портфель с заданными параметрами после кластеризации построить не удалось.

Поскольку важным свойством практического применения любой техники является предсказуемость результата, попробуем построить модели E от t и L от R_c и t. Также в моделях учтём фактор исходной размерности задачи n. Лучшими (и не содержащими при этом незначимых коэффициентов при уровне значимости 0,95) оказались следующие модели:

1. Для всех наборов данных в целом:

$$E = 372{,}2t - 2420t^2 + 5800t^3 - 16{,}2\frac{n}{1000}$$

для данной модели:

$$R^2 = 0{,}68;\ \ F\text{-значение} = 175{,}0;\ \ \text{число степеней свободы } df = 314$$

t-значения: для t = 3,83; для t^2 = 4,10; для t^3 = 5,84; для n = 3,39

и

$$L = -0{,}341 + 3{,}75t^3 + 0{,}33R_c - 0{,}033\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}69;\ \ F\text{-значение} = 238{,}0;\ \ df = 314$$

t-значения: для постоянного члена = 8,40; для t^3 = 25,7; для R_c = 9,86

2. Для данных об акциях, торгуемых на Нью-Йоркской фондовой бирже в

2004–2006 годах:

$$E = 131{,}7 + 1636t - 9084t^2 + 16703t^3 - 164\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}88;\ \ F\text{-значение} = 218{,}0;\ \ df = 110$$

t-значения: для постоянного члена = 2,53; для t = 3,74; для t^2 = 4,77; для t^3 = 6,49; для n = 5,19

и

$$L = -0{,}47 + 4{,}07t^3 + 0{,}41R_c$$

для данной модели

$$R^2 = 0{,}69;\ \ F\text{-значение} = 127{,}9;\ \ df = 112$$

t-значения: для постоянного члена = 6,59; для t^3 = 15,13; для R_c = 6,81

3. Для данных об акциях, торгуемых на бирже NASDAQ в 2004–2006 годах:

$$E = 18{,}3 + 267{,}2t - 1536{,}2t^2 + 3248{,}3t^3 - 26\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}99;\ \ F\text{-значение} = 2356{,}3;\ \ df = 121$$

t-значения: для постоянного члена = 3,46; для t = 5,59; для t^2 = 7,44; для t^3 = 11,87; для n = 7,64

и

$$L = -1{,}91t^2 + 8{,}26t^3 + 0{,}31R_c - 0{,}273\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}87;\ \ F\text{-значение} = 221{,}4;\ \ df = 122$$

t-значения: для t^2 = 3,34; для t^3 = 6,13; для R_c = 7,65; для n = 6,93

4. Для данных об акциях компаний из рейтинга S&P в 2014–2016 годах:

$$E = -898{,}4 - 1306{,}7t^2 + 6420{,}5t^3 - 1{,}87n$$

для данной модели

$$R^2 = 0{,}84;\ \ F\text{-значение} = 136{,}7;\ \ df = 73$$

t-значения: для постоянного члена = 3,97; для t^2 = 3,30; для t^3 = 6,49; для n = 4,02

и

$$L = -1{,}65 - 2{,}65t^2 - 3{,}45t^3 + 0{,}213R_c - 2{,}88\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}67;\ \ F\text{-значение} = 39{,}9;\ \ df = 72$$

t-значения: для постоянного члена = 4,53; для t^2 = 4,44; для t^3 = 2,31; для R_c = 4,03; для n = 4,05

Таким образом, общее резюме по использованию метода одиночной связи следующее — несмотря на возможность получения хороших результатов и высокую скорость объединения в кластеры, практическое применение затрудняется сложностью качественного прогнозирования его эффективности («подстроить» алгоритм под набор данных возможно, но сложно — а ещё качество этой подстройки будет зависеть от особенностей самих данных). Кроме того, сам характер связи может, в отдельных случаях, нарушаться (точнее, не подтверждаться эмпирически из-за того, что одна связь «перетянула» на себя влияние другой) — как, например, в случае с отрицательным коэффициентом при t^2 и t^3 (причём одновременно) в модели показателя снижения качества портфеля L для набора данных акций компаний из рейтинга S&P 500. Это обусловлено высокой вероятностью попадания в один кластер ценных бумаг с противоположной изменчивостью.

3.2.4. Влияние предлагаемого алгоритма при кластеризации по методу полной связи

Как можно видеть, объединение в кластеры при использовании метода полной связи происходит достаточно неравномерно и медленнее, чем при кластеризации по методу одиночной связи. В то же время, на всех трёх наборах данных кластеры выделяются достаточно хорошо, а для данных об акциях компаний Нью-Йоркской фондовой биржи и биржи NASDAQ — ещё и достаточно стабильно.

Поскольку объединение в кластеры при использовании метода полной связи происходит медленнее, чем при кластеризации по методу одиночной связи, в целом рост E и L при возрастании t и R_{ic} происходит медленнее, чем при использовании метода одиночной связи.

В то же время, эта «медленность» выделения кластеров приводит к достаточно хорошей предсказуемости результатов применения алгоритма, в частности, можно построить следующие модели (приведены лучшие модели, не имеющие незначимых переменных при уровне значимости 0,95):

1. Для всех наборов данных в целом:

$$E = 1{,}49 + 33{,}9t^2 - 0{,}31\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}88;\ \ F\text{-значение} = 1498{,}8;\ \ df = 375$$

t-значения: для постоянного члена = 14,93; для t^2 = 54,6; для n = 3,71

и

$$L = -0{,}151 + 0{,}128t + 0{,}291t^3 + 0{,}133R_c - 0{,}0095\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}63;\ \ F\text{-значение} = 161{,}2;\ \ df = 373$$

t-значения: для постоянного члена = 11,55; для t = 4,41; для t^3 = 2,17; для R_c = 12,93; для n = 4,16

2. Для данных об акциях, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах:

$$E = 8{,}81 + 39{,}2t^2 - 5{,}6\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}98;\ \ F\text{-значение} = 3465;\ \ df = 123$$

t-значения: для постоянного члена = 14,90; для t^2 = 82,25; для n = 12,87

и

$$L = 0{,}14 + 0{,}21t + 0{,}19R_c - 0{,}28\frac{n}{1000}$$

для данной модели:

$$R^2 = 0{,}75;\ \ F\text{-значение} = 125{,}2;\ \ 122$$

t-значения: для постоянного члена = 3,151; для t = 13,26; для R_c = 10,17; для n = 9,82

3. Для данных об акциях, торгуемых на бирже NASDAQ в 2004–2006 годах:

$$E = 4{,}02 + 6{,}31t + 34{,}28t^3 - 2{,}8\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}99;\ \ F\text{-значение} = 4915{,}0;\ \ df = 122$$

t-значения: для постоянного члена = 16,13; для t = 17,09; для t^3 = 19,97; для n = 14,02

и

$$L = -0{,}121t + 0{,}565t^2 + 0{,}062R_c - 0{,}005\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}91;\ \ F\text{-значение} = 325{,}3;\ \ df = 122$$

t-значения: для t = 2,33; для t^2 = 5,48; для R_c = 7,21; для n = 5,30

4. Для данных об акциях компаний из рейтинга S&P в 2014–2016 годах:

$$E = -13{,}365 + 44{,}213t^2 - 19{,}925t^3 + 0{,}03n$$

для данной модели

$$R^2 = 0{,}99;\ \ F\text{-значение} = 3644{,}0;\ \ df = 122$$

t-значения: для постоянного члена = 9,00; для t^2 = 17,79; для t^3 = 3,40; для n = 9,77

и

$$L = -0{,}484 + 0{,}375t^2 + 0{,}137R_c - 0{,}689\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}826;\ \ F\text{-значение} = 198{,}8;\ \ df = 122$$

t-значения: для постоянного члена = 6,08; для t^2 = 20,50; для R_c = 12,57; для n = 4,26

Таким образом, несмотря на то, что применение алгоритма с кластеризацией по методу полной связи далеко не всегда приведёт к очень сильному ускорению процесса оптимизации, она также позволяет получить достаточно хорошие, а главное — предсказуемые результаты.

3.2.5. Влияние предлагаемого алгоритма при кластеризации по методу средней связи

При кластеризации по методу средней связи объединение в кластеры происходит немного более быстро и «гладко», чем при использовании метода полной связи, но всё же не так быстро, как при использовании метода одиночной связи. Результаты, во многом, обусловлены скоростью объединения акций в кластеры — средней между методами одиночной и полной связей.

И, скорее всего, этим же «средним» положением метода средней связи обусловлена моделируемость E и L — можно построить модели, но их качество будет хуже, чем у аналогичных моделей для метода полной связи:

1. Для всех наборов данных в целом:

$$E = 2{,}72 + 169t^3 - 1{,}008\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}86;\ \ F\text{-значение} = 1128{,}0;\ \ df = 369$$

t-значения: для постоянного члена = 11,78; для t^3 = 47,37; для n = 5,03

и

$$L = -0{,}208 + 1{,}516t^3 + 0{,}186R_c$$

для данной модели

$$R^2 = 0{,}32;\ \ F\text{-значение} = 89{,}49;\ \ df = 369$$

t-значения: для постоянного члена = 5,61; для t^3 = 12,18; для R_c = 5,95

2. Для данных об акциях, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах:

$$E = 19{,}25 + 196{,}60t^3 - 12{,}87\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}97;\ \ F\text{-значение} = 2406{,}0;\ \ df = 123$$

t-значения: для постоянного члена = 12,75; для t^3 = 68,40; для n = 11,55

и

$$L = 4{,}595t - 21{,}00t^2 + 31{,}04t^3 + 0{,}324R_c - 0{,}495\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}61;\ \ F\text{-значение} = 40{,}47;\ \ df = 121$$

t-значения: для t = 3,51; для t^2 = 3,68; для t^3 = 4,06; для R_c = 4,99; для n = 6,68

3. Для данных об акциях, торгуемых на бирже NASDAQ в 2004–2006 годах:

$$E = 6{,}02 + 5{,}15t + 90{,}4t^3 - 4{,}3\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}99;\ \ F\text{-значение} = 8256{,}0;\ \ df = 122$$

t-значения: для постоянного члена = 17,40; для t = 10,03; для t^3 = 37,89; для n = 15,61

и

$$L = 0{,}394t^2 + 0{,}008R_c - 0{,}008\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}92; F = 489{,}8;\ \ df = 123$$

t-значения: для t^2 = 20,54; для R_c = 9,00; для n = 8,80

4. Для данных об акциях компаний из рейтинга S&P в 2014–2016 годах:

$$E = -28{,}012 + 22{,}540t - 87{,}173t^2 + 304{,}980t^3 + 0{,}058n$$

для данной модели

$$R^2 = 0{,}99;\ \ F\text{-значение} = 6527{,}0;\ \ df = 115$$

t-значения: для постоянного члена = 12,45; для t = 3,46; для t^2 = 3,09; для t^3 = 8,11; для n = 12,70

и

$$L = -1{,}091 + 0{,}555t^2 + 0{,}177R_c - 1{,}850\frac{n}{1000}$$

для данной модели

$$R^2 = 0{,}64;\ \ F\text{-значение} = 71{,}76;\ \ df = 116$$

t-значения: для постоянного члена = 5,92; для t^2 = 12,72; для R_c = 6,79; для n = 4,92

Скорее всего, причины такой сложной прогнозируемости в том, что метод полной связи практически гарантирует, что в один кластер не попадут ценные бумаги с противоположной изменчивостью, тогда как при проведении кластеризации по методу средней связи такая ситуация (ведущая к значительному повышению меры риска портфеля V) всё же гораздо более возможна (если у двух бумаг с противоположной изменчивостью есть «общие» схожие ценные бумаги).

Говоря о моделируемости зависимостей показателей E и L от параметров алгоритма, можно сделать вывод о том, что зависимость на всех наборах данных «в целом» построить достаточно сложно, даже если учитывать фактор исходной размерности n. Причина этого, с одной стороны, в том, что на разных рынках кластеры выделяются по-разному, а с другой — в том, что и для каждого из рынков в отдельности смоделировать зависимость может быть не слишком просто. Как следствие, многие из построенных уравнений регрессии, даже являясь значимыми (и в целом, и с точки зрения значимости каждого из коэффициентов в

отдельности), имеют коэффициент детерминации R^2 в пределах 0,63–0,69, то есть их объясняющая способность не очень высокая (хоть и реалистичная). В то же время, например, для данных по бирже NASDAQ зависимости моделируются очень хорошо (коэффициент детерминации R^2 только для одного из уравнений 0,87, а для всех остальных $R^2 > 0{,}9$, что говорит о высокой объясняющей способности моделей). Причина этого, скорее всего, в том, что биржа NASDAQ по своему составу относительно однородна, следствием чего является отсутствие при проведении кластеризации каких-то резких «скачков» как в размерности, так и в снижении возможностей диверсификации портфеля. С другой стороны, для остальных рынков, насколько можно предполагать, есть некоторые «точки перелома» для t, в которых объединяются два кластера, имеющие в своём составе ценные бумаги с противоположной изменчивостью, что, в свою очередь, приводит к значительному росту L.

Подводя итог данному разделу, можно сделать следующие выводы:

1. Для качественной реализации алгоритма производительность и оптимизация имеют значение
2. В то же время, хорошо реализованный алгоритм позволяет выполнять поставленную задачу — делать скорость расчётов на устаревшей технике сравнимой со скоростью расчётов на современной, дорогой технике
3. Качество работы алгоритма возможно контролировать, но для этого необходима подготовительная работа, возможно, с историческими ценами того же рынка. Лучше всего контролируется качество на относительно однородных данных, так как в них меньше ценных бумаг с противоположной изменчивостью
4. Лучше всего алгоритм контролируется при кластеризации по методу полной связи, немного хуже — при кластеризации по методу средней связи, хуже всего — при кластеризации по методу одиночной связи.

Заключение

В результате проведённого исследования было выявлено, что наилучшим способом сокращения размерности в задаче составления оптимального инвестиционного портфеля является иерархическая кластеризация. Вычислять расстояние между отдельными ценными бумагами необходимо с использованием меры расстояния на основе коэффициента парной корреляции между ними. Полученный в результате алгоритм можно применять на практике. В дальнейшем исследования стоит направить на совершенствование методов кластеризации (как самих алгоритмов, так и способов вычисления расстояний), чтобы сделать результаты более предсказуемыми. Также стоит обратить внимание на гладкие алгоритмы кластеризации, использование которых сейчас затруднено из-за отсутствия меры расстояния между ценными бумагами, которая являлась бы метрикой. Другой важной будущей задачей является обобщение алгоритма на задачу оптимизации «в целом», чтобы расширить сферу возможного будущего применения алгоритма.

Список литературы

[1] Markowitz H. Portfolio Selection. // The Journal of Finance. 1952. vol. 7:1. p. 77–91

[2] Diebold, F., Doherty N. Hering R. The known, the unknown, and the unknowable in financial risk management: Measurement and theory advancing practice. Princeton : Princeton University Press, 2010. 347 p.

[3] Дубровин В. И. Оськив О. И. Модели и методы оптимизации выбора инвестиционного портфеля // Радиоэлектроника, информатика, управление. 2008. Т. 1, С. 49–60

[4] Chaitanya J. Markowitz Portfolio Optimization [Электронный ресурс]. URL: `https://chaitjo.github.io/markowitz/` (дата доступа: 29.05.2020).

[5] Вишневер В. Я. Некоторые особенности конкуренции на российском фондовом рынке // Вестник Самарского государственного экономического университета 2016. № 140. С. 13–15.

[6] Музыкантов Д. С какой суммы начинать торговать на бирже [Электронный ресурс]. URL: `https://vc.ru/finance/63970-s-kakoy-summy-nachinat-torgovat-na-birzhe` (дата доступа: 30.05.2020).

[7] Орлов А. И. Прикладная статистика. Учебник. М. : Издательство «Экзамен», 2004. 656 с.

[8] Прикладная статистика: классификация и снижение размерности: Справочное издание. Айвазян С. А, Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Под ред. С. А. Айвазяна. М. : Финансы и статистика, 1989. 607 с.

[9] Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования [Электронный ресурс]. URL: `http://www.machinelearning.ru/wiki/images/c/ca/Voron-ML-Clustering.pdf` (дата доступа: 31.05.2020).

[10] Бронштейн Е. М., Ишманов И. Н. Формирование портфеля ценных бумаг на основе распознавания образов // Финансовая аналитика: проблемы и решения. 2010. №. 11(35). С. 35–39.

[11] Тюхова Е. М., Сизых Д. С. Использование кластерного анализа для формирования портфеля ценных бумаг в инвестиционных системах (робоэдвайзерах) // Управление развитием крупномасштабных систем MLSD'2019 М., 2019 г. / отв. ред. С.Н. Васильев. М., 2019. С. 300–303.

[12] Tola V., et. al. Cluster analysis for portfolio optimization Journal of Economic Dynamics and Control. 2008. Vol. 32(1). P. 235–258

[13] León D., et. al. Clustering algorithms for Risk-Adjusted Portfolio Construction. // Procedia Computer Science 2017. Vol. 108. P. 1334–1343

[14] López de Prado, M. A robust estimator of the efficient frontier. [Электронный ресурс] URL: https://ssrn.com/abstract=3469961 (дата доступа: 30.05.2020).

[15] Falcone J.-L., Albuquerque P. A Correlation-Based Distance. [Электронный ресурс] URL:http://www.arxiv.org/abs/cs.IR/0402061 (дата доступа: 31.05.2020).

[16] Полетаев А. Ю., Спиридонова Е. М. Иерархическая кластеризация как метод снижения размерности в задаче оптимизации инвестиционного портфеля Марковица // Моделирование и анализ информационных систем 2020. №. 1(27). С. 62–71.

[17] Diamond S., Boyd S. CVXPY: A Python-embedded modeling language for convex optimization // Journal of Machine Learning Research 2016. Vol. 17(83). P. 1–5

[18] Agrawal A., Verschueren R., Diamond S., Boyd S. A rewriting system for convex optimization problems // Journal of Control and Decision 2018. Vol. 5(1).P. 42–60

[19] Virtanen P., et. al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python // Nature Methods 2020. Vol. 17. P. 261–272

[20] Behnel S., Bradshaw R., Citro C. Dalcin L. Seljebotn D. S. Smith K. Cython: The Best of Both Worlds // Computing in Science Engineering 2011. Vol. 13(2). P. 31–39

[21] van der Walt S, Colbert S. C., Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation // Computing in Science & Engineering 2011. Vol. 13(2). P. 22–30

Приложение А

Исходный код реализации предложенного алгоритма на Python

Основной модуль portfolio_optimization.py

```
import numpy as np
import numpy_indexed as npi
import cvxpy as cp
from functools import partial
from scipy.cluster.hierarchy import fcluster, linkage
from scipy.stats import gmean

from cythonized.recalculation import recalculate_data,
        recalculate_returns


def optimize_use_clustering(data, value, t, dist, returns,
        crit='return_bound'):
    clusters = form_clusters(data, t, dist)
    clustered_data = recalculate_data(data, clusters)
    clustered_returns = recalculate_returns(returns, clusters)
    w = optimize_port(clustered_data, value, clustered_returns, crit)
    x = clusters_weights(w, clusters)
    return x


def form_clusters(x, t, dist, return_count=False):
    link = linkage(x, metric='correlation', method=dist)
    res = fcluster(Z=link, t=t, criterion='distance')
    if return_count:
        res = res, np.unique(res).max()
    return res


def recalculate_data_npi(data, clusters):
    calc_average_change = partial(np.apply_along_axis,
        func1d=gmean, axis=0)
    clustered = np.empty((clusters.max(), data.shape[1]))
```

```
34      groups = npi.group_by(clusters).split(data)
35      for i, clust in enumerate(groups):
36          clustered[i] = calc_average_change(arr=clust)
37      return clustered
38
39
40  def recalculate_returns_npi(returns, clusters):
41      calc_average_change = partial(np.apply_along_axis,
42          func1d=gmean, axis=0)
43      clustered = np.empty((clusters.max(),))
44      groups = npi.group_by(clusters).split(returns)
45      for i, clust in enumerate(groups):
46          clustered[i] = calc_average_change(arr=clust)
47      return clustered
48
49
50  def optimize_port(data, value, returns, criterion='return_bound'):
51      CV = np.cov(data)
52      weights_var = cp.Variable(data.shape[0])
53      return_var = returns * weights_var
54      risk_var = cp.quad_form(weights_var, CV)
55      if criterion == "risk_bound":
56          prob = cp.Problem(cp.Maximize(return_var),
57              [risk_var <= cp.Parameter(name="max_risk",
58               value=value, nonneg=True),
59               cp.sum(weights_var) == 1, weights_var >= 0])
60      elif criterion == "return_bound":
61          prob = cp.Problem(cp.Minimize(risk_var),
62              [return_var >= cp.Parameter(name="min_ret",
63                value=value, nonneg=True),
64               cp.sum(weights_var) == 1, weights_var >= 0])
65      elif criterion == "risk_return":
66          prob = cp.Problem(cp.Maximize(return_var -\
67              cp.Parameter(name="gamma",
68               value=value, nonneg=True) * risk_var),
69              [cp.sum(weights_var) == 1, weights_var >= 0])
70      prob.solve()
71      return weights_var.value
72
73
```

```
74 def clusters_weights(W, clusters):
75     _, clusters_sizes = np.unique(clusters, return_counts=True)
76     function = np.vectorize(lambda x: (W / clusters_sizes)[x - 1])
77     return function(clusters)
78 
79 
80 def calc_risk(w: np.ndarray, cv: np.ndarray) -> float:
81     if len(w.shape) == 1:
82         w = w.reshape(1, w.shape[0])
83     return (w.T.dot(w) * cv).sum()
84 
85 
86 def calc_revenue(w: np.ndarray, d: np.ndarray) -> float:
87     return w.dot(d)
```

Модуль cythonized.recalculation.pyx

```
1  cimport cython
2  
3  cimport numpy as np
4  import numpy as np
5  from libc.math cimport pow
6  
7  
8  def recalculate_data(np.ndarray np_data, np.ndarray np_clusters):
9      cdef long k = np.unique(np_clusters).max()
10     cdef long n = np_data.shape[0], m = np_data.shape[1]
11     cdef long i = 0, j = 0
12     cdef double [:,:] data = np_data
13     cdef long [:] clusters = np_clusters.astype(np.int64)
14     cdef double [:,:] res = np.ones((k, m))
15     cdef long [:] clusters_sizes = np.zeros(k, dtype=np.int64)
16 
17     for i in range(n):
18         clusters_sizes[clusters[i] - 1] += 1
19         for j in range(m):
20             res[clusters[i] - 1][j] *= data[i][j]
21 
22     for i in range(k):
23         for j in range(m):
24             res[i][j] = pow(res[i][j], 1 / clusters_sizes[i])
25     return np.array(res)
```

```
26
27
28 def recalculate_returns(np.ndarray np_returns,
29         np.ndarray np_clusters):
30     cdef long k = np.unique(np_clusters).max()
31     cdef long n = np_returns.shape[0]
32     cdef long i = 0
33     cdef double [:] returns = np_returns
34     cdef long [:] clusters = np_clusters.astype(np.int64)
35     cdef double [:] res = np.ones((k,))
36     cdef long [:] clusters_sizes = np.zeros(k, dtype=np.int64)
37
38     for i in range(n):
39         clusters_sizes[clusters[i] - 1] += 1
40         res[clusters[i] - 1] *= returns[i]
41
42     for i in range(k):
43         res[i] = pow(res[i], 1 / clusters_sizes[i])
44     return np.array(res)
```

Приложение Б

Результаты применения предлагаемого алгоритма

- T_0 — время работы программы для составления оптимального портфеля без применения предлагаемого алгоритма
- T_p — время работы программы для составления оптимального портфеля с применением предлагаемого алгоритма и пересчётом цен, реализованным на *Python* и *NumPy*
- T_p — время работы программы для составления оптимального портфеля с применением предлагаемого алгоритма и пересчётом цен, реализованным на *Cython*
- T_p — время работы программы для составления оптимального портфеля с применением предлагаемого алгоритма, для пересчёта цен используется *numpy-indexed*

Таблица Б.1

Средние результаты применения предлагаемого метода на акциях, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах при кластеризации по методу одиночной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	2,4	0,00	17,60	6,51	3,85	16,11	3,32	0,85	0,47	1,60
1,10	0,10	2,4	0,01	16,88	6,65	3,49	14,98	2,54	0,82	0,47	1,57
1,15	0,10	2,4	0,01	17,06	6,68	3,61	15,08	2,61	0,86	0,47	1,59
1,20	0,10	2,4	0,01	16,67	7,10	3,89	15,38	2,37	0,87	0,47	1,57
1,25	0,10	2,4	0,02	16,61	6,51	3,26	14,84	2,49	0,77	0,41	1,49
1,30	0,10	2,4	0,02	14,69	6,77	3,61	15,04	2,16	0,80	0,42	1,57
1,05	0,15	4,6	0,01	17,36	4,62	1,73	8,20	3,36	0,56	0,24	0,84
1,10	0,15	4,6	0,02	17,54	4,76	1,61	7,92	2,65	0,58	0,24	0,85
1,15	0,15	4,6	0,02	17,63	4,75	1,70	7,88	2,75	0,54	0,22	0,81
1,20	0,15	4,6	0,03	17,26	4,80	1,69	7,79	2,48	0,56	0,22	0,81
1,25	0,15	4,6	0,04	14,93	4,85	1,63	7,77	2,33	0,55	0,24	0,80
1,30	0,15	4,6	0,1	14,26	5,02	1,67	7,78	2,16	0,55	0,22	0,83
1,05	0,20	8,6	0,02	17,38	4,10	1,31	4,86	3,31	0,50	0,18	0,51
1,10	0,20	8,6	0,03	18,52	4,51	1,35	4,49	2,55	0,50	0,19	0,49
1,15	0,20	8,6	0,04	17,79	4,32	1,43	4,70	2,82	0,50	0,17	0,50
1,20	0,20	8,6	0,1	17,97	4,38	1,31	4,49	2,57	0,51	0,18	0,48
1,25	0,20	8,6	0,1	15,96	4,28	1,35	4,48	2,41	0,52	0,17	0,48
1,30	0,20	8,6	0,1	13,74	4,55	1,47	4,63	2,07	0,50	0,18	0,48
1,05	0,25	16,8	0,03	22,47	4,06	1,26	2,55	3,23	0,48	0,18	0,33
1,10	0,25	16,8	0,04	18,83	4,22	1,19	2,82	2,57	0,48	0,16	0,33
1,15	0,25	16,8	0,1	19,28	4,17	1,42	2,77	2,70	0,46	0,18	0,31
1,20	0,25	16,8	0,1	18,28	4,32	1,46	2,68	2,49	0,48	0,19	0,32
1,25	0,25	16,8	0,1	16,12	4,14	1,34	2,54	2,42	0,49	0,17	0,33
1,30	0,25	16,8	0,1	15,54	4,28	1,36	2,67	2,19	0,51	0,18	0,34
1,05	0,30	37,4	0,06	19,06	3,86	1,26	1,77	3,25	0,45	0,17	0,22
1,10	0,30	37,4	0,1	19,13	4,18	1,26	1,66	2,61	0,48	0,20	0,24
1,15	0,30	37,4	0,1	16,57	4,08	1,27	1,81	2,63	0,45	0,15	0,22
1,20	0,30	37,4	0,1	18,63	4,05	1,17	1,79	2,48	0,47	0,19	0,23
1,25	0,30	37,4	0,2	16,90	3,95	1,26	1,63	2,40	0,48	0,19	0,22
1,30	0,30	37,4	0,3	13,36	4,09	1,33	1,78	2,17	0,47	0,18	0,24
1,05	0,35	80,4	0,1	22,98	3,74	1,50	1,11	3,19	0,43	0,14	0,17
1,10	0,35	80,4	0,1	18,03	3,88	1,45	1,19	2,58	0,43	0,14	0,17
1,15	0,35	80,4	0,2	18,58	3,74	1,14	1,26	2,72	0,43	0,14	0,17
1,20	0,35	–	–	–	–	–	–	–	–	–	–
1,25	0,35	–	–	–	–	–	–	–	–	–	–
1,30	0,35	–	–	–	–	–	–	–	–	–	–
1,05	0,40	219,0	0,2	17,38	3,70	1,17	0,97	3,23	0,44	0,14	0,15
1,10	0,40	–	–	–	–	–	–	–	–	–	–
1,15	0,40	–	–	–	–	–	–	–	–	–	–
1,20	0,40	–	–	–	–	–	–	–	–	–	–
1,25	0,40	–	–	–	–	–	–	–	–	–	–
1,30	0,40	–	–	–	–	–	–	–	–	–	–

Таблица Б.2

Средние результаты применения предлагаемого метода на акциях, торгуемых на бирже NASDAQ в 2004–2006 годах при кластеризации по методу одиночной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,7	0,01	8,52	8,19	4,97	20,04	1,45	1,18	0,79	2,07
1,10	0,10	1,7	0,01	10,29	7,02	3,75	18,66	1,59	0,88	0,48	1,80
1,15	0,10	1,7	0,01	12,10	7,76	4,06	18,56	1,77	0,91	0,53	1,85
1,20	0,10	1,7	0,01	13,50	7,26	4,50	18,69	1,78	0,96	0,58	1,85
1,25	0,10	1,7	0,01	12,82	7,32	4,31	18,51	1,83	0,89	0,56	1,87
1,30	0,10	1,7	0,01	10,17	7,97	3,78	18,73	1,58	0,95	0,55	1,85
1,05	0,15	2,9	0,01	8,58	4,66	1,65	11,06	1,49	0,64	0,28	1,10
1,10	0,15	2,9	0,01	11,67	5,17	1,54	11,26	1,62	0,58	0,25	1,03
1,15	0,15	2,9	0,02	11,37	5,03	1,65	11,11	1,72	0,58	0,25	1,05
1,20	0,15	2,9	0,02	11,72	4,88	1,60	10,63	1,78	0,58	0,24	1,02
1,25	0,15	2,9	0,02	11,81	4,75	1,79	10,81	1,84	0,60	0,29	1,08
1,30	0,15	2,9	0,0	11,17	5,03	1,84	10,86	1,58	0,57	0,26	1,03
1,05	0,20	5,0	0,02	9,37	3,78	1,08	6,84	1,48	0,49	0,18	0,64
1,10	0,20	5,0	0,02	9,54	3,82	1,15	6,94	1,84	0,50	0,17	0,64
1,15	0,20	5,0	0,03	11,85	3,94	1,06	6,64	1,74	0,47	0,17	0,61
1,20	0,20	5,0	0,0	13,01	4,17	1,15	6,64	1,68	0,49	0,18	0,61
1,25	0,20	5,0	0,0	12,00	4,34	1,14	6,81	1,92	0,50	0,17	0,62
1,30	0,20	5,0	0,0	10,52	4,08	1,18	6,18	1,57	0,50	0,19	0,63
1,05	0,25	9,1	0,03	9,54	3,55	0,92	4,43	1,46	0,42	0,17	0,38
1,10	0,25	9,1	0,03	9,22	3,57	0,90	3,74	1,66	0,41	0,18	0,39
1,15	0,25	9,1	0,0	12,22	3,65	0,88	4,18	1,80	0,44	0,17	0,38
1,20	0,25	9,1	0,0	12,27	4,01	1,05	3,99	1,79	0,45	0,16	0,42
1,25	0,25	9,1	0,0	9,37	3,92	0,92	4,07	1,95	0,44	0,15	0,41
1,30	0,25	9,1	0,1	9,80	3,74	0,93	4,35	1,61	0,43	0,15	0,38
1,05	0,30	16,0	0,04	9,26	3,35	0,81	2,97	1,42	0,41	0,15	0,27
1,10	0,30	16,0	0,1	10,54	3,88	0,87	2,67	1,68	0,42	0,15	0,27
1,15	0,30	16,0	0,1	11,55	3,81	0,84	2,73	1,76	0,42	0,14	0,27
1,20	0,30	16,0	0,1	10,98	3,35	0,90	2,96	1,75	0,42	0,16	0,28
1,25	0,30	16,0	0,1	10,50	3,83	0,91	2,50	1,95	0,42	0,15	0,28
1,30	0,30	16,0	0,1	9,88	3,99	0,96	2,53	1,61	0,44	0,15	0,28
1,05	0,35	31,6	0,1	8,43	3,23	0,76	1,67	1,48	0,42	0,16	0,20
1,10	0,35	31,6	0,1	10,84	3,57	0,87	1,64	1,76	0,42	0,15	0,20
1,15	0,35	31,6	0,1	12,24	3,43	0,99	1,55	1,82	0,40	0,14	0,20
1,20	0,35	31,6	0,2	12,1	3,4	1,0	1,7	1,7	0,4	0,2	0,2
1,25	0,35	31,6	0,2	12,2	3,6	1,0	1,6	1,9	0,4	0,2	0,2
1,30	0,35	31,6	0,2	11,2	3,6	1,0	1,5	1,6	0,4	0,1	0,2
1,05	0,40	56,0	0,1	9,10	3,31	0,83	1,29	1,48	0,38	0,12	0,15
1,10	0,40	56,0	0,2	11,3	3,4	1,0	1,2	1,8	0,4	0,1	0,2
1,15	0,40	56,0	0,2	12,9	3,2	1,0	1,2	1,8	0,4	0,1	0,2
1,20	0,40	56,0	0,3	10,7	3,7	0,9	1,3	1,8	0,4	0,1	0,2
1,25	0,40	56,0	0,3	11,9	3,3	1,0	1,2	1,9	0,4	0,1	0,2
1,30	0,40	56,0	0,4	11,0	3,0	0,9	1,2	1,6	0,4	0,1	0,2

Таблица Б.3

Средние результаты применения предлагаемого метода на акциях компаний из рейтинга S&P 500 в 2014–2016 годах при кластеризации по методу одиночной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	2,4	0,01	1,18	1,90	0,39	4,92	0,22	0,26	0,11	0,52
1,10	0,10	2,4	0,02	1,52	1,67	0,38	5,00	0,25	0,24	0,11	0,51
1,15	0,10	2,4	0,02	1,18	1,90	0,38	5,02	0,25	0,25	0,11	0,52
1,20	0,10	2,4	0,03	1,27	1,88	0,41	5,12	0,27	0,24	0,10	0,52
1,25	0,10	2,4	0,03	1,20	1,78	0,39	5,06	0,25	0,24	0,09	0,50
1,30	0,10	2,4	0,03	1,53	1,78	0,53	5,09	0,25	0,24	0,13	0,55
1,05	0,15	4,3	0,03	1,14	1,52	0,26	2,89	0,23	0,21	0,06	0,32
1,10	0,15	4,3	0,03	1,19	1,49	0,29	2,92	0,25	0,22	0,06	0,34
1,15	0,15	4,3	0,04	1,19	1,52	0,19	2,92	0,26	0,24	0,06	0,31
1,20	0,15	4,3	0,05	1,25	1,58	0,32	2,94	0,28	0,23	0,07	0,31
1,25	0,15	4,3	0,08	1,22	1,61	0,23	2,92	0,26	0,24	0,06	0,32
1,30	0,15	–	–	–	–	–	–	–	–	–	–
1,05	0,20	9,4	0,06	1,10	1,43	0,23	1,50	0,21	0,21	0,04	0,19
1,10	0,20	9,4	0,06	1,28	1,23	0,20	1,47	0,25	0,20	0,05	0,19
1,15	0,20	9,4	0,08	1,38	1,36	0,33	1,43	0,25	0,19	0,08	0,19
1,20	0,20	–	–	–	–	–	–	–	–	–	–
1,25	0,20	–	–	–	–	–	–	–	–	–	–
1,30	0,20	–	–	–	–	–	–	–	–	–	–
1,05	0,25	24,5	0,08	1,34	1,23	0,28	0,77	0,23	0,16	0,03	0,09
1,10	0,25	24,5	0,09	1,38	1,18	0,26	0,74	0,23	0,16	0,03	0,07
1,15	0,25	–	–	–	–	–	–	–	–	–	–
1,20	0,25	–	–	–	–	–	–	–	–	–	–
1,25	0,25	–	–	–	–	–	–	–	–	–	–
1,30	0,25	–	–	–	–	–	–	–	–	–	–
1,05	0,30	55,2	0,13	1,11	1,07	0,16	0,36	0,21	0,13	0,03	0,05
1,10	0,30	–	–	–	–	–	–	–	–	–	–
1,15	0,30	–	–	–	–	–	–	–	–	–	–
1,20	0,30	–	–	–	–	–	–	–	–	–	–
1,25	0,30	–	–	–	–	–	–	–	–	–	–
1,30	0,30	–	–	–	–	–	–	–	–	–	–
1,05	0,35	–	–	–	–	–	–	–	–	–	–
1,10	0,35	–	–	–	–	–	–	–	–	–	–
1,15	0,35	–	–	–	–	–	–	–	–	–	–
1,20	0,35	–	–	–	–	–	–	–	–	–	–
1,25	0,35	–	–	–	–	–	–	–	–	–	–
1,30	0,35	–	–	–	–	–	–	–	–	–	–
1,05	0,40	–	–	–	–	–	–	–	–	–	–
1,10	0,40	–	–	–	–	–	–	–	–	–	–
1,15	0,40	–	–	–	–	–	–	–	–	–	–
1,20	0,40	–	–	–	–	–	–	–	–	–	–
1,25	0,40	–	–	–	–	–	–	–	–	–	–
1,30	0,40	–	–	–	–	–	–	–	–	–	–

Таблица Б.4

Средние результаты применения предлагаемого метода на акциях, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах при кластеризации по методу полной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,6	0,00	19,42	11,68	8,34	25,85	3,18	1,54	1,15	2,80
1,10	0,10	1,6	0,00	17,92	11,80	8,37	25,59	2,56	1,42	1,08	2,68
1,15	0,10	1,6	0,00	18,32	11,21	8,31	25,39	2,60	1,32	0,95	2,64
1,20	0,10	1,6	0,00	16,35	11,39	8,25	25,31	2,38	1,34	1,02	2,61
1,25	0,10	1,6	0,00	16,16	10,59	7,18	24,84	2,13	1,38	0,93	2,59
1,30	0,10	1,6	0,01	13,66	10,91	7,31	24,78	2,09	1,41	0,99	2,66
1,05	0,15	2,1	0,00	20,12	7,75	4,57	17,19	3,01	0,93	0,55	1,81
1,10	0,15	2,1	0,01	15,96	7,43	4,14	16,84	2,56	0,90	0,55	1,76
1,15	0,15	2,1	0,01	17,45	7,06	3,97	17,08	2,59	0,89	0,48	1,78
1,20	0,15	2,1	0,01	16,14	7,55	3,98	17,12	2,47	0,91	0,53	1,81
1,25	0,15	2,1	0,02	15,88	7,40	4,46	17,19	2,26	0,92	0,55	1,80
1,30	0,15	2,1	0,02	13,78	8,29	4,92	17,43	2,20	0,93	0,51	1,81
1,05	0,20	2,8	0,01	16,84	6,41	3,53	13,16	3,42	0,72	0,37	1,33
1,10	0,20	2,8	0,02	17,30	6,13	2,86	12,55	2,44	0,75	0,34	1,31
1,15	0,20	2,8	0,02	16,88	6,39	2,74	12,75	2,61	0,73	0,35	1,29
1,20	0,20	2,8	0,03	17,30	6,22	2,91	12,77	2,53	0,73	0,34	1,31
1,25	0,20	2,8	0,04	16,22	5,92	2,70	12,40	2,33	0,72	0,35	1,32
1,30	0,20	2,8	0,05	13,89	5,75	2,36	12,20	2,15	0,67	0,33	1,27
1,05	0,25	3,7	0,01	17,37	5,17	2,26	9,93	3,26	0,65	0,28	1,03
1,10	0,25	3,7	0,02	18,05	5,30	2,03	9,49	2,59	0,62	0,27	1,00
1,15	0,25	3,7	0,03	17,87	5,48	2,08	9,77	2,59	0,61	0,27	0,99
1,20	0,25	3,7	0,04	16,30	5,18	2,17	9,48	2,46	0,62	0,26	1,00
1,25	0,25	3,7	0,05	16,18	5,23	2,08	9,59	2,31	0,62	0,27	1,00
1,30	0,25	3,7	0,07	14,30	5,35	2,07	9,46	2,16	0,63	0,26	0,99
1,05	0,30	4,7	0,02	16,82	4,88	1,70	8,07	3,18	0,58	0,25	0,79
1,10	0,30	4,7	0,03	17,63	4,87	1,69	7,89	2,59	0,58	0,23	0,78
1,15	0,30	4,7	0,03	18,23	5,05	1,80	7,61	2,67	0,57	0,25	0,80
1,20	0,30	4,7	0,05	17,54	4,93	1,90	7,63	2,48	0,59	0,22	0,81
1,25	0,30	4,7	0,06	16,57	4,72	1,65	7,48	2,26	0,57	0,24	0,81
1,30	0,30	4,7	0,08	14,28	5,02	1,93	7,51	2,11	0,59	0,23	0,82
1,05	0,35	6,0	0,02	20,25	4,42	1,78	6,07	3,47	0,54	0,23	0,68
1,10	0,35	6,0	0,03	17,57	4,68	1,62	6,48	2,56	0,55	0,20	0,64
1,15	0,35	6,0	0,04	15,62	4,64	1,57	6,15	2,75	0,54	0,20	0,65
1,20	0,35	6,0	0,06	16,92	4,74	1,70	6,15	2,52	0,55	0,23	0,66
1,25	0,35	6,0	0,07	15,27	4,46	1,69	5,96	2,37	0,56	0,22	0,68
1,30	0,35	6,0	0,08	13,25	4,62	1,75	6,11	2,11	0,55	0,22	0,64
1,05	0,40	7,5	0,03	16,18	4,18	1,42	5,47	3,44	0,54	0,21	0,56
1,10	0,40	7,5	0,04	17,38	4,53	1,45	5,22	2,59	0,54	0,21	0,57
1,15	0,40	7,5	0,06	16,74	4,32	1,58	5,28	2,56	0,53	0,19	0,55
1,20	0,40	7,5	0,07	17,22	4,46	1,48	5,10	2,49	0,55	0,22	0,55
1,25	0,40	7,5	0,10	14,71	4,50	1,44	5,03	2,43	0,51	0,20	0,56
1,30	0,40	7,5	0,12	14,48	4,60	1,54	5,09	2,13	0,54	0,17	0,56

Таблица Б.5

Средние результаты применения предлагаемого метода на акциях, торгуемых на бирже NASDAQ в 2004–2006 годах при кластеризации по методу полной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,3	0,01	10,77	12,27	9,35	26,64	1,48	1,57	1,26	2,96
1,10	0,10	1,3	0,01	11,04	9,57	6,02	25,65	1,68	1,45	0,99	2,65
1,15	0,10	1,3	0,01	11,83	10,38	6,68	25,13	1,75	1,29	0,92	2,57
1,20	0,10	1,3	0,01	12,88	10,57	7,65	26,49	1,74	1,36	1,04	2,66
1,25	0,10	1,3	0,01	11,52	10,06	8,27	26,11	1,90	1,40	1,06	2,73
1,30	0,10	1,3	0,01	11,45	11,18	7,17	26,19	1,50	1,32	0,98	2,60
1,05	0,15	1,7	0,01	10,65	9,26	5,15	20,92	1,43	1,09	0,70	2,05
1,10	0,15	1,7	0,01	10,92	8,96	5,19	19,91	1,74	0,98	0,63	1,99
1,15	0,15	1,7	0,01	11,38	8,31	4,52	19,96	1,73	0,97	0,57	1,95
1,20	0,15	1,7	0,01	12,69	7,56	4,44	19,71	1,70	0,92	0,61	1,90
1,25	0,15	1,7	0,01	10,25	8,28	4,67	20,43	1,84	0,99	0,66	1,94
1,30	0,15	1,7	0,01	11,00	7,96	5,34	19,98	1,58	0,96	0,61	1,93
1,05	0,20	2,1	0,01	9,70	6,50	2,76	15,50	1,54	0,75	0,42	1,46
1,10	0,20	2,1	0,01	9,90	6,00	2,79	15,39	1,66	0,73	0,39	1,44
1,15	0,20	2,1	0,01	11,18	6,09	2,50	14,79	1,67	0,67	0,35	1,41
1,20	0,20	2,1	0,02	10,69	6,29	2,80	15,03	1,78	0,72	0,39	1,40
1,25	0,20	2,1	0,02	13,54	6,29	3,26	15,07	1,90	0,74	0,40	1,44
1,30	0,20	2,1	0,02	9,95	6,06	2,68	14,95	1,65	0,74	0,38	1,45
1,05	0,25	2,7	0,02	9,90	5,63	2,10	11,60	1,46	0,63	0,27	1,14
1,10	0,25	2,7	0,02	10,38	4,85	1,99	12,70	1,65	0,61	0,27	1,08
1,15	0,25	2,7	0,02	12,27	5,69	1,84	11,35	1,69	0,59	0,26	1,11
1,20	0,25	2,7	0,02	11,87	5,20	2,15	11,55	1,78	0,65	0,28	1,11
1,25	0,25	2,7	0,02	13,49	5,06	2,23	11,36	1,86	0,61	0,29	1,12
1,30	0,25	2,7	0,03	10,66	5,49	2,20	11,52	1,56	0,63	0,28	1,09
1,05	0,30	3,5	0,02	9,05	4,60	1,57	9,98	1,40	0,57	0,23	0,89
1,10	0,30	3,5	0,02	10,42	4,40	1,62	9,85	1,64	0,55	0,24	0,87
1,15	0,30	3,5	0,03	11,26	4,97	1,42	9,59	1,79	0,56	0,22	0,88
1,20	0,30	3,5	0,03	11,37	5,05	1,42	9,66	1,75	0,55	0,23	0,88
1,25	0,30	3,5	0,04	12,72	4,54	1,70	9,17	1,85	0,55	0,23	0,91
1,30	0,30	3,5	0,04	10,65	4,90	1,61	9,44	1,59	0,57	0,25	0,90
1,05	0,35	4,3	0,03	9,64	4,13	1,41	7,95	1,44	0,52	0,23	0,73
1,10	0,35	4,3	0,03	11,98	4,41	1,33	7,94	1,65	0,51	0,22	0,72
1,15	0,35	4,3	0,04	11,35	4,36	1,40	7,95	1,80	0,53	0,20	0,74
1,20	0,35	4,3	0,05	11,64	4,67	1,44	7,35	1,76	0,51	0,19	0,73
1,25	0,35	4,3	0,06	11,81	4,57	1,45	7,71	1,85	0,50	0,19	0,71
1,30	0,35	4,3	0,06	9,82	4,50	1,38	7,76	1,59	0,51	0,22	0,72
1,05	0,40	5,3	0,03	10,56	3,98	1,35	6,82	1,53	0,49	0,19	0,61
1,10	0,40	5,3	0,04	10,76	4,41	1,36	6,33	1,83	0,50	0,19	0,61
1,15	0,40	5,3	0,05	12,30	4,33	1,29	6,48	1,74	0,47	0,20	0,60
1,20	0,40	5,3	0,06	9,75	4,18	1,13	6,42	1,89	0,50	0,18	0,59
1,25	0,40	5,3	0,07	12,29	4,23	1,31	6,50	1,83	0,48	0,18	0,60
1,30	0,40	5,3	0,08	9,93	4,30	1,34	6,20	1,67	0,51	0,19	0,63

Таблица Б.6

Средние результаты применения предлагаемого метода на акциях компаний из рейтинга S&P 500 в 2014–2016 годах при кластеризации по методу полной связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,6	0,01	1,15	2,14	0,58	7,49	0,22	0,29	0,14	0,75
1,10	0,10	1,6	0,01	1,48	2,10	0,63	7,47	0,24	0,29	0,14	0,77
1,15	0,10	1,6	0,01	1,30	2,10	0,58	7,57	0,24	0,30	0,14	0,77
1,20	0,10	1,6	0,01	1,19	2,13	0,53	7,68	0,25	0,30	0,14	0,74
1,25	0,10	1,6	0,02	1,15	2,19	0,51	7,57	0,24	0,27	0,12	0,78
1,30	0,10	1,6	0,02	1,52	1,88	0,63	7,62	0,25	0,30	0,14	0,76
1,05	0,15	2,0	0,01	1,18	1,99	0,45	5,78	0,23	0,27	0,11	0,60
1,10	0,15	2,0	0,01	1,22	1,79	0,53	5,79	0,25	0,27	0,13	0,59
1,15	0,15	2,0	0,02	1,51	1,65	0,57	5,90	0,25	0,28	0,09	0,60
1,20	0,15	2,0	0,02	1,17	1,94	0,40	5,97	0,25	0,27	0,10	0,60
1,25	0,15	2,0	0,02	1,57	2,18	0,51	5,90	0,24	0,27	0,12	0,58
1,30	0,15	2,0	0,03	1,10	1,95	0,42	5,83	0,23	0,26	0,11	0,61
1,05	0,20	2,7	0,02	1,28	1,67	0,36	4,25	0,23	0,23	0,11	0,48
1,10	0,20	2,7	0,02	1,41	1,82	0,35	4,49	0,24	0,25	0,10	0,46
1,15	0,20	2,7	0,03	1,40	1,69	0,36	4,35	0,24	0,24	0,11	0,48
1,20	0,20	2,7	0,03	1,18	1,82	0,27	4,43	0,24	0,23	0,07	0,47
1,25	0,20	2,7	0,03	1,19	1,64	0,39	4,56	0,24	0,23	0,08	0,45
1,30	0,20	2,7	0,05	1,38	1,70	0,42	4,45	0,24	0,23	0,08	0,46
1,05	0,25	3,6	0,02	1,19	1,72	0,26	3,41	0,23	0,22	0,08	0,36
1,10	0,25	3,6	0,03	1,34	1,60	0,30	3,61	0,24	0,20	0,08	0,36
1,15	0,25	3,6	0,03	1,20	1,70	0,23	3,50	0,24	0,21	0,07	0,37
1,20	0,25	3,6	0,04	1,14	1,76	0,27	3,63	0,26	0,23	0,06	0,37
1,25	0,25	3,6	0,04	1,27	1,69	0,33	3,63	0,24	0,22	0,08	0,40
1,30	0,25	3,6	0,06	1,46	1,68	0,30	3,41	0,24	0,23	0,09	0,37
1,05	0,30	4,6	0,03	1,53	1,34	0,34	2,81	0,22	0,21	0,08	0,28
1,10	0,30	4,6	0,04	1,34	1,61	0,29	2,81	0,25	0,22	0,05	0,30
1,15	0,30	4,6	0,04	1,11	1,51	0,23	2,69	0,26	0,23	0,07	0,29
1,20	0,30	4,6	0,05	1,25	1,61	0,29	2,79	0,27	0,22	0,04	0,32
1,25	0,30	4,6	0,05	1,25	1,66	0,30	2,90	0,23	0,23	0,08	0,31
1,30	0,30	4,6	0,07	1,30	1,61	0,29	2,76	0,25	0,22	0,08	0,33
1,05	0,35	5,7	0,04	1,18	1,39	0,24	2,33	0,24	0,22	0,07	0,27
1,10	0,35	5,7	0,04	1,23	1,51	0,25	2,40	0,25	0,21	0,05	0,28
1,15	0,35	5,7	0,05	1,24	1,54	0,32	2,25	0,26	0,20	0,05	0,26
1,20	0,35	5,7	0,06	1,26	1,46	0,24	2,33	0,25	0,21	0,06	0,25
1,25	0,35	5,7	0,07	1,35	1,48	0,25	2,24	0,24	0,23	0,06	0,28
1,30	0,35	5,7	0,09	1,33	1,54	0,34	2,35	0,27	0,23	0,07	0,27
1,05	0,40	6,9	0,04	1,07	1,43	0,21	2,01	0,26	0,20	0,07	0,23
1,10	0,40	6,9	0,05	1,25	1,55	0,21	1,94	0,24	0,20	0,05	0,22
1,15	0,40	6,9	0,06	1,04	1,30	0,19	1,89	0,24	0,20	0,06	0,23
1,20	0,40	6,9	0,07	1,09	1,22	0,20	1,74	0,25	0,21	0,05	0,24
1,25	0,40	6,9	0,09	1,15	1,31	0,30	1,81	0,24	0,21	0,05	0,23
1,30	0,40	6,9	0,11	1,30	1,36	0,26	2,01	0,26	0,23	0,06	0,23

Таблица Б.7

Средние результаты применения предлагаемого метода на акциях, торгуемых на Нью-Йоркской фондовой бирже в 2004–2006 годах при кластеризации по методу средней связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,7	0,00	23,73	10,00	6,68	22,68	3,31	1,26	0,87	2,45
1,10	0,10	1,7	0,00	17,32	9,43	6,72	22,22	2,52	1,14	0,83	2,35
1,15	0,10	1,7	0,00	17,23	10,38	6,84	22,74	2,62	1,24	0,89	2,43
1,20	0,10	1,7	0,00	17,20	9,85	6,64	22,58	2,50	1,21	0,82	2,39
1,25	0,10	1,7	0,01	15,58	9,86	6,85	22,24	2,39	1,31	0,85	2,46
1,30	0,10	1,7	0,01	13,77	8,90	6,30	21,92	2,04	1,21	0,85	2,38
1,05	0,15	2,5	0,00	20,24	6,45	3,49	14,40	3,20	0,82	0,44	1,57
1,10	0,15	2,5	0,01	17,59	7,02	3,45	14,66	2,58	0,84	0,44	1,57
1,15	0,15	2,5	0,01	18,46	6,86	3,61	14,94	2,62	0,81	0,42	1,54
1,20	0,15	2,5	0,01	16,91	6,81	3,38	14,80	2,35	0,82	0,42	1,51
1,25	0,15	2,5	0,02	16,44	6,61	3,55	14,70	2,15	0,85	0,42	1,56
1,30	0,15	2,5	0,03	13,58	6,54	3,43	14,49	2,12	0,80	0,41	1,53
1,05	0,20	3,5	0,01	21,07	5,14	2,38	10,01	3,11	0,63	0,28	1,05
1,10	0,20	3,5	0,02	16,54	5,52	2,22	10,27	2,49	0,65	0,27	1,04
1,15	0,20	3,5	0,02	18,95	5,37	2,04	10,14	2,64	0,67	0,28	1,07
1,20	0,20	3,5	0,03	17,54	5,25	2,20	9,99	2,55	0,63	0,29	1,07
1,25	0,20	3,5	0,05	16,43	5,29	2,21	9,87	2,37	0,65	0,28	1,06
1,30	0,20	3,5	0,07	14,03	5,61	2,41	9,91	2,11	0,66	0,31	1,06
1,05	0,25	5,1	0,01	22,14	4,74	1,81	7,24	3,56	0,59	0,26	0,77
1,10	0,25	5,1	0,02	17,19	4,99	1,62	7,05	2,56	0,58	0,22	0,75
1,15	0,25	5,1	0,03	18,43	4,94	1,86	7,46	2,57	0,55	0,23	0,77
1,20	0,25	5,1	0,04	16,87	5,12	1,89	7,06	2,61	0,58	0,25	0,79
1,25	0,25	5,1	0,07	14,85	4,71	1,82	7,05	2,38	0,55	0,22	0,74
1,30	0,25	5,1	0,08	13,18	4,87	1,75	7,09	2,12	0,56	0,22	0,77
1,05	0,30	7,3	0,02	19,09	4,26	1,57	5,48	3,30	0,52	0,19	0,56
1,10	0,30	7,3	0,03	17,09	4,83	1,52	5,22	2,61	0,51	0,23	0,57
1,15	0,30	7,3	0,03	17,22	4,61	1,54	5,34	2,73	0,53	0,20	0,58
1,20	0,30	7,3	0,05	17,38	4,46	1,68	5,11	2,54	0,50	0,20	0,55
1,25	0,30	7,3	0,07	16,24	4,47	1,60	5,11	2,32	0,51	0,22	0,57
1,30	0,30	7,3	0,09	13,79	4,53	1,64	5,26	2,11	0,57	0,19	0,58
1,05	0,35	10,3	0,02	18,56	4,33	1,61	4,03	3,28	0,48	0,20	0,44
1,10	0,35	10,3	0,03	17,45	4,40	1,56	3,97	2,56	0,53	0,19	0,45
1,15	0,35	10,3	0,04	18,41	4,41	1,55	4,00	2,70	0,48	0,20	0,43
1,20	0,35	10,3	0,06	17,37	4,38	1,61	3,95	2,53	0,51	0,19	0,43
1,25	0,35	10,3	0,08	16,55	4,33	1,67	3,89	2,20	0,51	0,21	0,46
1,30	0,35	10,3	0,11	14,29	4,46	1,52	3,78	2,05	0,50	0,19	0,44
1,05	0,40	14,2	0,07	22,16	4,08	1,64	3,12	3,38	0,49	0,19	0,37
1,10	0,40	14,2	0,13	17,71	4,43	1,47	3,25	2,55	0,48	0,18	0,37
1,15	0,40	14,2	0,18	17,77	4,11	1,45	3,43	2,74	0,47	0,18	0,37
1,20	0,40	14,2	0,21	16,21	4,14	1,42	3,31	2,53	0,49	0,20	0,35
1,25	0,40	14,2	0,24	14,87	4,17	1,52	3,34	2,28	0,50	0,17	0,38
1,30	0,40	14,2	0,27	13,97	4,57	1,47	3,19	2,04	0,52	0,18	0,37

Таблица Б.8

Средние результаты применения предлагаемого метода на акциях, торгуемых на бирже NASDAQ в 2004–2006 годах при кластеризации по методу средней связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,4	0,01	10,22	9,79	5,34	24,47	1,44	1,51	1,06	2,72
1,10	0,10	1,4	0,01	11,61	10,44	5,52	24,55	1,68	1,26	0,85	2,51
1,15	0,10	1,4	0,01	11,70	9,74	5,81	23,99	1,78	1,16	0,81	2,42
1,20	0,10	1,4	0,01	13,32	9,18	6,07	24,37	1,79	1,26	0,89	2,46
1,25	0,10	1,4	0,01	12,41	9,65	6,96	25,02	1,82	1,30	0,96	2,58
1,30	0,10	1,4	0,01	10,06	10,55	6,58	25,28	1,52	1,23	0,86	2,44
1,05	0,15	1,8	0,00	11,25	6,95	3,55	18,58	1,37	0,89	0,53	1,78
1,10	0,15	1,8	0,01	10,10	7,38	3,83	17,98	1,69	0,86	0,52	1,78
1,15	0,15	1,8	0,01	11,38	7,34	3,78	18,34	1,71	0,87	0,48	1,72
1,20	0,15	1,8	0,01	12,66	7,18	3,88	17,87	1,78	0,85	0,48	1,69
1,25	0,15	1,8	0,01	11,29	7,39	4,04	18,10	1,85	0,87	0,52	1,77
1,30	0,15	1,8	0,01	11,38	7,17	3,83	17,84	1,56	0,82	0,49	1,71
1,05	0,20	2,5	0,01	8,65	5,59	2,58	13,31	1,47	0,70	0,36	1,24
1,10	0,20	2,5	0,01	10,83	5,61	2,01	12,83	1,64	0,63	0,29	1,17
1,15	0,20	2,5	0,01	11,40	5,24	1,78	12,27	1,72	0,65	0,30	1,21
1,20	0,20	2,5	0,02	12,43	5,71	2,58	12,62	1,69	0,63	0,32	1,19
1,25	0,20	2,5	0,02	12,00	5,38	2,57	12,40	1,94	0,65	0,34	1,19
1,30	0,20	2,5	0,02	10,37	5,55	2,41	12,45	1,60	0,64	0,31	1,19
1,05	0,25	3,5	0,02	9,55	4,31	1,45	9,51	1,39	0,56	0,24	0,88
1,10	0,25	3,5	0,02	9,70	4,71	1,42	9,40	1,75	0,56	0,21	0,88
1,15	0,25	3,5	0,02	10,74	4,83	1,40	9,11	1,77	0,53	0,23	0,84
1,20	0,25	3,5	0,02	12,45	4,54	1,64	8,92	1,76	0,54	0,24	0,87
1,25	0,25	3,5	0,03	12,84	4,61	1,58	9,15	1,90	0,54	0,22	0,87
1,30	0,25	3,5	0,03	11,29	4,53	1,54	8,93	1,58	0,54	0,22	0,86
1,05	0,30	4,7	0,03	9,43	4,05	1,36	7,75	1,60	0,52	0,21	0,67
1,10	0,30	4,7	0,03	11,40	4,05	1,22	7,38	1,76	0,51	0,20	0,69
1,15	0,30	4,7	0,04	12,51	4,60	1,43	7,09	1,80	0,49	0,19	0,66
1,20	0,30	4,7	0,04	10,88	4,56	1,22	7,18	1,81	0,51	0,20	0,68
1,25	0,30	4,7	0,05	11,76	4,38	1,48	6,76	1,89	0,49	0,20	0,68
1,30	0,30	4,7	0,05	11,97	4,30	1,47	6,81	1,58	0,49	0,18	0,65
1,05	0,35	6,4	0,03	13,08	4,18	1,16	5,63	1,44	0,46	0,17	0,52
1,10	0,35	6,4	0,04	9,51	3,93	1,10	5,64	1,75	0,47	0,20	0,50
1,15	0,35	6,4	0,05	12,48	4,30	1,28	5,49	1,68	0,46	0,17	0,52
1,20	0,35	6,4	0,05	12,84	3,77	1,30	5,29	1,66	0,45	0,18	0,52
1,25	0,35	6,4	0,06	11,76	4,09	1,25	5,24	1,91	0,48	0,18	0,52
1,30	0,35	6,4	0,07	9,34	4,06	1,21	5,16	1,66	0,50	0,20	0,52
1,05	0,40	8,7	0,04	8,56	3,93	1,01	4,18	1,53	0,47	0,17	0,41
1,10	0,40	8,7	0,05	10,13	3,99	1,05	3,97	1,71	0,44	0,17	0,43
1,15	0,40	8,7	0,06	10,19	3,93	1,23	3,90	1,76	0,45	0,17	0,43
1,20	0,40	8,7	0,07	11,80	3,85	1,11	4,19	1,76	0,46	0,19	0,42
1,25	0,40	8,7	0,09	12,77	3,88	1,22	4,07	1,92	0,46	0,17	0,40
1,30	0,40	8,7	0,10	11,68	3,88	1,23	3,99	1,64	0,44	0,19	0,41

Таблица Б.9

Средние результаты применения предлагаемого метода на акциях компаний из рейтинга S&P 500 в 2014–2016 годах при кластеризации по методу средней связи

				«Устаревший»				«Современный»			
R	t	E	L	T_0	T_p	T_c	T_n	T_0	T_p	T_c	T_n
1,05	0,10	1,7	0,01	1,17	1,97	0,54	6,83	0,23	0,30	0,15	0,73
1,10	0,10	1,7	0,01	1,20	2,00	0,63	6,93	0,26	0,28	0,09	0,72
1,15	0,10	1,7	0,01	1,28	2,04	0,59	7,04	0,23	0,28	0,14	0,70
1,20	0,10	1,7	0,01	1,19	2,07	0,52	6,87	0,25	0,30	0,13	0,71
1,25	0,10	1,7	0,02	1,19	2,09	0,48	6,99	0,24	0,30	0,11	0,72
1,30	0,10	1,7	0,02	1,54	1,98	0,59	6,90	0,25	0,29	0,12	0,69
1,05	0,15	2,4	0,02	1,08	1,87	0,34	5,02	0,22	0,25	0,11	0,51
1,10	0,15	2,4	0,02	1,40	1,74	0,40	5,01	0,24	0,25	0,11	0,48
1,15	0,15	2,4	0,02	1,40	1,84	0,42	4,90	0,25	0,23	0,10	0,50
1,20	0,15	2,4	0,02	1,16	1,85	0,44	5,03	0,24	0,24	0,08	0,52
1,25	0,15	2,4	0,03	1,24	1,79	0,32	5,02	0,24	0,25	0,09	0,53
1,30	0,15	2,4	0,04	1,48	1,80	0,49	4,93	0,23	0,25	0,07	0,54
1,05	0,20	3,4	0,02	1,26	1,65	0,30	3,62	0,23	0,23	0,10	0,38
1,10	0,20	3,4	0,02	1,18	1,59	0,33	3,72	0,26	0,24	0,09	0,40
1,15	0,20	3,4	0,02	1,24	1,76	0,28	3,61	0,24	0,22	0,10	0,38
1,20	0,20	3,4	0,03	1,21	1,64	0,28	3,77	0,28	0,23	0,08	0,39
1,25	0,20	3,4	0,03	1,19	1,72	0,30	3,70	0,26	0,21	0,07	0,40
1,30	0,20	3,4	0,06	1,46	1,60	0,33	3,55	0,25	0,23	0,09	0,38
1,05	0,25	5,0	0,04	1,22	1,43	0,30	2,60	0,23	0,23	0,08	0,28
1,10	0,25	5,0	0,04	1,19	1,56	0,36	2,62	0,24	0,21	0,06	0,28
1,15	0,25	5,0	0,04	1,27	1,57	0,25	2,53	0,24	0,21	0,04	0,29
1,20	0,25	5,0	0,05	1,32	1,67	0,25	2,43	0,24	0,21	0,05	0,27
1,25	0,25	5,0	0,06	1,37	1,47	0,26	2,45	0,25	0,21	0,06	0,30
1,30	0,25	5,0	0,10	1,16	1,37	0,33	2,64	0,23	0,22	0,08	0,29
1,05	0,30	7,2	0,05	1,22	1,35	0,28	1,94	0,25	0,20	0,06	0,23
1,10	0,30	7,2	0,05	1,34	1,39	0,26	1,91	0,24	0,21	0,05	0,23
1,15	0,30	7,2	0,06	1,23	1,52	0,26	1,90	0,27	0,21	0,06	0,21
1,20	0,30	7,2	0,07	1,14	1,38	0,19	1,96	0,26	0,20	0,06	0,23
1,25	0,30	7,2	0,09	1,39	1,32	0,22	1,94	0,24	0,20	0,08	0,22
1,30	0,30	7,2	0,13	1,16	1,38	0,23	1,97	0,26	0,20	0,06	0,23
1,05	0,35	10,2	0,06	1,30	1,39	0,21	1,38	0,24	0,22	0,07	0,19
1,10	0,35	10,2	0,06	1,47	1,35	0,31	1,41	0,25	0,20	0,05	0,19
1,15	0,35	10,2	0,07	1,38	1,43	0,26	1,35	0,23	0,19	0,06	0,19
1,20	0,35	10,2	0,08	1,29	1,35	0,23	1,38	0,24	0,21	0,05	0,19
1,25	0,35	10,2	0,11	1,02	1,23	0,17	1,21	0,25	0,21	0,08	0,19
1,30	0,35	–	–	–	–	–	–	–	–	–	–
1,05	0,40	14,6	0,07	1,22	1,22	0,18	0,97	0,22	0,19	0,06	0,15
1,10	0,40	14,6	0,07	1,22	1,27	0,23	1,05	0,24	0,19	0,04	0,13
1,15	0,40	14,6	0,08	1,08	1,30	0,16	0,97	0,24	0,19	0,05	0,14
1,20	0,40	14,6	0,09	1,36	1,30	0,19	1,04	0,23	0,18	0,04	0,13
1,25	0,40	14,6	0,13	1,11	1,29	0,22	0,94	0,24	0,19	0,05	0,15
1,30	0,40	–	–	–	–	–	–	–	–	–	–