# **Детальное описание данных и IT части**

Прежде чем уходить в глубокую математику хотелось бы объяснить принципы наши принципы симуляции рынка. Мы в своем исследовании рассматривали российский рынок акций как наиболее доступный в плане информации и все, о чем будет сказано ниже актуально именно относительно к нему, но большинство тезисов справедливы и к другим рынкам.

Итак, мы будем рассматривать основную торговую сессию MOEX – 10 00 – 18 45, помимо этой сессии есть еще аукционы (утренний и вечерний), вечерняя сессия и остановки торгов, их довольно сложно воспроизводить, и мы в своей работе их рассматривать не будем. К тому же маркет мейкеры обычно котируются в основную торговую сессию.

Иногда это происходит в рамках договоров маркет мейкерства с биржей. Биржа таким образом пытается дать ликвидность по неликвидным инструментам либо в неликвидное время (например, на вечерней сессии) и платит за это трейдерам. На 2021 год этим в основном занимаются трейдеры Сбербанка и БКС, их ордеры можно довольно легко обнаружить, взяв почти любую неликвидную облигацию и в нем будут примерно на 5 или 6 уровне очень большие объемы, притом, что стакан в целом пустой. Биржа предоставляет всем участникам рынка, особенно с колокацией данные в нескольких форматах:

1. Все трейды, происходящие по инструментам примерно в следующем формате:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TRADENO | SECCODE | TIME | BUYORDERNO | SELLORDERNO | PRICE | VOLUME |
| 2563532985 | VTBR | 100000 | 8663 | 30212 | 0.06966 | 100000 |

В данных указывается либо SECCODE либо IsIn инструмента по которому можно их распознать. Трейды как правило приходят пачками особенно, если какой-то крупный маркет ордер забрал большую ликвидность стакана и снял много мелких сделок.

1. Снапшот стакана и его апдейты с определенной задержкой:

Table

Description automatically generated

Полное изменение стакана на всех уровнях. Как правило такие данные очень тяжеловесные и в интерфейсе трейдеров такие данные ограничиваются несколькими уровнями, до 10. Частота примерно раз в 100 мс по ликвидным инструментам, если не применяются для HFT стратегий.

1. Best bid/ask – урезанная версия снапшота, присылающая изменения лишь по 1 уровню бида и аска. Предоставляется при урезанной подписке, может использоваться не крупными игроками.
2. Full order Log или Level III данные. Предоставляются на некоторых рынках (на московской бирже только для фьючерсов на момент 2021 года), представляют собой все ордера в обезличенном виде, приходящие на биржу по конкретным инструментам. Исторические данные такие можно купить, но в real time редко даются. Это самые низкоуровневые из доступных данных по бирже. Именно на них мы и будем проводить бектест, по ним можно полностью воспроизводить действия, проходящие на бирже. Образец данных:

Table

Description automatically generated

В отличии от трейдов здесь есть другие поля.

BUYSELL – “S” заявка ask на продажу, “B” заявка бид на покупку.

ACTION – “0” - снятие заявки, “1” - постановка заявки, “2” – трейд.

По таким данным можно воспроизвести все. Есть небольшой нюанс в этих данных, они не совсем полные. Во-первых, есть айсберг заявки по которым выставляются неправильные объемы, о них будет сказано дальше. Во-вторых, типы ордеров бывают не только постановка и снятие но и также Fill or Kill и другие специфичные типы, которые имеют свои условия, но такой информации тоже не дается. Однако эти ордера незначительно ломают симуляцию и можно их частично вычислить. О таких алгоритмах будет рассказано в части бектеста.

Важно понимать, что трейдеры выполняющие задачи маркет мейкерства часто отталкиваются от того какие данные и с какой задержкой они получают, далеко не все игроки способны себе обеспечить быстрый высококачественный поток данных. Для этого требуется очень серьезная IT база, которую далеко не все игроки способны себе обеспечить. Поэтому на этом рынке существует множество вендорской IT продукции, которая решает некоторые проблемы, но требует поддержки и часто большую плату. Например, такую продукцию предоставляет EPAM, Murex и тд.

Стоит понимать, что экономия на данном этапе может поставить крест на любом HFT и эффективной торговле. Если вдруг IT инфраструктура откажет работать в нагруженный день, как например в марте 2020 года, это может повлечь очень большие убытки. Примером таких IT проблем может быть например экономия на Position Keeping системе (обеспечивает контроль над текущей позицией трейдера), если вдруг она откажет в работе то такие модели как Inventory стратегия Avellaneda Stoikov мгновенно прекратит работу в решающий момент, и все хорошие P&L результаты получавшиеся в их работе достичь будет крайне проблематично.

Мы очень подробно останавливаемся на этом, потому что любому участнику рынка стоит понимать, что реальная торговля — это не только борьба моделей любой сложности, но и борьба обеспеченностью IT. Есть многие HFT модели, которые основываются на борьбе за скорость вплоть до наносекунд и их требуется сильно упрощать чтобы хоть как-то приблизиться к такой скорости.

Конечно, это прежде всего актуально роботам арбитражерам, которые используются большинством крупных игроков. Примерами таких алгоритмов являются известные торговля индекса против акций в нем, или скажем торговля акции против его депозитарной расписки на другой бирже. Однако для маркет мейкера стоящего узким спредом и особенно действующего агрессивно, как в моделях с RL агентами скорость тоже может быть очень критична.

Вся research IT база у нас написана на Python. В дальнейшем в эксплуатацию входят в самой простой версии для примитивных алгоритмов Excel, с подпиской на данные с биржи иногда с подпиской на Bloomberg. Для более прокачанных вещей как наши алгоритмы используется низкоуровневые языки программирования – C#, Java, C++. Инфраструктура Московской биржи, например очень сильно завязана на Windows системы и часто это бывает ограничением для некоторых алгоритмов. При написании алгоритмов мы будем частично упоминать особенности необходимые для учета при их написании, если захочется попробовать их применить.

# **Принципы проведения бектеста**

Существуют разные принципы проведения бектеста и чтобы показать преимущество нашего подхода пробежимся по тому как его проводят в других работах и какие предположения используют.

1. Бектест на Open/Close результатах дня

Наверно самый часто встречающийся подход из всех других.

Его идея довольно проста. В открытых источниках, например в yahoo Finance есть результаты торгов в начале и конце дня в формате – Open/Close/MinPrice/maxPrice. Соответственно для стратегий, ориентированных на большой промежуток времени, можно предположить, что алгоритм может торговать в промежутке min – max цены и принимает торговые решения преимущественно на такой time Series дате. Скажем алгоритм может зайти в бумагу и выйти из нее через полгода.

Либо другой подход, который предлагает множество ПО – Мы строим технические индикаторы на основе Close цен. Это может быть пересечение скользящих средних, RSI или что-то в этом роде. Чтобы увидеть такой бектест достаточно взять почти любую книгу по теханализу, очень часто там используется именно он. Само собой такой подход содержит очень сильные допущения, ограничения и имеет множество недостатков. Основные минусы выглядят так:

* Практически никак не учитывается влияние на рынок наших ордеров
* Не очень понятно какую цену использовать при торговле за день, если объем достаточно большой, то исполнять его по close price невозможно.
* Никакие Intraday события не учитываются вообще
* Разрывы между днями в цене могут сломать алгоритм
* Высокая зависимость и предвзятость в сторону того по какой цене алгоритм может торговать в течении дня
* Не понятно как учитывать Stop Loss при высокой волатильности
* Иногда неадекватная реакция на значительные падения и рост

Тем не менее такие решения бесплатны и предлагаются в большинстве торговых терминалах.

1. Бектест на свечах с различным timeframe.

Graphical user interface, chart

Description automatically generated

Является логичным продолжением первого типа. Здесь уменьшена гранулярность данных, но за низкую гранулярность часто надо доплачивать в терминалах и такие данные найти уже в открытом доступе сильно сложнее.

Такой бектест есть в trading View, его можно увидеть справа. Здесь показан пример бектеста стратегии пересечения скользящих средних на акциях аэрофлота за 2020 год. Множество подобных решений есть и в питоновских библиотеках и почти в половине статей используются такие данные. Он имеет почти все те же минусы что и предыдущий подход, хоть и менее ярко выражен. Он может подойти для простых стратегий не предполагающих быстрых перевыставлений в течение дня и узкого спреда. Здесь не требуется глубокой IT подготовки в принципе и поэтому такой подход достаточно популярен.

1. Бектест на Trade Log и/или с Level 1 Best Bid/Ask данных

Многие участники рынка, не имеющие других данных, выбирают этот подход и в статьях его можно встретить особенно где пишут про RL. Суть его заключается в следующем – Мы в каждый момент времени выставляем свои бид и аск ордера, если трейд произошел выше нашего аска или ниже нашего бида мы предполагаем что он Chart, scatter chart

Description automatically generatedпроизошел с нами и учитываем его как наш трейд. Визуализацию такого подхода я тоже здесь представлю, на примере котирования по скользящей средней +- спред. Здесь можно увидеть что трейды в таком случае происходят с заметно большей частотой и такой подход значительно выигрывает у предыдущих. Он достаточно неплох и в некоторых случаях может не сильно расходится с рынком. Но тут так же есть пару проблем:

* Полностью игнорируются текущие заявки в стакане, если учитываются только трейды.

Если учитывать еще и стакан, а такие данные есть далеко не у всех, то есть другие проблемы:

* Бектест становится очень долгим даже без алго заявок, снапшот стакана очень тяжеловесный. Лучшие решения на C++ предлагают симуляцию дневной торговли за 3 часа.
* Гранулярность в данных все равно как минимум в 300 мс, в таком случае пропадут все “мгновенные” ордера от роботов – постановка и снятие произойдут мгновенно, таких заявок очень много и их игнорирование значительно снижает качество.
* Очередь нашей заявки становится сложно отслеживать – Все ордера исполняются по принципу FIFO. То есть при равной цене ордеров первым исполнится тот который раньше пришел, этот аспект придется проигнорировать

Однако все вышеописанные проблемы могут быть неактульны по неликвидным бумагам и такой бектест будет иметь право на жизнь. Если апдейты происходят редко то работать с этим вполне можно.

1. Бектест на Full Order Log или на Level III данных.

На данный момент это самая детальная и приближенная к рынку симуляция торгов и мы в своей работе будем использовать именно ее. Такой бектест могут себе позволить только крупные игроки и мы его сможем провести только на старых данных. Рассмотрим его принципы более детально:

* Алгоритм смотрит на текущий стакан и учитывает текущие трейды и на основе них принимает решения. Происходит полное воспроизведение событий на рынке и в идеальном случае можно будет просто переключить алгоритм с симулированного рынка на реальный, и он даже не заметит разницы.
* В такой системе можно учитывать 3 уровня задержек. Первое – задержка в получении маркет данных в 3-5 мс. Этот уровень даже достаточно большой и с запасом, но он вполне подходит как для плохих сценариев. Второе задержка на отправку ордеров. Тоже можно заложить около 3 мс, так как алгоритм не мгновенно отправляет ордера. И третья актуальная скорее для крупных игроков – задержка между решением об отправке ордера и его реальная отправка в виде FIX сообщения. Задержка чисто техническая и, может быть, не актуальна для маленьких хедж фондов. Ее смысл в следующем – алгоритм существует внутри большой IT инфраструктуры банка, в которой есть свои системы бек офиса, риск контроля, position keeping и они как правило очень старые их ограничения могут быть существенными. Например, одной из самых популярных поставщиков систем бэкофиса для банка является компания Диасофт в подавляющем большинстве банков используется ее старое ПО. Их система не способна обрабатывать слишком большое число трейдов и заявок и алго без задержек в такой системе работать не сможет. Поэтому мы заложим здесь задержку в 100 мс.
* По данным FOL можно вычислить айсберг исполнения по следующей функции – если мы видим, что ордер выставился на маленький объем, а реально исполнился на гораздо больший объем, то очевидно, что это айсберг и его настоящий объем скрыт. Можно ретроспективно заменить его объем на правильный. Косвенно так же можно их вычислить по их показанному объему – 1, 10 или 100. В документации биржи часто есть ограничения на минимально показанный объем, это как правило 100 и айсберги обычно стоят с этим объемом. Эту операцию мы заранее проделали и исправили в данных айсберги.
* Так как здесь происходит очень точечный учет влияния на рынок то двигать его агрессивно все же не стоит. Потому что чем больше наше влияние на рынок тем больше вероятность что участники рынка бы на нас отреагировали в прошлом. Мы заложим в наших алго не агрессивное исполнение и будем показывать результаты, где это выполняется.
* Чтобы не было проблем с дивидендами и разрывами цен между днями мы предполагаем, что алго в конце дня закрывает все позиции по цене последнего трейда.

# **Используемые базовые алгоритмы и бенчмарки**

Прежде чем тестировать сложные алгоритмы маркет мейкинга, мы попробуем проделать бектест для простых алгоритмов и усложненного алгоритма Avellaneda Stoikov. Всего их будет 3 – Basic MM, Bollinger MM, Inventory MM. Рассмотрим каждый из них.

**Chart

Description automatically generated**Basic MM. Этот алгоритм возможно самый популярный и самый простой для целей маркет мейкинга. Его логика заключается в следующем – он считает reference price и относительно него стоит фиксированным спредом. Справедливая цена это как правило разновидность средней цены последних трейдов. Она может быть простой средней или как-то взвешенной на объем или на время. Спред может быть ориентировачно - +- 0,1% относительно справедливой цены. То есть если справедливая цена 100 и спред 0,2%. То маркет мейкер будет стоять – 99,9 бид и 100,1 аск и переставляться на каждый новый трейд с рынка. Его можно встретить под следующими именами – стратегия Полосы, Envelopes band. Его визуализация представлена справа.

**Chart

Description automatically generated**

Bollinger bands MM. Является логичным продолжением простой версии. В отличии от простого ММ учитывает стандартное отклонение цены и в случае значительных колебаний, оперативно расширяет спред. В каноничной версии он считает среднее по последним 20 тикам и стоит спредом в 2 стандартных отклонения. Количество тиков и стандартное отклонение можно брать и другим. Визуализация также справа.

Inventory MM. Алгоритм основан на идеях Avellaneda Stoikov. Он будет отличатся в деталях от каноничной версии, потому что каноничная версия содержит в себе тяжелую оптимизационную задачу, которую быстрый алго не может решать за приемлемое время. Но тем не менее основную идею он содержит. В своей интерпретации этого алго мы заложили следующее – спред, также плавающий в зависимости от стандартного отклонения; если алго набирает позицию, то он старается быстрее от нее избавиться, и наоборот если он много шортит то старается набрать позицию. Делает это за счет смещения вверх-вниз бид и аска относительно справедливой цены – например если алго много закупил акций, то он будет ставить цену ниже, тем самым стараясь продать свою позицию; если у алго нет позиции, то ведет себя как Bollinger MM;

# **Результаты оптимизации и бектеста бенчмарков**

Бенчмарк Алгоритмы достаточно легковесны, и мы можем для своего исследования провести их бектест на промежутке 5 лет. У нас есть данные с 03.2015-03.2016 и 01.2017-01.2021. Мы покажем лучшие результаты на этом промежутке после наших оптимизаций. Перед этим расскажем о принципах оптимизации гиперпараметров для 3 наших алго.

У Basic MM мы оптимизировались по количеству тиков для расчета SMA (reference price), и ширину спреда. Результаты представлены внизу. Стоить отметить стохастичность результатов, особенно в марте 2020. В какой-то степени им повезло и за счет этого периода они и вышли вперед. Вообще данные результаты хороший пример того, что Маркет мейкинг в целом скорее убыточен и поэтому биржи за него и доплачивают участникам. Простые алгоритмы на долгом промежутке убыточны.

Graphical user interface, chart

Description automatically generated

Теперь посмотрим на результаты Bollinger MM. Параметры для оптимизации такие же - количество тиков в SMA, минимальный спред, и добавится новый – количество стандартных отклонений.

Graphical user interface, application

Description automatically generated

###### Graphical user interface, text, application, email Description automatically generated

И наконец посмотрим на последнюю пока самую успешную Inventory стратегию для LKOH. Мы ее оптимизировали по разным скользящим средним (SMA, WMA, EMA), а так все как у Bollinger. Результаты получились менее стабильными, но довольно занятными.

Graphical user interface, chart

Description automatically generated

Перед тем как суммировать результаты, еще отметим, что в каждый отдельный день результаты могут быть хорошими. Но черные лебеди(март 2020) или просто время показывают, что переиграть рынок очень сложная задача. Например, посмотрим на хороший день для стратегии Bollinger:

Graphical user interface, chart, histogram

Description automatically generated

Алго хорошо сумел заработать на спреде и может появиться желание его использовать. Но есть одно, но – многие бенчмарк стратегии, которые мы перебирали и которые действовали агрессивно, почти все буквально за пару плохих дней теряли все деньги. Эти дни могли долго не наступать и это могло выстрелить скажем через 3 года, но это все равно проявлялось.

По мере усложнения алгоритма у него получалось действовать более оперативно на рыночную ситуацию.