

S&P 500 stock data - Time Series Analysis

Henrico Brum
Jorge Valverde

06 de Junho de 2019

Resumo

Nessa proposta apresentamos uma metodologia para prever movimentação no mercado financeiro usando a base de dados S&P500, que contém informações de valores de ações em um intervalo de cinco anos. Nosso objetivo principal é o de conseguir predizer da melhor maneira possível o próximo ponto de uma série temporal composta pelos últimos trinta dias do valor de abertura de uma dada companhia. Usaremos grafos de visibilidade para gerar representações das séries temporais a fim de observar comportamentos semelhantes e regressão através de Vetores de Suporte e Multi-Layer Perceptrons para identificar o próximo ponto da série. Além disso, consideramos informações auxiliares geradas através de outras informações da base de dados a fim de melhorar a eficiência da regressão.

1 Introdução

A base de dados S&P 500 contém informações de valores de ações em um intervalo de cinco anos. A base de dados contém os valores de abertura, fechamento, o maior valor obtido no dia e o mais baixo, assim como o volume de transações durante a abertura da bolsa.

Nosso objetivo principal será o de treinar um modelo preditivo que consiga prever o valor de abertura de uma determinada companhia no próximo dia usando dados anteriores. Como a base conta com dados de 500 empresas diferentes que atuam em segmentos diferentes e possuem comportamentos diferentes, temos também como objetivo buscar alternativas para identificar empresas similares (com comportamento de valorização e desvalorização similar) para usar como treinamento de um modelo que consiga melhores resultados para certos tipos de companhias.

Esperamos combinar diferentes abordagens de visualização em grafos e agrupamento para restringir conjuntos de treino que favoreçam a predição de um determinado tipo de companhia. Nosso processo precisa ser independente de domínio ou de informações externas para que possa ser replicado em qualquer base de dados de natureza semelhante.

2 Definição do Problema

A predição de valores no mercado de ações é considerada um grande desafio na área de predição de séries temporais [7]. Muito disso se deve a imprevisibilidade do mercado e do grande número de variáveis que compõem um valor de compra/venda de uma determinada companhia (tamanho da empresa, *marketing*, desastres naturais, declarações à imprensa e balanço de fim de ano). Para atingir tal meta definimos nossos objetivos como:

- Descobrir representações de dados capazes de induzir valores futuros de ações da bolsa de valores
- Utilizar grafos e medidas de similaridade para agrupar companhias com comportamentos similares na base de dados S&P 500
- Treinar modelos de regressão utilizando séries temporais de empresas similares (agrupadas anteriormente)
- Avaliar os modelos de regressão usando dados do mesmo agrupamento não contidos no conjunto de treinamento

3 Revisão da literatura

O interesse em aplicar a teoria de redes para estudar séries temporais cresceu com o sucesso da aplicação da teoria de redes em muitos campos da ciência. Segundo [4], os métodos de transformação de séries temporais em redes estão divididos em três classes: redes de proximidade, redes de visibilidade e redes de transição.

- Redes de proximidade: Nesta abordagem, as subsequências são representadas por vértices e as conexões são feitas usando alguma função de distância. Este é um dos mais utilizados e é subdividido em três categorias: redes cíclicas, redes de correlação e redes de recorrência.
 - Redes cíclicas: As redes são geradas a partir de séries pseudo-periódicas, como batimentos cardíacos e fala humana. Um ciclo é uma subsequência entre dois mínimos locais. Por exemplo, no eletrocardiograma, cada batimento cardíaco representa um ciclo. A série é dividida em m ciclos. Cada vértice representa um ciclo e as conexões entre vértices são realizadas de acordo com a força de correlação temporal ρ entre os ciclos da série. Essa correlação mede quão semelhantes são os ciclos [15].
 - Redes de correlação: A criação da rede consiste em obter todas as subsequências de tamanho l de uma série temporal. Para cada par de subsequências S_i e S_j , calcular a correlação r_{ij} entre esses segmentos. Cada segmento é representado por um vértice e uma aresta é criada entre dois vértices i e j se $r_{ij} \geq r_c$, onde $r_c \in [0, 1]$ é um parâmetro limitante [14].

- Redes de recorrência: Um espaço de fase de uma série temporal é um espaço matemático que representa todos os estados possíveis de uma série temporal. Quando esses estados são plotados em função do tempo, o caminho traçado por esses estados consecutivos é chamado de trajetória. Quando uma trajetória retorna a um ponto (ou área) que já passou por ela, então esse tempo é chamado de recorrência. Desta forma, é possível representar todos os pontos de recorrência de um sistema usando uma rede de recorrência [9].
- Redes de visibilidade: Cada observação na série é representada por um vértice e as conexões são feitas se um dado vértice puder “ver” outro. As observações da série são plotadas usando plotagem de barras e uma aresta é criada entre dois vértices i e j se for possível traçar uma linha entre os pontos da série que os representam sem cruzar por nenhuma barra. Uma vantagem desse método é que a rede é invariante para afinar transformações na série. Ao aplicar esse método a uma série periódica, a rede resultante é regular. Séries aleatórias geram redes aleatórias [8].
- Redes de transição: Dada uma série temporal $X = \{x_1, \dots, x_n\}$ ela pode ser discretizada em intervalos e representada por um conjunto de símbolos $L = \{L_1, \dots, L_K\}$. A partir destes símbolos, é possível construir um grafo ponderado dirigido completo onde os vértices são os símbolos de L e o peso entre dois símbolos L_i e L_j é definido pela probabilidade condicional $p(X_{i+1} \in L_i | X_i \in L_j)$. Diferente das outras abordagens, a topologia de rede gerada por esse método depende apenas do método de segmentação escolhido [10].

4 Metodologia

4.1 Diferenças entre *open-close* e *high-low*

Tendo como objetivo prever o preço de abertura de uma empresa no próximo dia, dado um período de tempo anterior, precisamos primeiramente averiguar quais variáveis utilizar como entrada do modelo a ser treinado. Na base de dados existem cinco valores, sendo que quatro são passíveis de análise direta, são eles: valor de abertura, valor de fechamento, valores mais alto e mais baixo durante o período de abertura da bolsa.

Buscamos descobrir se existe uma relação entre as diferenças entre os valores de abertura e fechamento, assim como na diferença entre os valores mais altos e mais baixos obtidos no dia. A expressa concatenação de todos os valores poderia prejudicar o regressor ao aumentar a dimensionalidade dos dados de entrada, o que aumenta consideravelmente o número de iterações necessárias para se obter uma regressão confiável.

Utilizamos duas abordagens para essa comparação - a primeira faz uso de *Dynamic Time Warping* (DTW), enquanto a segunda é baseada em modelagem das séries temporais com redes complexas (redes de visibilidade).

Em ambas comparamos a série temporal de uma companhia dada pela diferença entre os valores de abertura e fechamento (*open-close*) e a diferença entre os valores mais altos e baixos obtidos no dia (*high-low*).

4.1.1 Usando DTW

Diversas medidas de similaridade entre séries temporais são encontradas na literatura [13], dentre elas a DTW é uma medida de similaridade entre séries temporais que calcula a distância entre duas séries de dados [1] de uma maneira rápida e eficiente. A ideia foi gerar duas séries temporais distintas contendo a distância medida pela DTW entre as taxas de abertura e fechamento e valor mais alto e baixo no dia para observar o quanto próximos esses valores se encontram.

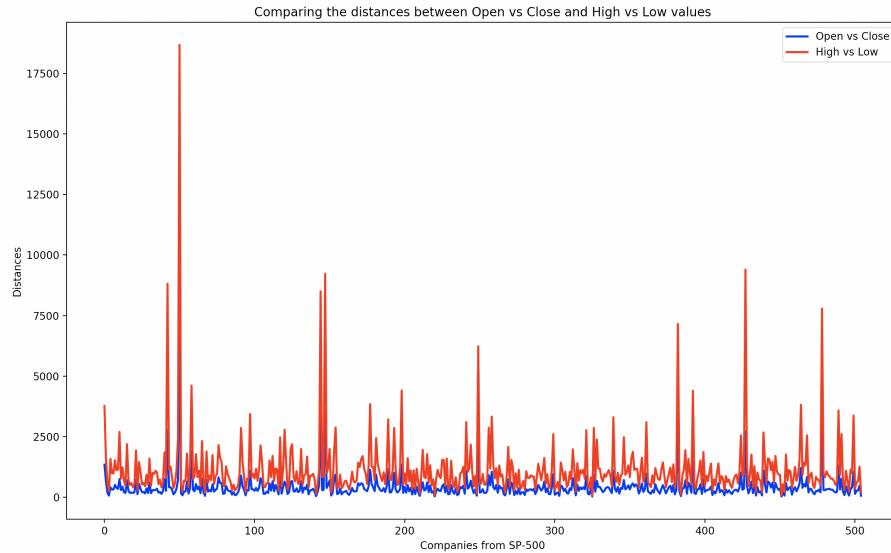


Figura 1: Comparaçao das distancias entre os preços das acções Open-Close e High-Low para cada companhia

Na Figura 1 no eixo X vemos cada uma das empresas, enquanto que no Y observamos a distância calculada da DTW em relação a um dos dois casos (*open-close* e *high-low*). A divergência entre valores de abertura e fechamento se manteve muito próxima de zero o que significa que esses dados variam muito pouco e, por conta disso, a utilização de um ou outro influencia pouco na predição do próximo valor da sequência. Na diferença entre valores mais altos e baixos no dia, podemos observar valores próximos em quase todas as companhias, exceto por alguns *outliers* que provavelmente são causados por dados erroneamente inseridos ou eventos de baixíssima ocorrência na base de dados.

4.1.2 Usando redes complexas: redes de visibilidade

Outra abordagem utilizada foi uma modelagem de redes complexas usando redes de visibilidade para identificar propriedades de séries temporais modeladas como grafos. Grafos são estruturas baseadas em Vértices (V) que representam instâncias de uma base de dados, no nosso caso valores de empresas, e Arestas (A) que conectam vértices.

Um dos desafios é encontrar uma maneira de representar séries temporais usando grafos. Para tal usamos Redes de Visibilidade [8], onde cada valor de uma série temporal é visto como um vértice e uma aresta é formada para todos os vértices em que se possa traçar uma reta que não sobreponha outro valor da série. A Figura ?? ilustra o processo de cálculo de arestas entre vértices, que também pode ser definido pela equação:

$$y_c = y_b + (y_a - y_b) \frac{(t_b - t_c)}{(t_b - t_a)} \quad (1)$$

Podemos ver na Figura 2 um exemplo do processo de criação de uma rede de visibilidade a partir de uma série temporal.

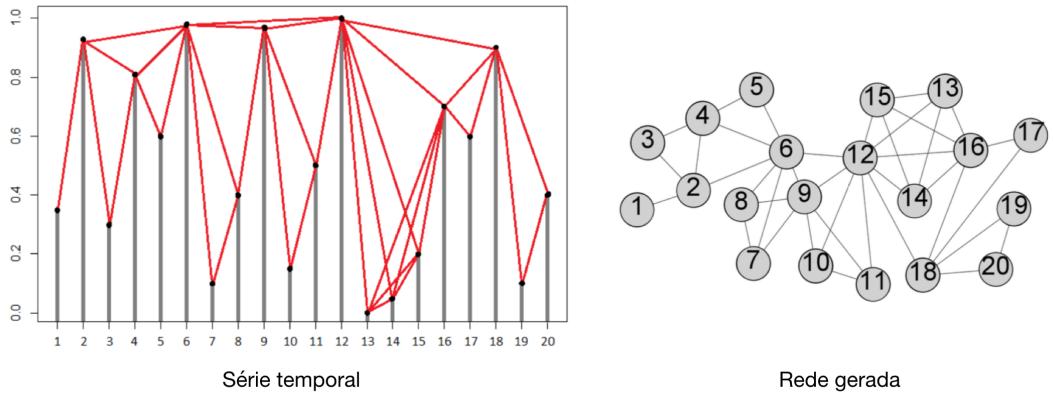


Figura 2: A primeira figura mostra um exemplo de uma série temporal a ser convertida em uma rede complexa. A série é representada como barras verticais. A segunda figura mostra a rede gerada a partir da série temporal.

Para comparar as diferenças entre abertura e fechamento, e valor mais alto e baixo diário, criamos diferentes redes para cada valor - isso fez com que tivéssemos quatro redes para cada empresa da lista, gerada a partir da série temporal de cada valor (mais de 2000 grafos). Em seguida consideramos o seguinte processo:

- Avaliar cada rede usando diferentes medidas de redes complexas (*avg. degree*, *avg. clustering coefficient*, *avg. path length*, número de comunidades e modularidade)

- As medidas de redes nos permitem determinar se dois grafos tem estruturas semelhantes ou diferentes. Por exemplo, se o *average degree* ou o *average clustering coefficient* de dois grafos é bem parecido, pode significar que duas redes têm estruturas similares .
- Então devemos usar essas medidas para comparar os grafos para os valores de *Open-Close* e *High-Low*. Testamos se os valores de cada medida de rede variam muito para cada comparação que estamos fazendo.

Na Tabela 1 mostramos os resultados obtidos de comparar as redes dos valores *Open* e *Close*, enquanto na Tabela 2 comparamos os resultados das redes dos valores *High* e *Low*. Da mesma forma, na Figura 3 comparamos os resultados obtidos de cada medida de rede (avg. degree, avg. clustering coefficient, avg. path length e número de comunidades) para os valores de Open, Close, High e Low. Os resultados mostram comportamentos bem similares para os diferentes valores dos preços das ações (seja qual for a medida de rede usada).

	Pearson corr.	Avg. Open	Avg. Close	Std. Open	Std. Close
Avg. degree	0.9746	13.0814	12.9753	2.5738	2.6695
Avg. clust. coeff.	0.8965	0.7112	0.7129	0.0132	0.0128
Avg. path length	0.8850	4.5396	4.5799	0.5270	0.5423
Communities	0.7175	16.2297	16.3366	3.9984	4.0533
Modularity	0.9766	0.7207	0.7188	0.0810	0.0832

Tabela 1: Comparando os valores Open e Close S&P 500.

	Pearson corr.	Avg. High	Avg. Low	Std. High	Std. Low
Avg. degree	0.9728	13.9078	14.0229	2.8070	3.1015
Avg. clust. coeff.	0.8680	0.7057	0.7031	0.0154	0.0147
Avg. path length	0.8568	4.4692	4.5569	0.5352	0.5524
Communities	0.6711	16.2237	16.5821	3.6472	4.0179
Modularity	0.9744	0.7201	0.7136	0.0831	0.0853

Tabela 2: Comparando os valores High e Low S&P 500.

4.2 Agrupamento dos dados

O objetivo de fazer agrupamento dos dados foi encontrar e agrupar aquelas companhias que tenham comportamentos similares. Por tanto seria interessante encontrar possíveis grupos de empresas cujo padrão de compra de ações seja semelhante ao longo do tempo. Inicialmente, pensamos que as séries temporais das empresas poderiam ser agrupadas de acordo com o setor a que elas pertencem. Os setores que pertencem às empresas são os seguintes: Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financials,

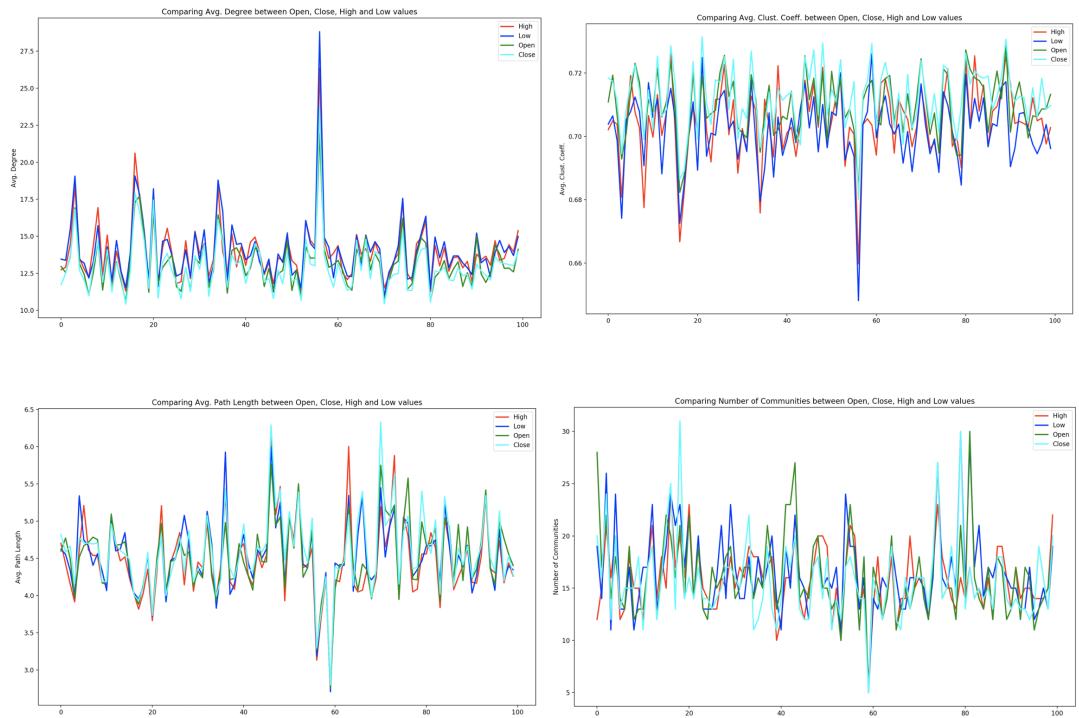


Figura 3: Comparação dos resultados obtidos para cada medida de rede para todos os valores de Open, Close, High e Low. Cada figura mostra os resultados de 100 empresas que foram escolhidas aleatoriamente.

Health Care, Industrials, Information Technology, Materials, Real State, e Utilities.

Para fazer a análise de agrupamento dos dados, propomos usar dois tipos de redes complexas: redes de visibilidade e redes de K-vizinhos mais próximos. Depois usamos o algoritmo de K-means (nas redes de visibilidade) e métodos de detecção de comunidades em redes (nas redes de k-vizinhos) para encontrar os grupos de companhias com comportamentos similares. Nesta abordagem pretendemos comparar se os grupos detectados pelos métodos de redes tem correlação alguma com os sectores que as empresas pertencem, caso contrario, fazer uma análise dos grupos encontrados.

4.2.1 Usando redes de visibilidade

O processo para fazer o analise de agrupamento usando redes de visibilidade é explicado a seguir:

- Selecionamos um grupo das redes que foram geradas na etapa anterior. Para esta abordagem usamos as redes que foram geradas para os valores de abertura.
- Usamos medidas de redes complexas para caracterizar cada companhia. Por tanto, temos um vetor representativo de cada companhia com os valores de *avg. degree*, *avg. clustering coefficient*, *avg. path length*, número de comunidades e modularidade.
- Usamos o K-means para agrupar as companhias segundo suas medidas de rede [3].
- Realizamos diferentes testes para encontrar o número ótimo de clusters. Usamos a medida de Silhouette Score para avaliar a qualidade dos clusters encontrados. Na Figura 4 mostramos a distribuição dos dados para diferentes números de clusters ($k = 5, k = 10, k = 15, k = 30$). Finalmente, na Tabela 3 mostramos os resultados obtidos da avaliação usando o Silhouette score. A tabela e as figuras dos clusters gerados mostram que os melhores agrupamentos são encontrados quando o valor de k diminui. A qualidade dos grupos piora com valores de k maiores.

Clusters	Silhouette score
5	0.4883
10	0.4293
15	0.4041
30	0.3734

Tabela 3: Resultados da avaliação da qualidade dos clusters usando o Silhouette score

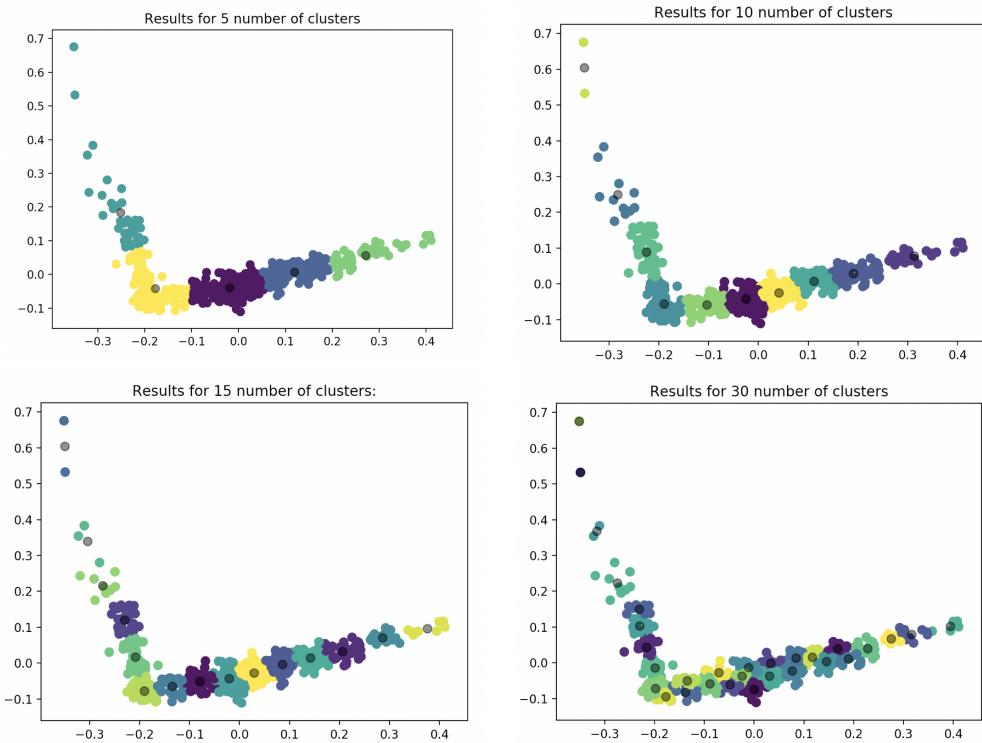


Figura 4: Análise do número de clusters encontrados usando as redes de visibilidade considerando o valor de abertura (Open)

4.2.2 Usando a rede baseada em *K*-vizinhos mais próximos (*k*-NN)

Nesta abordagem usamos novamente redes complexas, mas mudamos a forma de gerar as redes. Usamos a rede baseada em *k*-vizinhos mais próximos (KNN) [5]. Aqui geramos uma rede única que representa todas as companhias para cada valor das ações (*Open*, *Close*, *High* e *Low*). O processo de criação da rede é mostrado a seguir:

- Geramos uma única rede que representa todas as companhias.
- Cada nó está representado pela série temporal de cada companhia.
- A arista é a distância entre duas séries temporais. Escolhemos a medida de distância DTW (devido ao seu bom desempenho em eficiência e tempo em outros trabalhos)
- A criação da rede está baseada no algoritmo k-NN. Portanto os nós da rede estarão conectados aos seus k-vizinhos mais próximos.

Nesta abordagem, selecionamos os valores de abertura (*Open*) para a criação da rede k-NN. Depois, é necessário fazer vários testes para determinar o valor ideal do parâmetro *k*. Selecione os valores para *k* = 10, *k* = 20 e *k* = 50. Finalmente, o processo para fazer o análise de agrupamento usando redes do tipo k-NN é explicado a seguir:

- Geração da rede de k-vizinhos mais próximos.
- Usamos algoritmos de detecção de comunidades em redes complexas. Comunidades, ou clusters, geralmente são grupos de vértices com maior probabilidade de serem conectados uns aos outros do que para membros de outros grupos, embora outros padrões sejam possíveis [6]. Portanto, estes algoritmos permitem encontrar grupos de nós (companhias) com comportamentos similares. Escolhemos três dos algoritmos mais conhecidos para detecção de comunidades: Multilevel [2], Label propagation [11] e Infomap [12].
- Analisamos as comunidades que foram encontradas para conseguir identificar possíveis companhias com comportamentos similares.

Na Figura 5 e na Figura 6 mostramos as redes geradas pelo método dos *k* vizinhos mais próximos para *k* = 10 e *k* = 20. O primeiro grafo de cada figura representa a rede das companhias e os sectores que essas companhias pertencem. Podemos ver que os diferentes sectores estão espalhados por toda a rede e não especificamente em um único grupo. O segundo grafo de cada figura mostra como o algoritmo de detecção de comunidades divide a rede em vários grupos de companhias.

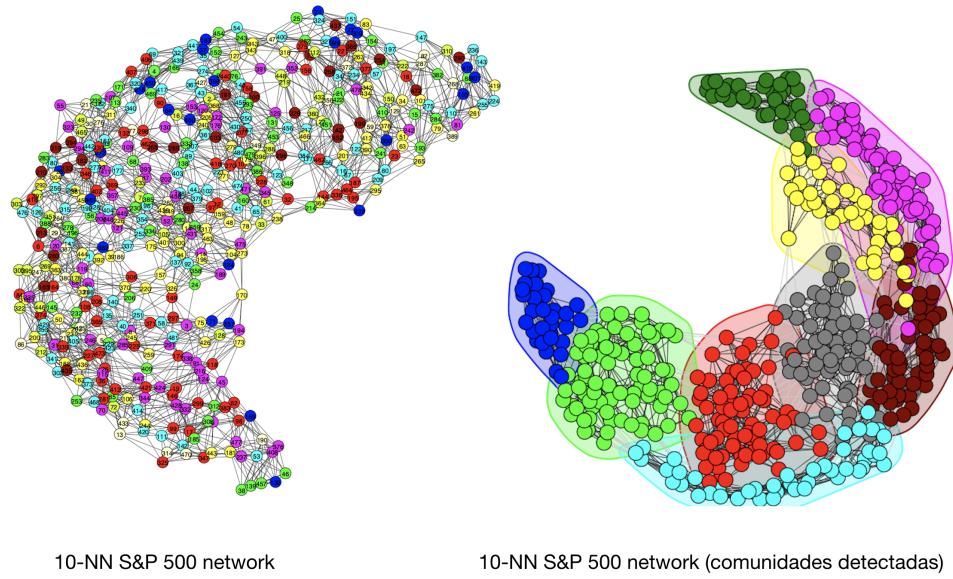


Figura 5: Rede gerada usando os k -vizinhos mais próximos ($k = 10$) para os valores de abertura. Cada nó representa a série temporal de uma companhia e as arestas são a distância (DTW) entre duas companhias. Na primeira rede, as cores dos nós representam o sectores de cada companhia. A segunda rede mostra as companhias agrupadas pelo método de detecção de comunidades (método multilevel).

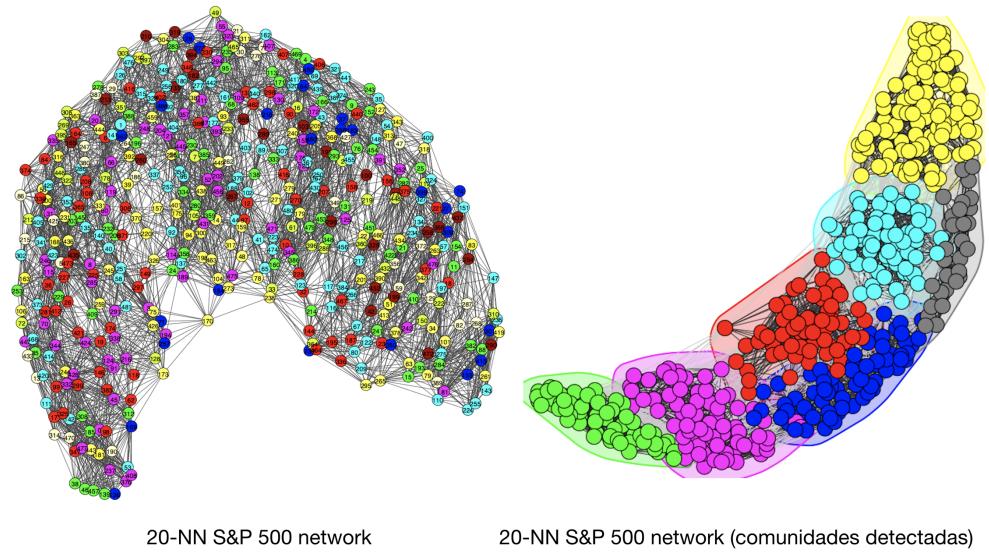


Figura 6: Rede gerada usando os k -vizinhos mais próximos ($k = 20$) para os valores de abertura. Cada nó representa a série temporal de uma companhia e as arestas são a distância (DTW) entre duas companhias. Na primeira rede, as cores dos nós representam o sectores de cada companhia. A segunda rede mostra as companhias agrupadas pelo método de detecção de comunidades (método multilevel).

4.3 Predição dos valores do próximo dia

Para avaliar o quanto bem os agrupamentos são gerados vamos fazer uma regressão de um período posterior da série buscando predizer o valor de abertura do próximo dia.

Utilizamos a abordagem *Knowledge Discover in Databases* (KDD) para estipular passos incrementais para a regressão dos valores - iniciamos pela clusterização dos dados (seguindo as abordagens apresentadas anteriormente), pré-processamos as séries temporais, treinamos os modelos de regressão e avaliamos usando três métricas que se aplicam a nossa tarefa.

4.3.1 Transformação dos dados e agrupamento

Utilizamos os *clusters* obtidos nas etapas anteriores para agrupar empresas que possuam comportamentos semelhantes. A intuição por trás dessa decisão é de que modelos treinados com séries temporais similares tendem a gerar modelos mais otimizados para determinados comportamentos. Identificados esses comportamentos podemos gerar modelos que sejam mais eficientes para empresas que tem uma variação muito alta, ou baixa, ou que tem um crescimento contínuo ou sazonais semelhantes - essas características não são definidas empiricamente, mas sim obtidas após o agrupamento descrito na Seção 4.2.

Sendo nossa tarefa de identificação de movimento em mercado de ações, os valores de cada empresa incrementam de período para período. Uma empresa pode começar com um valor de ação 15,23 e encerrar o dia variando em cerca de 1 ponto para mais ou menos. Entretanto para métodos de aprendizado de máquina consideram muito a variação dos valores (algumas empresas iniciam o dia com valor 15 enquanto em outros casos os valores permeiam grandes dezenas), então decidimos transformar nossos dados somente na variação do dia anterior. Para exemplificar essa situação podemos ver na Figura 7 duas séries temporais retiradas presentes no mesmo *cluster* que possuem grandezas diferentes.

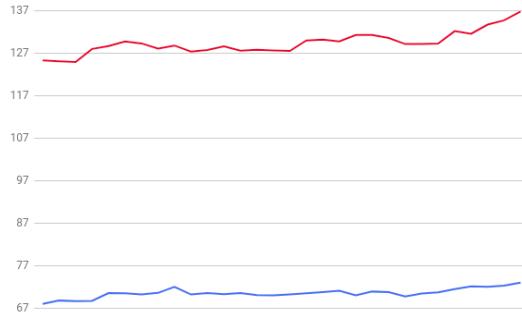


Figura 7: Duas séries temporais em escalas diferentes.

Para reduzir o problema de empresas com maiores valores que outras usamos

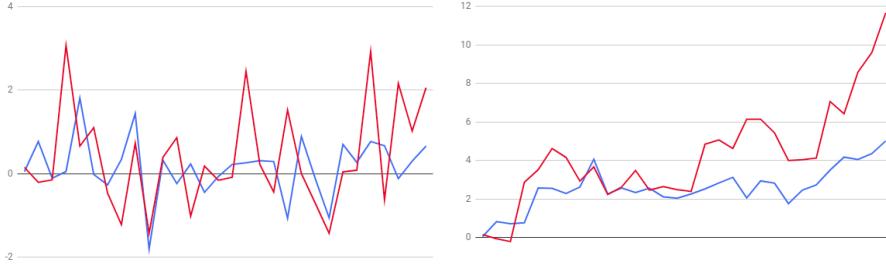


Figura 8: A figura da esquerda mostra a série temporal sendo processada com a diferença entre o momento t para $t - 1$. Na direita vemos o somatório da diferença no momento t para $t - 1$ com a diferença do momento $t - 1$ para $t - 2$. As séries processadas em ambas as imagens são as mesmas presentes na Figura 7

duas abordagens – utilizar somente a diferença dos valores em vez do valor total; e usar a diferença dos valores, porém incrementando a série temporal ao longo do tempo. Para a primeira abordagem simplesmente decidimos que o momento atual será $t = t - t_{-1}$, dessa maneira somente as variações serão usadas para prever o valor do próximo dia. O valor a ser predito também será uma variação, ou seja, valores positivos indicam valorização das ações, enquanto que negativos indicam desvalorização.

A desvantagem é que sazonalidades, como uma alta constante ou grandes variações podem não ser observadas pelos modelos. Por exemplo, valores muito positivos no final da série devem indicar previsões favoráveis, porém pequenos valores intermediários podem atrapalhar os métodos de aprendizado de máquina. Por essa razão propomos a incrementação das diferenças, ou seja, a medida que a série se aproxima do seu final, ela carrega o somatório de todos os valores anteriores. Isso faz com que grandes aumentos sejam propagados para momentos futuros das séries.

Na Figura 8 podemos observar o processamento das séries e como ele aproxima mais as séries e permite que o modelo não se confunda muito pelas grandezas. Na Seção 5 apresentaremos os resultados obtidos com cada uma das abordagens.

Outro ponto importante é como tratar dados ausentes na base de dados. Nesse trabalho sempre que identificamos uma ausência de valor atribuímos o mesmo valor do dia anterior.

4.3.2 Regressores e Avaliação

Para a regressão utilizamos dois algoritmos de aprendizado de máquina - *Support Vector Machines* (SVM) e *Multi-layer Perceptron* (MLP). Para ambos os modelos utilizamos a biblioteca *sklearn*, assim como os hiperparâmetros *default* de cada modelo implementado.

A utilização dos modelos de regressão se deve à avaliação dos agrupamentos gerados. O tamanho de *clusters* foi variado, desde 4 agrupamentos até 20.

Uma desvantagem inesperada desta base de dados foi o número de dados que impossibilitou que executássemos a regressão usando todos os *clusters*, ou até mesmo toda a base (uma execução com 5 *folds* demorou cerca de 5 dias). Para possibilitar nosso processamento retiramos sub-conjuntos aleatórios de empresas de um mesmo *cluster* em cada execução – sendo *clusters* com tamanhos inferiores ao *threshold* usados em sua totalidade. Usamos dois *thresholds* diferentes, com 5 companhias e com 10.

Além do número de empresas usadas, também realizamos experimentos com diferentes tamanhos de exemplos de treinamento, ou seja, com mais ou menos dias para prever o próximo. Nossos experimentos consideraram um mês (trinta dias) ou um decêndio (dez dias). A intuição diz que modelos podem usar mais informação com mais dias para se analisar, porém os modelos podem ter dificuldade de convergir com muitos dados.

As medidas usadas para a avaliação dos modelos foram 3 – a medida POCID, a medida *Theil's U* e o Erro Médio Quadrático (MSE). O MSE se dá por

$$MSE = \frac{1}{N} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (2)$$

Essa métrica é muito utilizada em problemas de regressão, porém para o caso específico da predição de ações de bolsa de valores ela pode não ser a ideal. O MSE mede a distância quadrática do valor obtido na regressão para o valor real da ação, enquanto que essa variação pode fazer com que uma ação positiva ($\hat{Y} > 0$) origine uma expectativa negativa ($Y < 0$), o que nesse domínio conta muito para a tarefa. É muito melhor que um modelo de regressão realize a predição do valor 2.8 ao invés de -0.22 se o valor aberto no próximo dia for de 0.80 – o que o MSE penalizaria.

Para tal utilizamos outra medida chamada POCID que mede o quanto bem um valor predito identifica a correta orientação do próximo valor na série. A equação se dá por

$$POCID = 100 \frac{\sum_{j=1}^n D_j}{N}, \text{ onde} \quad (3)$$

$$D_j = \begin{cases} 1, & \text{if } (\text{target}_j - \text{target}_{j-1})(\text{output}_j - \text{output}_{j-1}) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

A métrica retorna uma porcentagem de quantas vezes a regressão acerta a direção da série temporal, ou seja, se o valor será maior ou menor que zero. Nessa medida buscamos maximizar o resultado obtido, sendo que quanto atingimos menos de 50 na POCID isso indica que erramos a maioria das séries preditas. Além do POCID, também usamos o *Theil's U* que dá uma ideia de quanto bem um modelo consegue se comportar em relação a um *baseline* simples para séries temporais, sua equação se dá por

$$U = \sqrt{\frac{\sum_{t=1}^{n-1} \left(\frac{\hat{Y}_{t+1} - Y_{t+1}}{Y_t} \right)^2}{\sum_{t=1}^{n-1} \left(\frac{Y_{t+1} - Y_t}{Y_t} \right)^2}}. \quad (4)$$

A Equação 4 compara a saída do regressor com um *baseline* simples que considera que o valor do próximo dia será exatamente o mesmo que o do dia atual, ou seja, o último elemento do nosso vetor de dias. Visto que a variação de valores é geralmente baixa, esse *baseline* é muito difícil de ser batido. A métrica fornece um número próximo a um (“1”) que, quando inferior a esse valor indica que o método utilizado funciona melhor do que o *baseline*. Por essa razão, essa métrica busca ser minimizada em nossos experimentos.

5 Experimentos e resultados

Nossos experimentos foram feitos usando três agrupamentos previamente descritos – eles serão chamados nessa Seção de 10NN, 20NN e 30NN, cada um obteve três possíveis agrupamentos usando métodos de detecção de comunidades. Os métodos são Multi Level, Label Propagation e INFOMAP.

Como dito anteriormente, cada execução buscou um número definido de companhias (5 ou 10) de cada agrupamento gerado. Além disso os dados foram representados usando 10 ou 30 dias e um pré-processamento foi feito respeitando duas abordagens – Variação de valores e Incrementação de diferença de valores.

Um *Baseline* foi utilizado a fim de comparação com os métodos diretamente e na medida de *Theil's U*. Os dois métodos de aprendizado de máquina usados na avaliação foram o SVR (Support Vector Regressor) e o Multilayer Perceptron. Todos os experimentos foram realizados cinco vezes, usando uma parcela de 80% para treino e 20% para teste em cada execução, porém não foi feita validação cruzada, ou seja, os dados foram colhidos aleatoriamente em cada execução (nunca misturando o treino com o teste, obviamente).

As tabelas exibidas nessa Seção representam a média de todas as execuções em todos os *clusters* e os melhores resultados em cada tipo de agrupamento estão destacados em cada tabela. A única exceção são os valores obtidos com a métrica *Theil's U*, onde destacamos todos os experimentos que superaram o *baseline*.

Nas Tabelas 4, 5 e 6 podemos ver a medida de MSE para as regressões. O menor valor obtido nessa métrica foi atingido pelo *baseline* com agrupamentos gerados usando o método INFOMAP com o agrupamento 20NN. É interessante observar que o pré-processamento de Variação das diferenças obteve os melhores valores, o que contrariou nossa intuição inicial de que os métodos se beneficiariam de dados incrementais nas séries. Em suma, os melhores valores foram atingidos usando menos companhias de cada *cluster* e o número de dias não influenciou muito nessa métrica. Alguns *outliers* na regressão causaram valores de MSE próximo a 14, o que seria muito trágico para se basear em uma aplicação real.

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	7.590	8.354	2.597	1.724	4.228	2.222	6.562	2.029	2.426
	10	10	2.327	6.170	2.904	4.354	2.940	3.121	4.639	3.133	2.129
	5	30	14.697	7.845	4.147	4.178	7.972	2.061	7.318	3.617	3.261
	10	30	3.669	8.445	7.835	2.643	4.918	2.500	17.609	25.115	5.168
Incremento	5	10	4.571	4.379	3.283	15.488	6.500	2.151	6.368	8.339	2.744
	10	10	3.466	6.141	3.712	2.333	3.756	2.945	2.083	17.340	7.159
	5	30	2.562	3.666	3.487	2.230	5.539	3.943	3.025	2.189	2.944
	10	30	6.569	6.487	5.710	5.382	2.780	2.694	2.428	9.047	3.035

Tabela 4: MSE com modelo MLP

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	7.271	7.393	2.442	1.597	3.953	1.987	5.844	1.868	2.273
	10	10	2.153	5.710	2.643	3.919	2.748	2.734	4.198	2.891	2.007
	5	30	8.821	5.890	3.052	2.825	5.000	1.491	4.963	2.441	2.175
	10	30	2.918	6.533	5.269	1.980	3.645	1.761	13.844	20.814	3.638
Incremento	5	10	4.225	4.239	3.052	13.948	6.296	2.110	6.114	7.078	2.596
	10	10	3.230	5.971	3.659	2.323	3.714	2.770	2.089	16.067	7.066
	5	30	2.284	3.364	2.963	1.999	4.960	3.574	2.524	1.941	2.612
	10	30	5.669	5.978	5.525	4.531	2.590	2.454	2.299	7.425	2.810

Tabela 5: MSE com modelo SVR

Uma métrica que foi abandonada durante os experimentos foi mensurar o número de vezes em que cada regressor acertava exatamente o valor do próximo dia. Essa métrica foi esquecida ao constatarmos que nenhum dos métodos sequer acertou uma única vez o valor correto. Somente o *baseline*, por sua natureza de replicar o último valor, conseguiu acertar algumas vezes e, constatamos se tratar de valores não informados na base de dados que possuíam a mesma abordagem para valoração.

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	7.269	7.390	2.418	1.564	3.916	1.976	5.849	1.836	2.234
	10	10	2.125	5.702	2.629	3.899	2.726	2.723	4.162	2.857	1.984
	5	30	8.801	5.890	3.053	2.818	4.981	1.490	4.943	2.443	2.165
	10	30	2.925	6.554	5.270	1.991	3.660	1.777	13.858	20.815	3.655
Incremento	5	10	4.182	4.208	3.013	13.904	6.256	2.076	6.085	7.032	2.575
	10	10	3.192	5.939	3.633	2.292	3.692	2.743	2.054	16.057	7.020
	5	30	2.267	3.343	2.934	1.969	4.931	3.553	2.507	1.928	2.586
	10	30	5.646	5.970	5.506	4.509	2.568	2.435	2.274	7.412	2.792

Tabela 6: MSE com baseline

Nas Tabelas 7 e 8 encontramos as médias de resultados da métrica POCID, que identifica orientação da série temporal. Nessa medida o *baseline* não foi avaliado, pois ele sempre possui o mesmo valor e nunca altera nem positiva nem negativamente.

O SVR obteve os melhores resultados, estando sempre superior ao 50, o que

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	51.884	51.827	50.568	50.847	50.698	51.671	50.640	51.514	51.506
	10	10	52.106	51.747	51.988	51.646	51.741	52.152	51.530	51.542	52.038
	5	30	51.693	52.087	52.070	51.899	51.558	52.622	51.076	52.565	51.669
	10	30	52.358	53.990	53.230	53.083	52.790	53.336	53.404	51.954	53.132
Incremento	5	10	51.064	51.202	51.053	50.687	50.989	50.916	51.195	50.761	50.929
	10	10	51.016	51.642	51.359	51.471	51.149	51.335	51.987	51.773	51.238
	5	30	51.812	51.128	50.663	50.734	50.587	51.225	53.046	53.298	50.513
	10	30	51.315	51.980	51.307	50.899	51.670	51.426	51.243	51.211	51.694

Tabela 7: Medida POCID com modelo SVR

indica que na maioria das séries ele acerta a positividade ou negatividade do valor do próximo dia. Usando a MLP, poucos resultados estiveram acima de 51, enquanto que com SVR atingimos o valor de 53.990 na métrica.

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	50.557	50.629	50.764	50.000	50.927	51.319	50.737	49.608	50.304
	10	10	50.891	50.855	50.478	50.669	51.186	50.132	50.889	49.706	50.784
	5	30	50.269	50.370	50.887	51.307	49.656	51.074	50.506	51.954	49.502
	10	30	50.397	51.351	51.242	51.024	51.063	51.042	50.814	51.354	50.753
Incremento	5	10	50.191	50.140	50.372	50.332	50.666	50.402	50.506	50.260	51.506
	10	10	50.495	50.864	50.829	50.120	50.815	50.946	51.697	50.841	50.357
	5	30	49.987	50.237	50.223	51.411	50.747	50.519	50.684	50.651	50.611
	10	30	51.277	51.096	50.782	51.597	50.797	50.855	51.469	51.211	49.641

Tabela 8: Medida POCID com modelo MLP

O *Theil's U* compara os métodos com o *baseline* proposto. Pela baixa variação dos valores, o *baseline* de se manter o mesmo valor do dia anterior se mostrou muito difícil de ser superado. Nas Tabelas 9 e 10 podemos observar os resultados, sendo que usando MLP nenhum experimento superou os resultados do *baseline*.

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	1.006	1.006	1.008	1.011	1.008	1.005	1.010	1.008	1.011
	10	10	1.004	1.001	1.005	1.005	1.005	1.004	1.008	1.007	1.002
	5	30	1.006	1.006	1.006	1.002	1.005	1.006	1.004	0.999	1.009
	10	30	0.999	0.992	0.996	0.997	0.995	0.997	1.000	1.000	1.002
Incremento	5	10	1.013	1.015	1.010	1.011	1.012	1.013	1.010	1.015	1.012
	10	10	1.010	1.010	1.010	1.009	1.007	1.010	1.010	1.009	1.012
	5	30	1.013	1.015	1.018	1.018	1.013	1.017	1.013	1.011	1.013
	10	30	1.012	1.010	1.012	1.014	1.015	1.012	1.015	1.013	1.012

Tabela 9: Medida *Theil's U* com modelo SVR

Usando SVR, como podemos ver na Tabela 9, a utilização de mais companhias e mais dias conseguiu superar o *baseline* em quase todos os agrupamentos. Observando também a Tabela 7 vemos que essa combinação atinge bons resultados em ambas as métricas, indicando que talvez seja uma excelente aposta no

mercado de ações usar mais dados e prever séries usando mais dias.

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10	1.044	1.052	1.049	1.051	1.040	1.056	1.060	1.057	1.054
	10	10	1.044	1.044	1.047	1.051	1.041	1.045	1.045	1.040	1.036
	5	30	1.233	1.220	1.215	1.219	1.213	1.193	1.211	1.219	1.198
	10	30	1.154	1.147	1.155	1.164	1.151	1.167	1.150	1.154	1.157
Incremento	5	10	1.019	1.024	1.024	1.025	1.020	1.022	1.025	1.036	1.018
	10	10	1.026	1.016	1.015	1.014	1.011	1.018	1.011	1.032	1.010
	5	30	1.061	1.051	1.062	1.067	1.054	1.056	1.070	1.064	1.061
	10	30	1.035	1.038	1.031	1.051	1.029	1.033	1.037	1.053	1.045

Tabela 10: Medida *Theil's U* com modelo MLP

6 Discussão

Nesse trabalho apresentamos três abordagens de agrupamento para séries temporais de ações de bolsa de valores. Além disso, usamos regressores treinados para mensurar o quanto bem uma previsão do valor do próximo dia pode ser performada. Nossos melhores resultados ainda se mantiveram muito próximos do *baseline*, porém conseguimos superá-lo em alguns experimentos com SVR.

A tarefa, apesar de difícil, mostra o potencial de métodos de Aprendizado de Máquina para predizer séries temporais e variações cambiais. Obviamente a variação de bolsa de valores acontece por fatores muito mais complexos do que os apresentados na base de dados (como obras, contratações e comunicados). Nem sempre é possível se prever quando uma ação vai diminuir ou subir, porém é possível identificar padrões nesses conjuntos de dados.

Referências

- [1] BERNDT, D. J., AND CLIFFORD, J. Using dynamic time warping to find patterns in time series. In *KDD workshop* (1994), vol. 10, Seattle, WA, pp. 359–370.
- [2] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R., AND LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*, 10 (2008), P10008.
- [3] DA SILVA, V. A. F. Time series analysis based on complex networks.
- [4] DONNER, R. V., SMALL, M., DONGES, J. F., MARWAN, N., ZOU, Y., XIANG, R., AND KURTHS, J. Recurrence-based time series analysis by means of complex network methods. *International Journal of Bifurcation and Chaos* 21, 04 (2011), 1019–1046.
- [5] FERREIRA, L., AND ZHAO, L. Time series clustering via community detection in complex networks.

- [6] FORTUNATO, S., AND HRIC, D. Community detection in networks: A user guide. *Physics reports* 659 (2016), 1–44.
- [7] KIM, K.-J. Financial time series forecasting using support vector machines. *Neurocomputing* 55, 1-2 (2003), 307–319.
- [8] LACASA, L., LUQUE, B., BALLESTEROS, F., LUQUE, J., AND NUÑO, J. C. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences* 105, 13 (2008), 4972–4975.
- [9] MARWAN, N., DONGES, J. F., ZOU, Y., DONNER, R. V., AND KURTHS, J. Complex network approach for recurrence analysis of time series. *Physics Letters A* 373, 46 (2009), 4246–4254.
- [10] NICOLIS, G., CANTU, A. G., AND NICOLIS, C. Dynamical aspects of interaction networks. *International Journal of Bifurcation and Chaos* 15, 11 (2005), 3467–3480.
- [11] RAGHAVAN, U. N., ALBERT, R., AND KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.
- [12] ROSVALL, M., AND BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.
- [13] SERRÍ, J., AND ARCOS, J. L. An empirical evaluation of similarity measures for time series classification. *Know.-Based Syst.* 67 (Sept. 2014), 305–314.
- [14] YANG, Y., AND YANG, H. Complex network-based time series analysis. *Physica A: Statistical Mechanics and its Applications* 387, 5-6 (2008), 1381–1386.
- [15] ZHANG, J., AND SMALL, M. Complex network from pseudoperiodic time series: Topology versus dynamics. *Physical review letters* 96, 23 (2006), 238701.