

Time series analysis

• • •

Henrico Brum
Jorge Valverde

Roteiro

1. Introdução
 - a. Objetivos
2. Metodologia
 - a. Distância entre Séries Temporais
 - b. Redes de Visibilidade
 - c. Diferenças entre open-close e high-low
 - i. Uso de medidas de distância entre séries temporais
 - ii. Redes complexas: redes de visibilidade
 - d. Agrupamento de dados
 - i. Redes complexas: redes de visibilidade e k-Means
 - ii. Redes complexas: redes de vizinhos mais próximos (k-NN)
 - e. Processamento e representação dos dados
 - f. Regressão e métricas
3. Experimentos e Resultados
4. Conclusões

Introdução

- Base de dados S&P 500
 - Séries temporais de valores de abertura, fechamento, valor mais alto, baixo e volume
 - Existem valores ausentes na tabela
 - Existem valores mal preenchidos
 - As empresas são de diferentes segmentos, porém seus comportamentos são diferentes

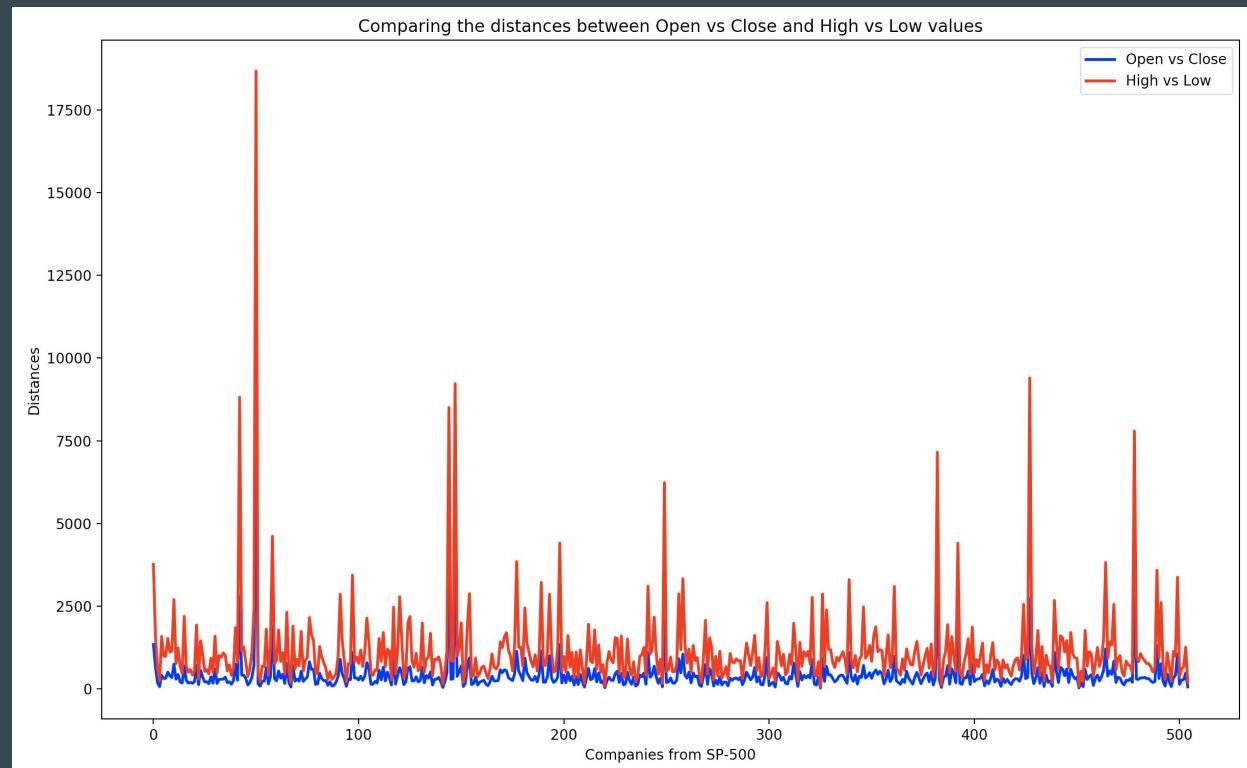
Introdução

- Objetivos
 - Utilizar grafos e medidas de similaridade para agrupar companhias com comportamentos similares
 - Descobrir representações de dados capazes de induzir valores futuros de ações da bolsa de valores
 - Treinar modelos de regressão utilizando séries temporais dos agrupamentos
 - Avaliar modelos na tarefa de predição do valor do próximo dia

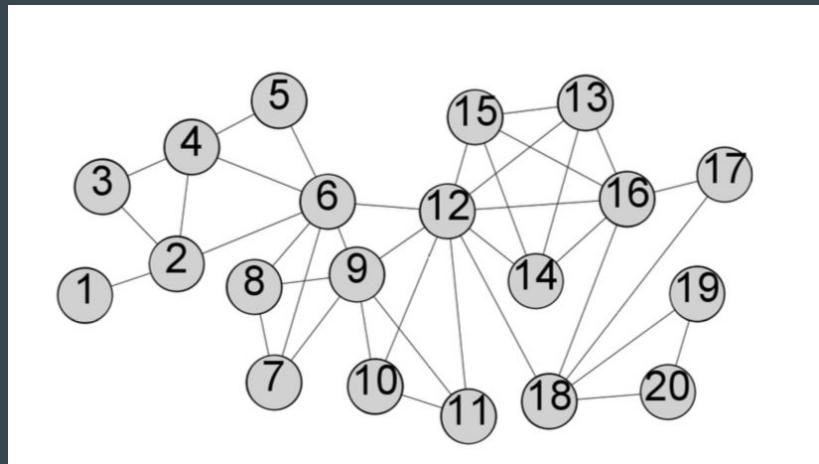
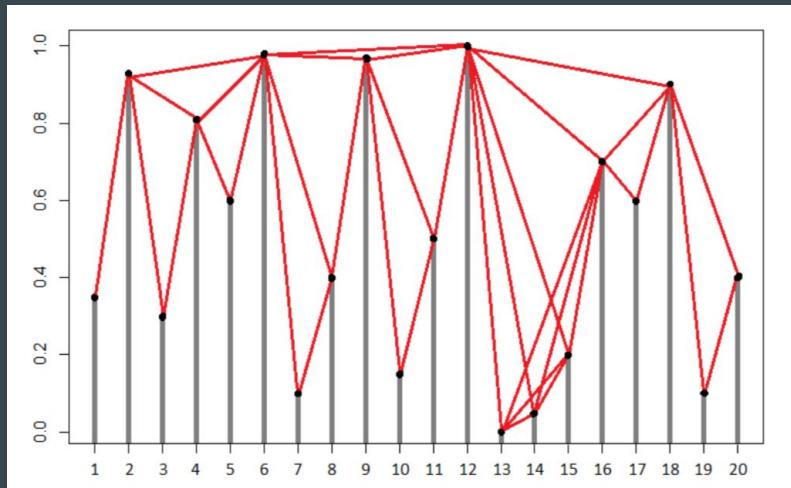
Diferenças entre open-close e high-low

- Usando Dynamic Time Warping (DTW)
- Usando redes complexas: redes de visibilidade

Diferenças entre open-close e high-low: DTW

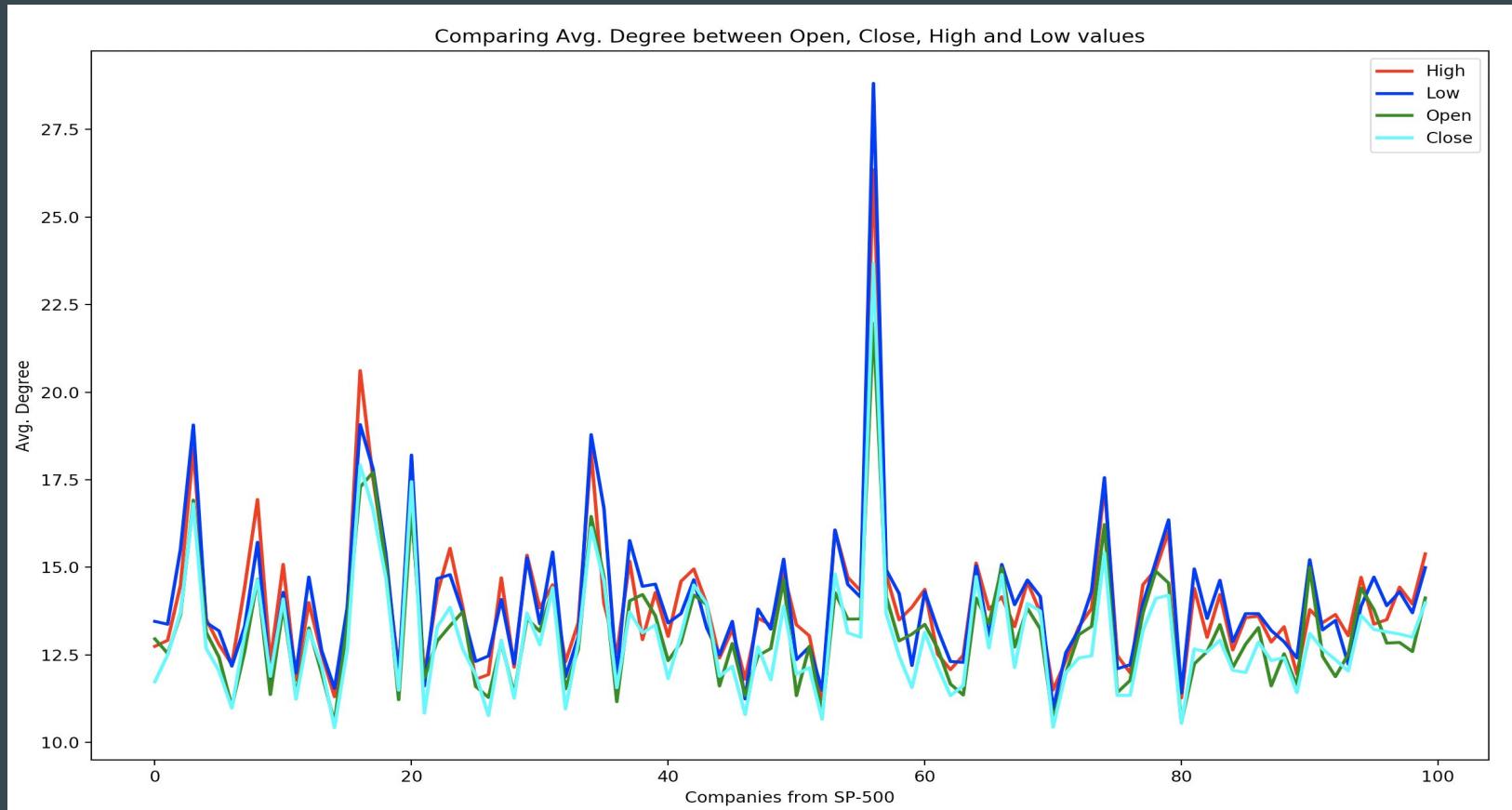


Diferenças entre open-close e high-low: Redes de visibilidade

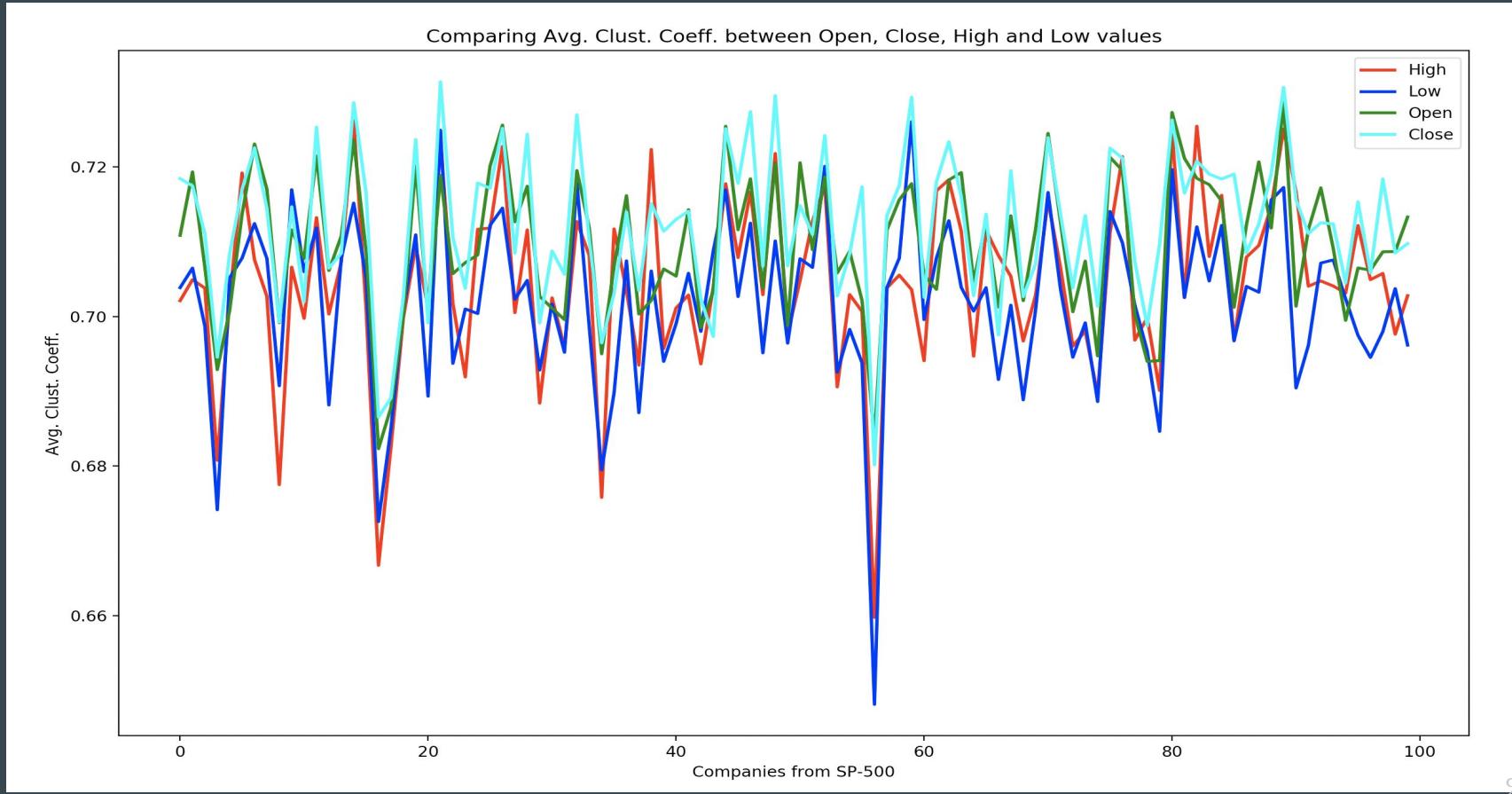


- 4 redes para cada companhia (Open, Close, High e Low)
- Medidas de rede: avg. degree, clustering coefficient, avg. path length, communities, modularity

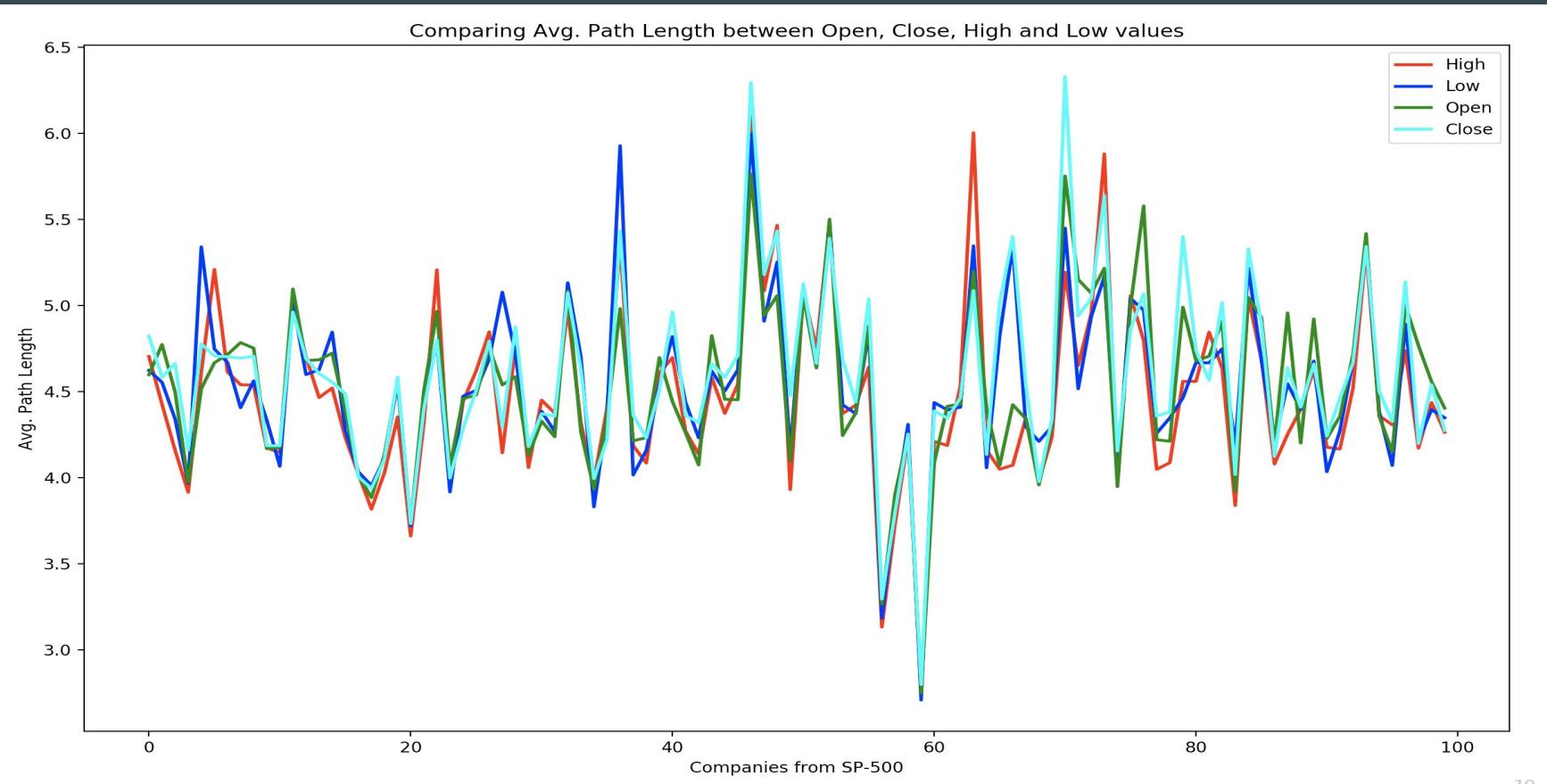
Redes de visibilidade: avg. degree



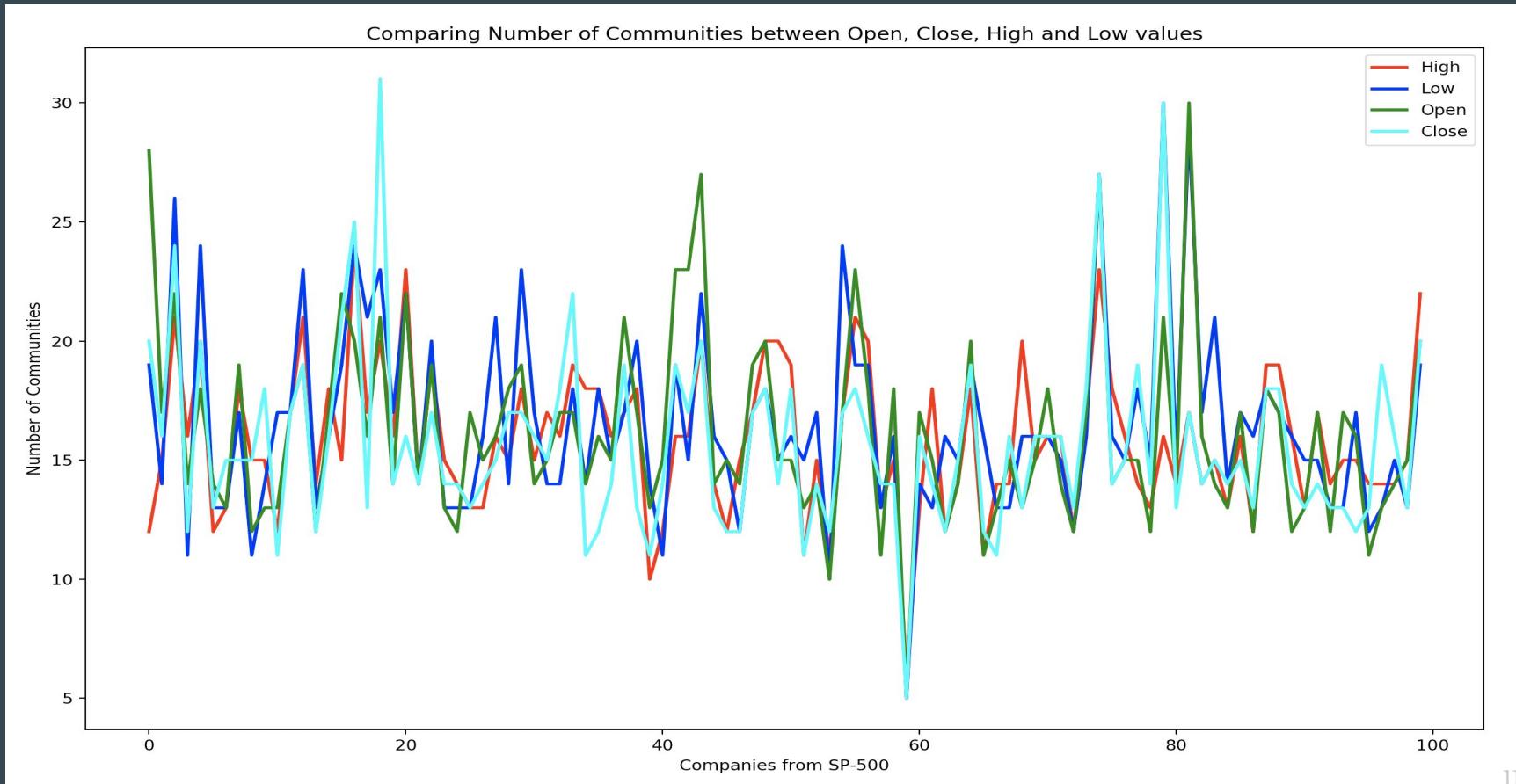
Redes de visibilidade: avg. clustering coefficient



Redes de visibilidade: avg. path length



Redes de visibilidade: communities



Redes de visibilidade: resumo

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

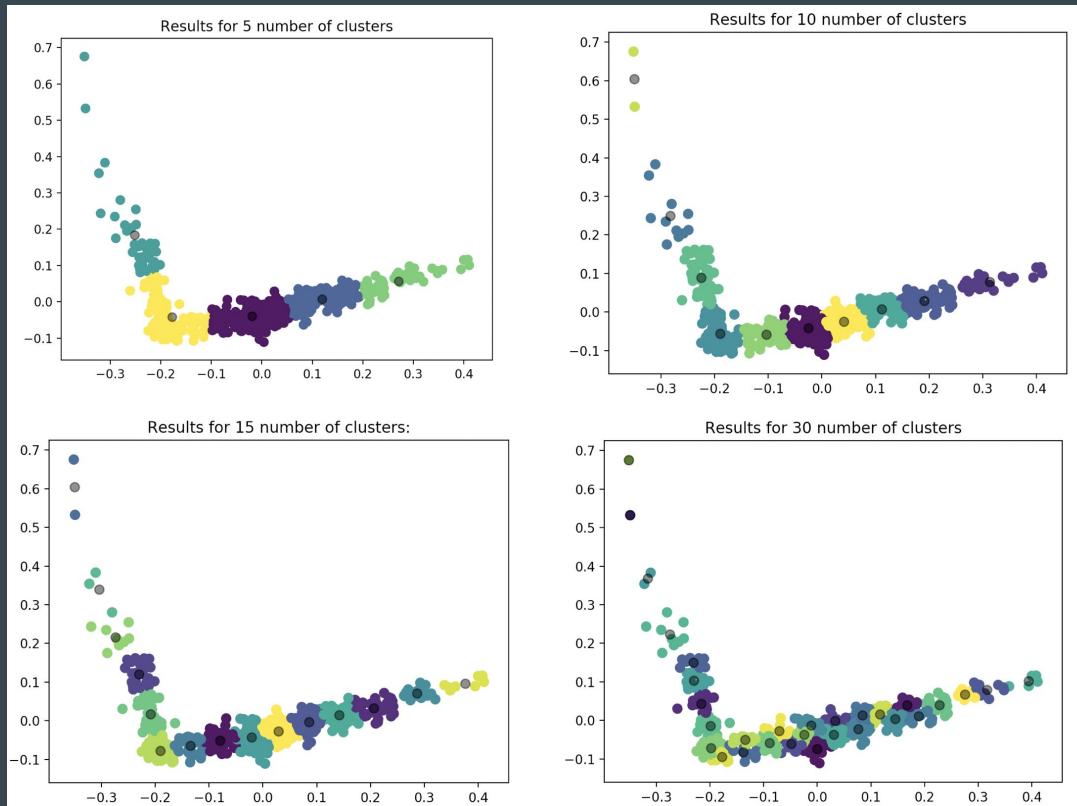
	Pearson corr.	Avg. High	Avg. Low	Std. High	Std. Low
Avg. degree	0.9728	13.9078	14.0229	2.8070	3.1015
Avg. clust. coeff.	0.8680	0.7057	0.7031	0.0154	0.0147
Avg. path length	0.8568	4.4692	4.5569	0.5352	0.5524
Communities	0.6711	16.2237	16.5821	3.6472	4.0179
Modularity	0.9744	0.7201	0.7136	0.0831	0.0853

Tabela 3: Comparando os valores High e Low S&P 500.

Agrupamento de dados

- Usando redes complexas: redes de visibilidade e k-means
- Usando redes complexas: redes de vizinhos mais próximos (k-NN) e algoritmos de deteção de comunidades

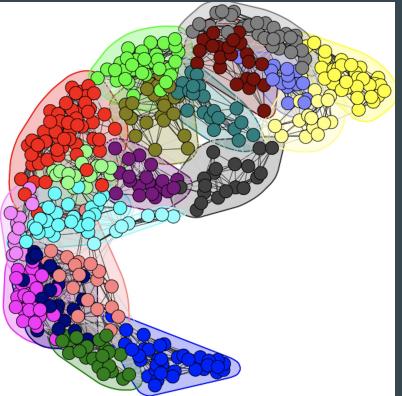
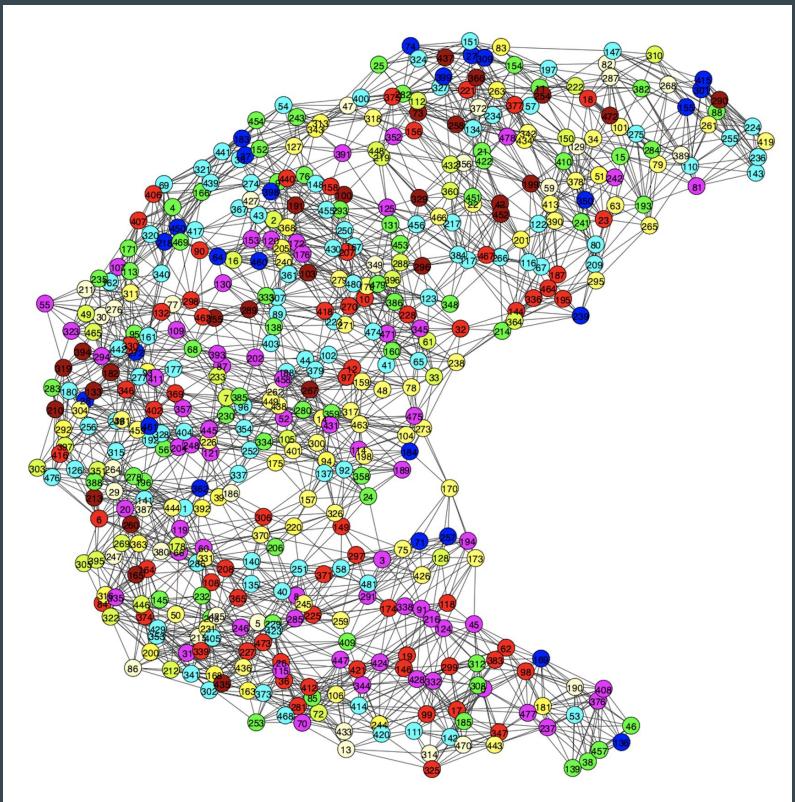
Agrupamento de dados: redes de visibilidade



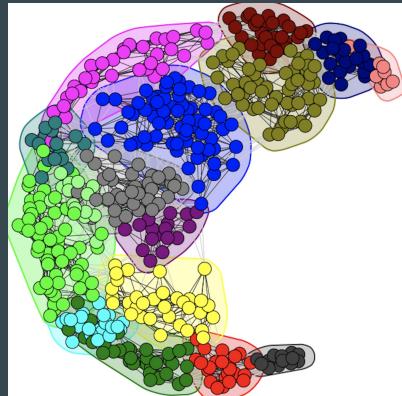
Agrupamento de dados: redes k-NN

- Nós: Série temporal de cada companhia
- Arestas: Distância (DTW) entre duas séries temporais
- Criação da rede: baseado nos k vizinhos mais próximos ($k=10, k=20, k=50$)
- Aplicamos algoritmos de detecção das comunidades nas redes geradas

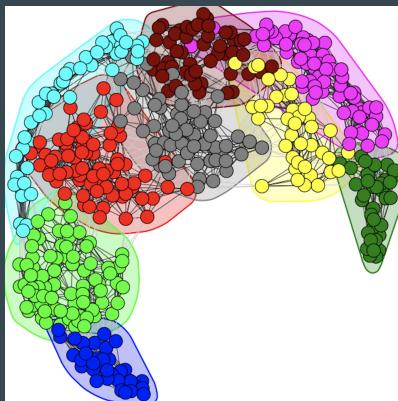
Agrupamento de dados: redes k-NN (k=10)



18 clusters

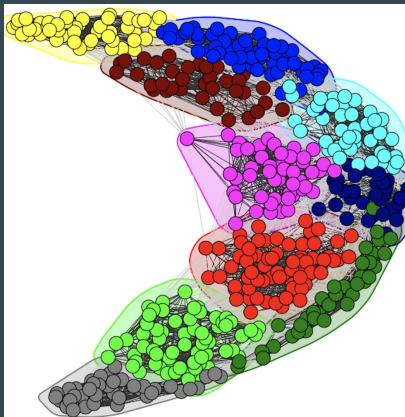
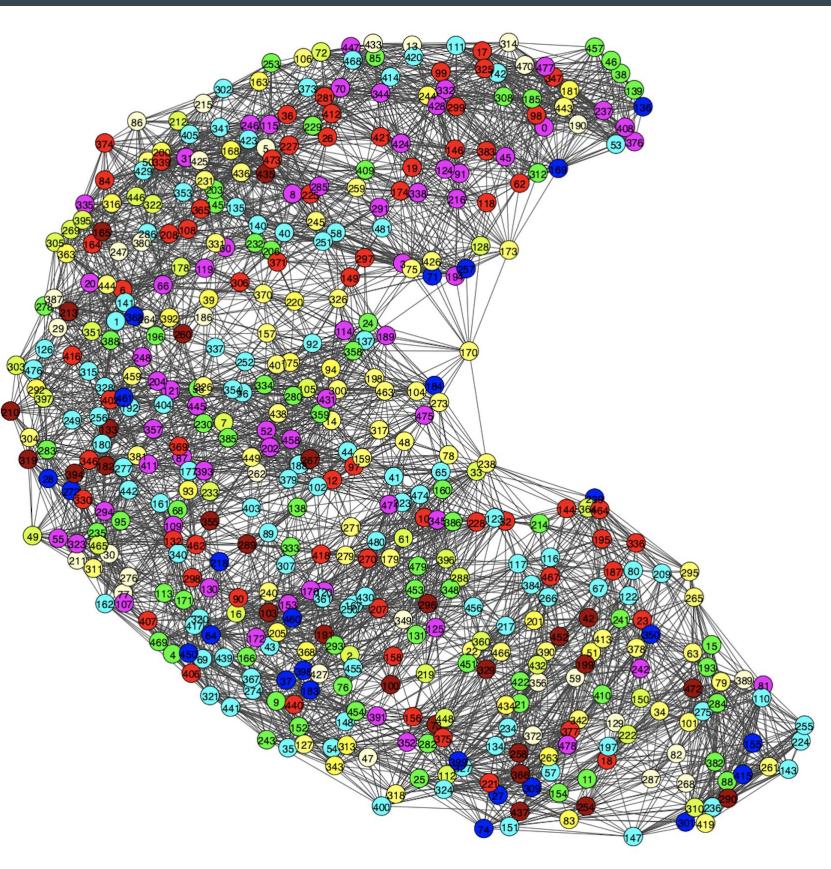


20 clusters

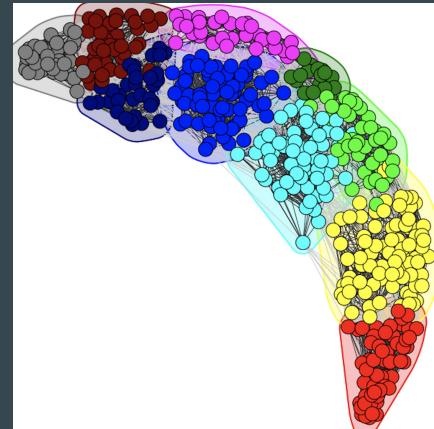


9 clusters

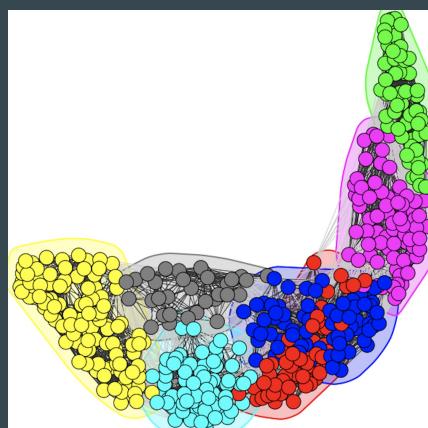
Agrupamento de dados: redes k-NN (k=20)



10 clusters

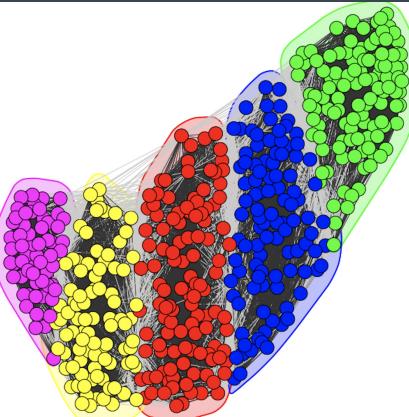
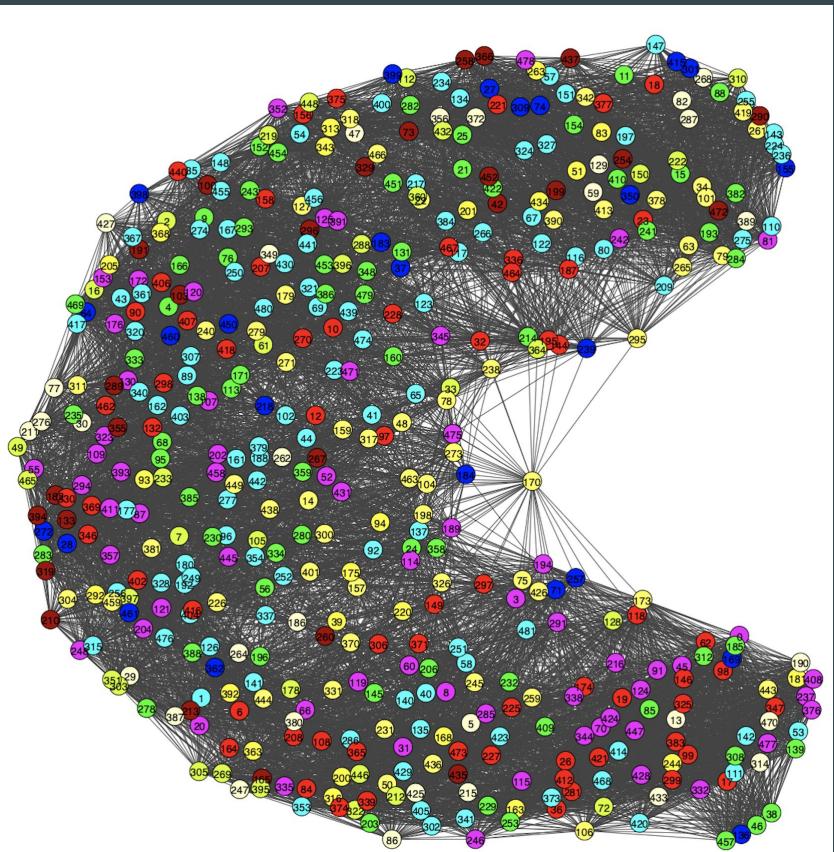


7 clusters

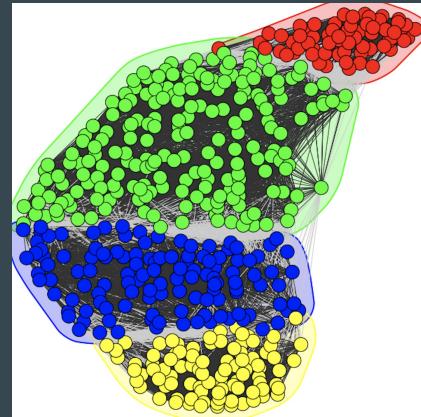


7 clusters

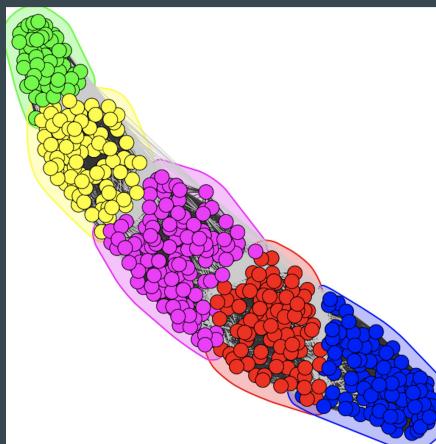
Agrupamento de dados: redes k-NN (k=50)



5 clusters



4 clusters



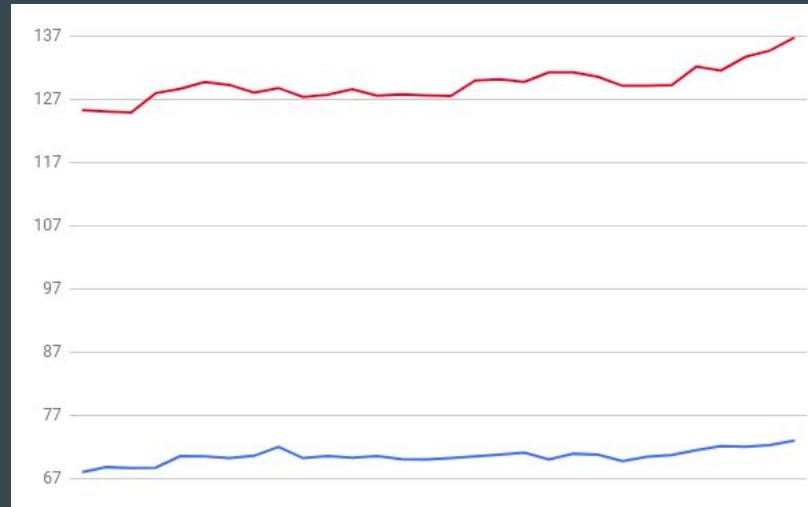
5 clusters

Processamento e Representação dos Dados

- Usar dias anteriores para prever o valor de abertura do próximo dia
- Eses valores podem variar muito de companhia para companhia e período para período

Processamento e Representação dos Dados

- Usar dias anteriores para prever o valor de abertura do próximo dia
- Esses valores podem variar muito de companhia para companhia e período para período

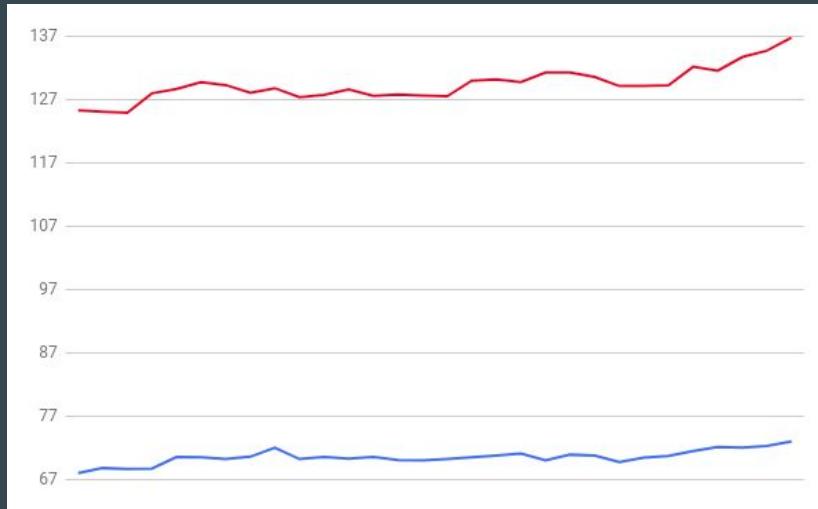


Processamento e Representação dos Dados

- Usar somente a variação (diferença) das séries
- Transparece as diferenças de grandezas entre diferentes companhias

Processamento e Representação dos Dados

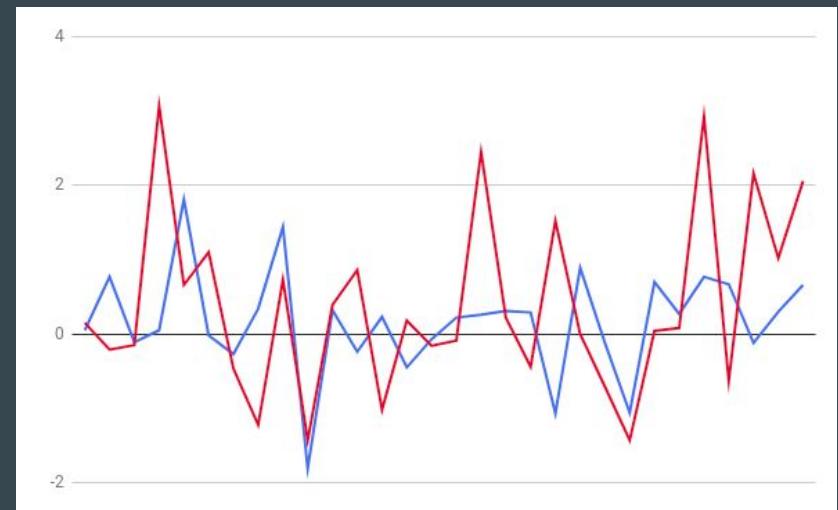
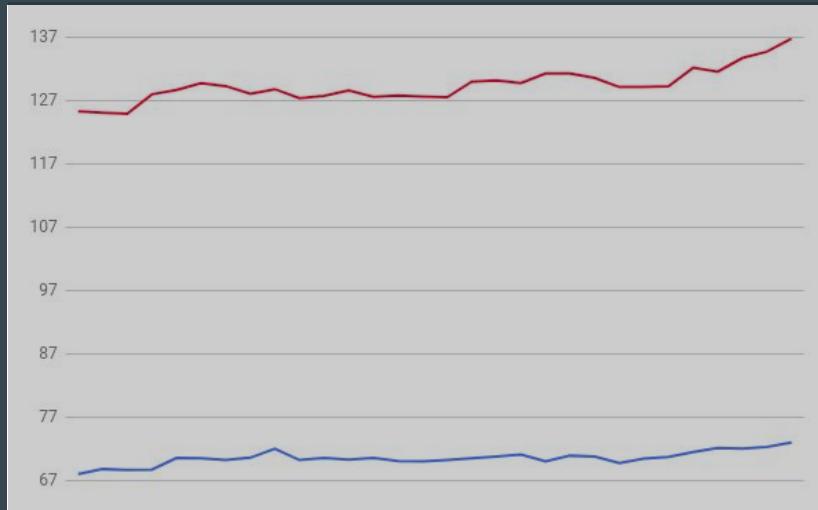
- Usar somente a variação (diferença) das séries
- Transparece as diferenças de grandezas entre diferentes companhias



Usando somente a diferença entre o valor e o dia anterior

Processamento e Representação dos Dados

- Usar somente a variação (diferença) das séries
- Transparece as diferenças de grandezas entre diferentes companhias



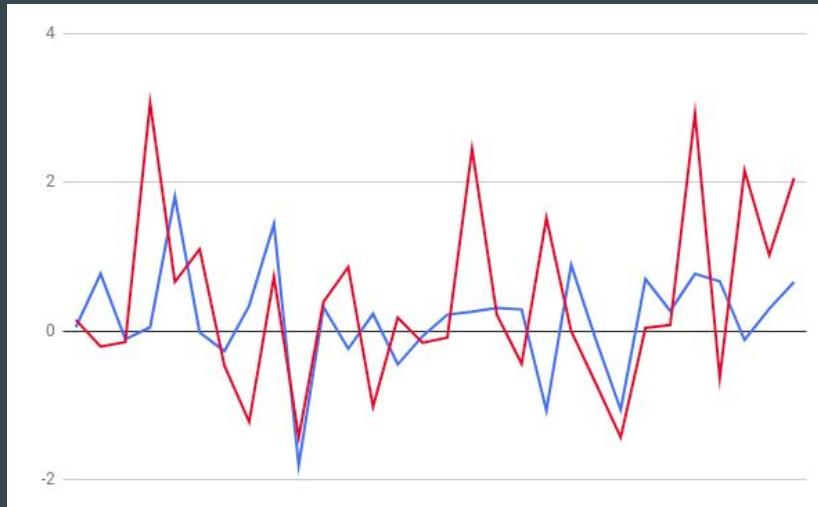
Usando somente a diferença entre o valor e o dia anterior

Processamento e Representação dos Dados

- Usar a variação (diferença) das séries de maneira incremental
- Garante que a continuidade da série seja percebida

Processamento e Representação dos Dados

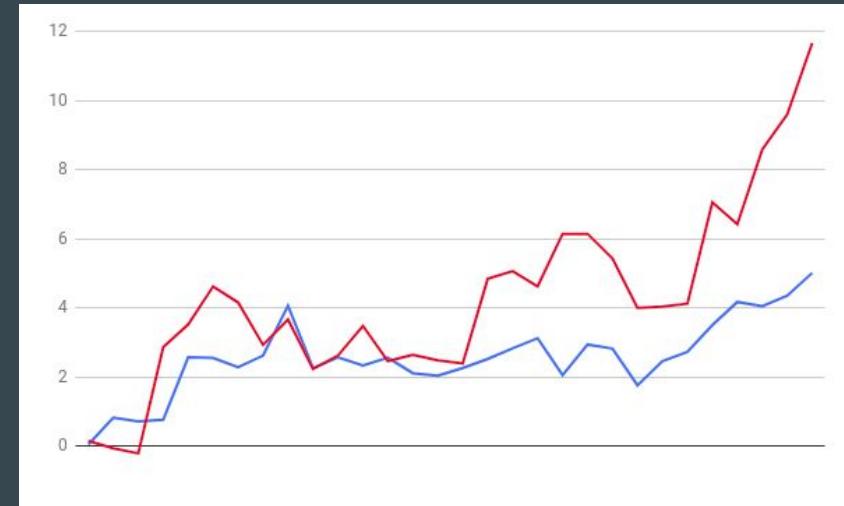
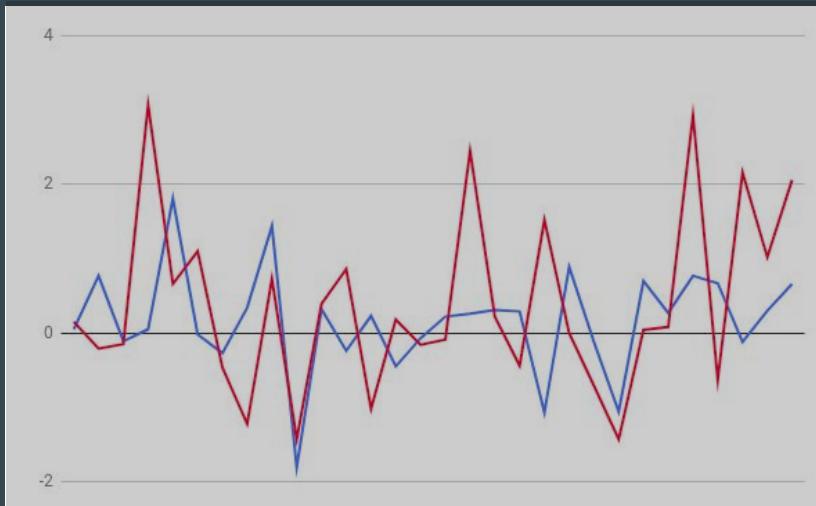
- Usar a variação (diferença) das séries de maneira incremental
- Garante que a continuidade seja percebida



Usando a diferença entre o valor e o dia anterior de maneira incremental

Processamento e Representação dos Dados

- Usar a variação (diferença) das séries de maneira incremental
- Garante que a continuidade seja percebida



Usando a diferença entre o valor e o dia anterior de maneira incremental

Processamento e Representação dos Dados

- Representamos os dados usando os dias anteriores à previsão
 - Variamos os dados usando um decêndio (10 dias) ou um mês (30 dias)
- Dados ausentes na base de dados foram substituídos pelo menor valor do dia anterior

Regressão e Métricas

- Utilizamos SVR e MLP para realizar a regressão
- Aliado a isso usamos um *baseline* para comparar os resultados
 - O *baseline* consiste em prever o próximo ponto com o mesmo valor do último dia da série

Regressão e Métricas

- As métricas usadas foram:
 - **MSE** - erro quadrático médio
 - **POCID** - medida de tendência da série
 - **THEIL'S U** - coeficiente de incerteza
 - **Acertos exatos** - número de acertos exatos dos modelos

Regressão e Métricas

- As métricas usadas foram:
 - **MSE** - erro quadrático médio
 - **POCID** - medida de tendência da série
 - **THEIL'S U** - coeficiente de incerteza

Resultados

- Três agrupamentos diferentes, com três métodos de comunidades
 - 10NN, 20NN e 50NN ~ MLV, LP e INFO
- Duas abordagens de processamento dos dados
 - Por variação da diferença e com incremento dessa variação
- Dois tamanhos de representação
 - Usando 10 dias e 30 dias

Resultados

- Cluster de tamanhos variados (de 4 companhias por *cluster* até 20)
 - Isso aumentou drasticamente o tempo de processamento
- Escolhemos um subconjunto aleatório com as companhias de cada *cluster*
 - Fizemos experimentos com 5 companhias e 10 companhias

Resultados

| 10NN 20NN 50NN

Resultados

10NN			20NN			50NN		
MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO

Resultados

Pré-proc.	10NN			20NN			50NN		
	MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO

Variação

Incremento

Resultados

Pré-proc.	Comp.	Dias	10NN			20NN			50NN		
			MLV	LP	INFO	MLV	LP	INFO	MLV	LP	INFO
Variação	5	10									
	10	10									
	5	30									
	10	30									
Incremento	5	10									
	10	10									
	5	30									
	10	30									

Resultados

- Avaliações com MSE
 - Com SVR

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

Resultados

- Avaliações com MSE
 - Com MLP

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

Resultados

- Avaliações com MSE
 - Com Baseline

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

Resultados

- Avaliações com POCID
 - Com SVR

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

Resultados

- Avaliações com POCID
 - Com MLP

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

Resultados

- Avaliações com Theil's U
 - Com SVR

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

Resultados

- Avaliações com Theil's U
 - Com MLP

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

Conclusões

- Dificilmente conseguimos superar o *baseline* da tarefa
- Os melhores resultados obtidos foram usando SVR com um número maior de dias para análise e mais companhias de cada *cluster*
- Os *clusters* gerados com MLV dificilmente obtiveram resultados promissores
- A divisão em *clusters* menores (10NN e 20NN) atingiram melhores resultados que usando menos *clusters* com mais companhias

Conclusões

- Dificilmente conseguimos superar o *baseline* da tarefa
- Os melhores resultados obtidos foram usando SVR com um número maior de dias para análise e mais companhias de cada *cluster*
- Os *clusters* gerados com MLV dificilmente obtiveram resultados promissores
- A divisão em *clusters* menores (10NN e 20NN) atingiram melhores resultados que usando menos *clusters* com mais companhias
- Nenhum de nós ficou rico durante a produção deste trabalho. :(

Time series analysis

• • •

Henrico Brum
Jorge Valverde