Recuperação Inteligente De Informações

02 - Análise Exploratória

Alexandre Herrero Matias Lucas da Silva Nolasco
Nicolas Abril
Departamento Acadêmico de Informática,
Universidade Tecnológica Federal do Paraná (UTFPR)

1 Tema

Título: Impacto da extinção das espécies.

A interação dos seres humanos com a natureza no período mais recente da história ajudou a contribuir para o desgaste dos recursos naturais e depredação de ambientes silvestres. Isso, por sua vez, tem apresentado diversas consequências, indo desde as alterações climáticas ao redor do globo até a extinção de algumas espécies. No entanto, apesar desse impacto já ser muito prejudicial, ele pode ainda ser potencializado. Isso acontece pelas interações que essas espécies têm em seu habitat natural, de forma que a extinção de um determinado grupo pode desequilibrar o ambiente onde eles comumente seriam encontrados e, consequentemente, também trazer prejuízos para outras espécies. Como existe uma grande quantidade de animais em risco após todos esses anos de uma relação predatória dos seres humanos com a natureza, seria interessante descobrir quais as espécies que podem causar o maior impacto à natureza caso seja extinta, para dessa forma descobrir a melhor maneira de concentrar os esforços para a preservação do meio ambiente.

Dentro desse cenário, esse trabalho tem como principal objetivo avaliar os impactos da extinção das espécies para assim identificar as mais vulneráveis. Para tal, serão utilizadas as informações disponibilizadas na Red List da IUCN e as interações fornecidas pela Global Biotic Interactions.

2 Equipe

Nome da equipe: Data Fund for Nature

A equipe é composta pelos seguintes membros:

- Alexandre Herrero Matias
- Lucas da Silva Nolasco
- Nicolas Abril

3 Obtenção e processamento de dados

Para o desenvolvimento desse projeto foram escolhidas duas bases de dados distintas, onde uma delas é a responsável por fornecer os dados da situação da espécie em termos de extinção e a outra fornece os dados das interações entre os diferentes grupos de seres vivos.

3.1 IUCN Red List

A União Interacional para Conservação da Natureza (IUCN) é um órgão fundado em 1964 responsável pela lista vermelha de animais em perigo de extinção, também conhecida como IUCN Red List, a qual conta com dados de uma grande diversidade de espécies, cobrindo desde animais e plantas até fungos. Por conta disso, a IUCN Red List se coloca como um indicador da vida e biodiversidade no mundo.

Para utilizar esses dados nas análises realizadas, o primeiro passo foi baixar essas informações diretamente da página da IUCN. Lá foram obtidas duas tabelas, uma contendo as informações sobre o estado das espécies com relação ao risco de extinção e outra contando com as localizações onde essas espécies podem ser encontradas ao redor do globo. Assim, com os dados em mãos, o passo seguinte foi carregá-los para se ter uma ideia das informações disponíveis. Para facilitar, algumas colunas que não seriam utilizadas em ambas as tabelas foram descartadas, como a versão do critério utilizado, ano e língua de publicação. Assim, para a tabela que indica o nível de extinção das espécies algumas das colunas mantidas foram: o nome científico da espécie, a categoria deles na Red List (indica o nível de ameaça de extinção sofrida por aquela espécie), a tendência da população (crescimento, estabilidade ou queda), se a espécie já estava possivelmente extinta e se ela está possivelmente extinta somente na natureza. O resultado pode ser visto na Tabela 1.

Já para a tabela contendo os dados de localização desses animais que fazem parte da Red List, também foram descartadas algumas colunas que não seriam utilizadas, mantendo somente as colunas com o nome científico da espécie, a latitude e a longitude. No entanto, havia animais que possuíam mais de um registro sobre a localização onde podia ser encontrado. Para resolver isso, as coordenadas foram agrupadas formando uma lista para cada animal. Uma outra alternativa seria tirar a média desses pontos, porém isso poderia distorcer um pouco a informação

Tabela 1: Exemplo da tabela resultante contendo os dados sobre a extinção das espécies.

scientificName	redlistCategory	population Trend	systems	realm	possibly Extinct	possiblyExtinct InTheWild
Heosemys annandalii	Critically Endangered	Decreasing	Terrestrial	Indomalayan	False	False
Hubbsina turneri	Critically Endangered	Decreasing	Freshwater	Neotropical	False	False
...	...	...	...	...	...	...
Hungerfordia pelewensis	Endangered	Unknown	Terrestrial	Oceanian	False	False
Ictalurus mexicanus	Vulnerable	Unknown	Freshwater	Neotropical	False	False

para espécies com duas coordenadas distantes. Por fim, o resultado foi o exemplo mostrado na Tabela 2.

Com isso, para obter a tabela final com os dados da Red List para todos os animais, ambas as tabelas foram combinadas com base no nome científico das espécies. Por fim, a tabela final foi salva em disco para possibilitar as análises que foram feitas posteriormente.

Tabela 2: Exemplo da tabela resultante contendo os dados sobre localização das espécies.

binomial	longitude	latitude
Abantis bicolor	[31.44849968, 31.7266674, 29.51413918, 31.72666...	[-28.89819527, -28.83063889, -31.607584, -28.83...
Abarema callejasii	[-75.57151548, -75.49367, -75.0166667, -75.0166...	[6.92677292, 7.06056, 6.0166667, 6.9, 7.1, 4.73...
Abarema josephi	[-75.67611, -75.5477557, -75.3636, -75.6625, -7...	[1.85694444, 6.1603885, 6.1719, 6.2325, 6.10972...
...	...	...
Abatia mexicana	[-98.040278, -97.02715158, -97.285, -97.220555,...	[20.124444, 19.430496, 19.841389, 19.830833, 19...
Abrahamia deflexa	[46.17361, 45.25, 47.233333, 45.78333, 46.81666...	[-15.71028, -16.83333, -15.066667, -15.93333, -...

3.2 Global Biotic Interactions

Já para os dados de interações entre as espécies, foi utilizada a base de dados disponibilizada pela Global Biotic Interactions (GloBI). Essa base é de acesso público e foi criada a partir da combinação de múltiplas outras fontes de dados também públicas. Os dados podem ser obtidos por meio de uma API, porém também há uma versão consolidada em um arquivo csv para pessoas que acabem precisando de uma grande quantidade para que assim elas evitem sobrecarregar os servidores com chamadas à API. Pela característica desse trabalho, optou-se pela segunda alternativa, utilizando o arquivo com todas as interações. O problema encontrado nesse caso foi o tamanho desse arquivo (um pouco mais de 14 GB), o que dificultava a sua manipulação. Para resolver isso, foi utilizada a ferramenta `miller` para descartar algumas colunas que não seriam utilizadas, como a fonte de onde elas foram extraídas e o autor do trabalho que disponibilizou essa interação.

Após isso, finalmente foi possível manipular os dados contidos no arquivo csv, do qual

foram removidas algumas outras colunas, mantendo somente as que indicavam o nome científico e o reino da espécie que executa a ação da interação, nome científico e o reino da espécie que sofre a ação da interação e o nome da ação que caracteriza a interação. Em seguida, foram removidas interações duplicadas, ou seja, foram descartadas todas as ações que possuíam a mesma espécie fonte, a mesma espécie alvo e a mesma ação. Por fim, foram descartadas as linhas onde alguma das espécies envolvidas estivesse marcada como `unidentified`. O resultado foi uma tabela semelhante ao exemplo apresentado na Tabela 3.

Tabela 3: Exemplo da tabela resultante contendo interações entre as espécies.

sourceTaxon SpeciesName	sourceTaxon KingdomName	interaction TypeName	targetTaxon SpeciesName	targetTaxon KingdomName
Andrena milwaukeensis	Animalia	visitsFlowersOf	Zizia aurea	Plantae
Andrena mandibularis	Animalia	visitsFlowersOf	Zanthoxylum americanum	Plantae
Andrena edwardsi	Animalia	visitsFlowersOf	Wyethia mollis	Plantae
...	...	...	...	...
Andrena mandibularis	Animalia	visitsFlowersOf	Viburnum dentatum	Plantae
Andrena milwaukeensis	Animalia	visitsFlowersOf	Viburnum lentago	Plantae

4 Cobertura e distribuição dos dados

Uma das informações úteis para se ter uma ideia mais clara da distribuição dos dados utilizados é ver onde as espécies listadas se localizam no mapa. Nesse sentido, a Figura 1 apresenta um mapa de calor com a distribuição das espécies apresentadas na Red List gerado a partir das informações de latitude e longitude listadas na base de dados. Nela é possível perceber que há uma boa distribuição de espécies ao redor do mundo, com exceção de alguns pontos como o Canadá, a Rússia, a Groenlândia e Antártica, regiões caracterizadas pelo clima mais frio. Apesar disso, como a base conta com espécies em uma grande diversidade de regiões, esse não será um aspecto que causará problemas.

Outra informação importante presente no conjunto de dados da Red List é a tendência de crescimento da população das espécies listadas. Pelo caráter dessa base de dados, é de se esperar que a maior parte das espécies presentes esteja em uma situação de queda, visto que é uma lista focada em espécies em risco de extinção. No entanto, seria interessante ter alguns exemplos fora dessa condição, o que possibilitaria comparações futuras caso surja a necessidade durante as análises. Como apresenta a Figura 2, os dados se organizam de maneira bem próxima ao esperado, com a parte majoritária das espécies apresentando uma tendência de queda nas

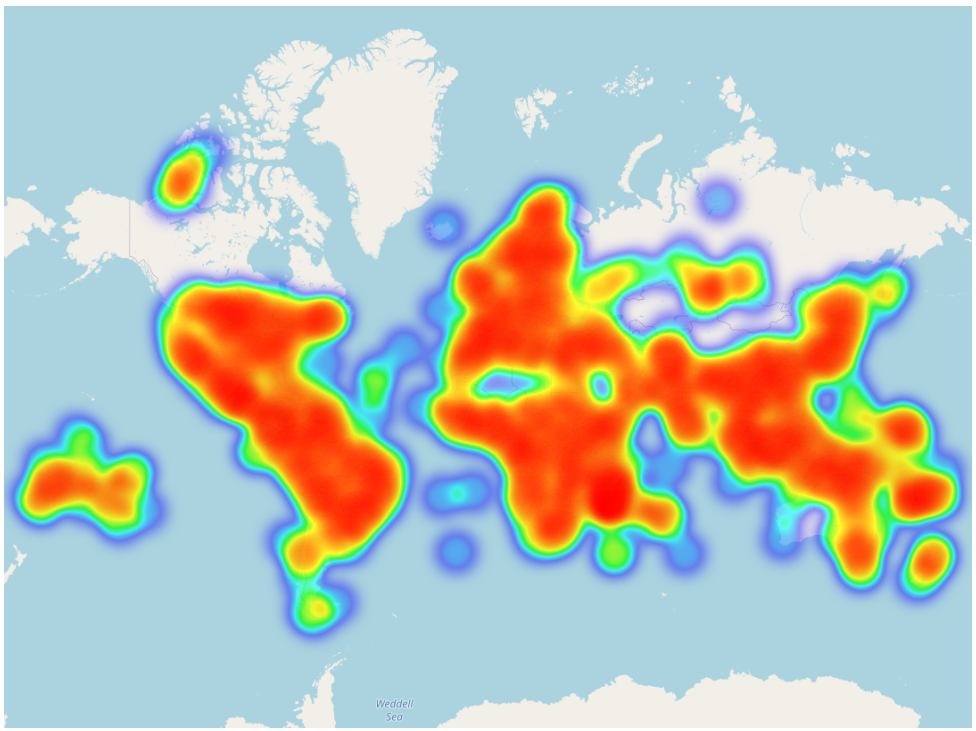

Figura 1: Mapa de calor contendo a distribuição ao redor do mundo dos animais presentes na Red List.

suas populações. Por outro lado, há ainda uma parcela considerável da qual a tendência da população é desconhecida. De qualquer forma, o ponto positivo fica para a porção de espécies em situação estável ou de aumento, o que, como explicado anteriormente, pode permitir análises de um ponto de vista diferente.

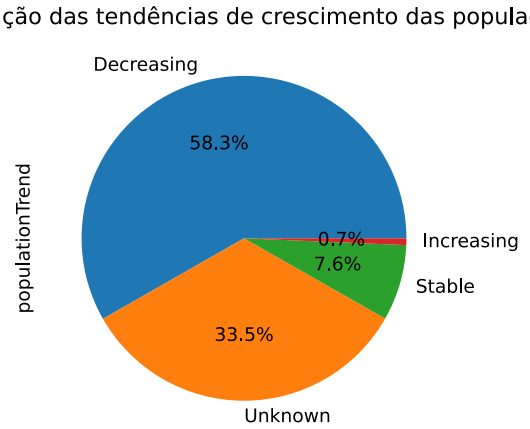

Figura 2: Distribuição das tendências de crescimento das populações.

Ainda nesse aspecto da Red List, além de saber a tendência da população, é importante saber o risco de extinção para uma determinada espécie para que futuramente seja possível identificar as espécies que podem causar maior impacto no caso de serem extintas. Novamente,

pelo caráter da base de dados, o esperado é que a maior parte das espécies listadas estejam em algum nível de vulnerabilidade, o que é confirmado pela distribuição apresentada na Figura 3. No gráfico apresentado fica claro que a maior parte dos grupos está em perigo ou em situação de vulnerabilidade, e que a grande maioria enfrenta algum nível de risco.

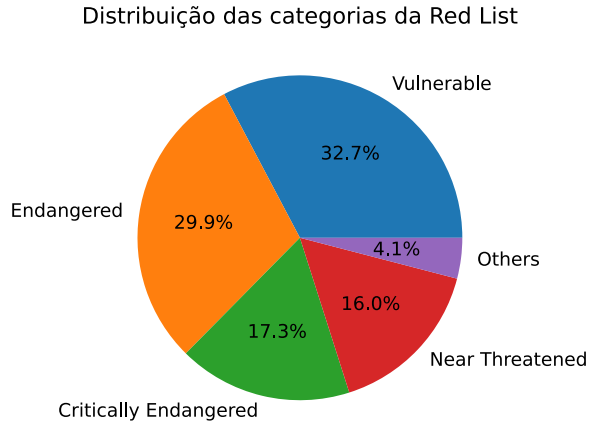

Figura 3: Distribuição das categorias da Red List.

Já com relação às interações entre as espécies, a Figura 4 apresenta as ações mais comuns observadas na base de dados utilizadas. O primeiro lugar disparado fica com a ação de comer, o que faz sentido visto que é uma ação que a maioria das espécies deve executar. Em seguida vem uma interação simples, mas a surpresa fica para a terceira colocação dessa lista, com a interação de hospedeiro. Essa ação em conjunto com a patógeno pode indicar a existência de espécies como vírus ou outros seres responsáveis por causas de doenças, indicando que a base realmente tem uma vasta quantidade de informações, não ficando limitada somente aos animais.

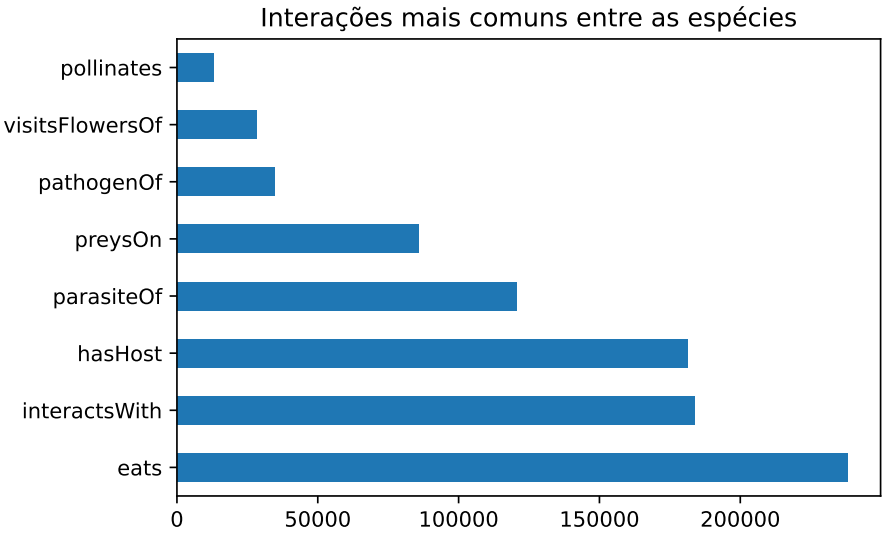

Figura 4: Distribuição das interações mais comuns entre as espécies

Essa hipótese se confirma ao olhar a distribuição do reino das espécies presentes na lista de interações, apresentada na Figura 5. Nela fica claro que a base é dominada por espécies de plantas, animais e fungos. No entanto, há ainda espécies do reino Orthornavirae, Bacteria e Protozoa, o que pode justificar a existência de interações como a "hospedeiro"observadas anteriormente. Isso mostra que, apesar de ser focada nas espécies mais comuns, essa base de dados traz uma variedade de outros grupos, o que pode enriquecer as análises.

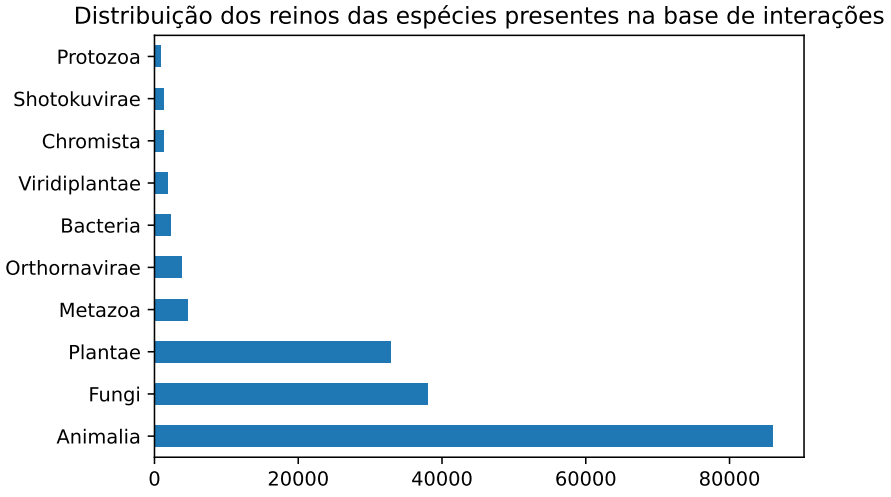

Figura 5: Distribuição dos reinos das espécies presentes na base de dados das interações

5 Análise Exploratória

Para esta atividade foram realizadas dez principais análises exploratórias dos dados obtidos de ambas as fontes. Para isso foram criados diversos gráficos de distribuição, gráficos de ocorrências além de um grafo de interações.

5.1 Distribuição dos Reinos animais que atuam como fonte de interações

A primeira análise realizada foi verificar a distribuição dos reinos dos seres vivos que atuam como fontes de interação. Para isto foi criado um gráfico contando a quantidade de de serves vivos de cada reino considerados como fontes de interação.

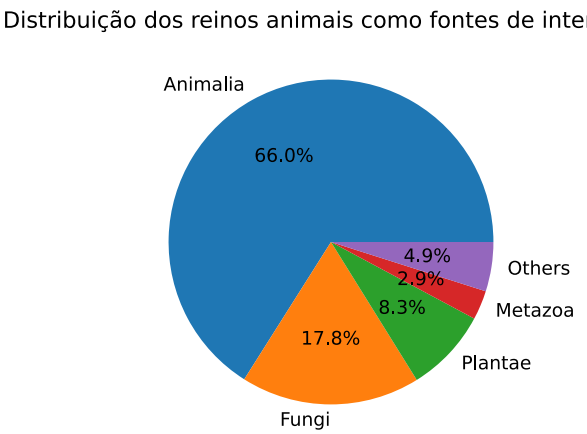

Figura 6: Gráfico das distribuições dos Reinos por quantidade de fontes de interação

O gráfico revela que a maior parte das ações das interações são realizadas pelas espécies do reino Animalia, seguido por Fungi e Plantae, além disso a população da maior parte das espécies listadas está em queda, com somente uma pequena parcela apresentando algum aspecto de crescimento ou estabilidade.

5.2 Distribuição dos Reinos animais que atuam como alvo de interações

A segunda análise realizada, de forma análoga a primeira, foi verificar a distribuição dos reinos dos seres vivos que atuam como alvos de interação. Para isto foi criado um gráfico contando a quantidade de de serves vivos de cada reino considerados como alvos de interação.

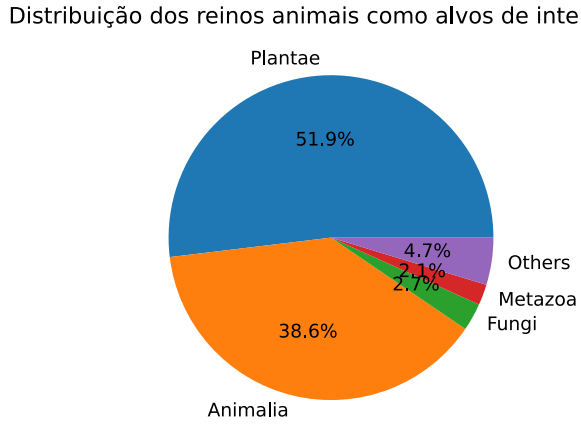

Figura 7: Gráfico das distribuições dos Reinos por quantidade de alvos de interação

O gráfico revela que, diferentemente da terceira análise, o reino Plantae toma a dianteira como maior alvo de de interação, seguido pelo Animalia e pelo Fungi.

5.3 Ocorrência de uma espécie como fonte de uma interação

Nesta análise foi verificado qual a espacie que apresenta a maior de quantidade de aparições como uma fonte de interação. Para isto foi construído um gráfico contendo as dez espécies que apresentam os maiores valores.

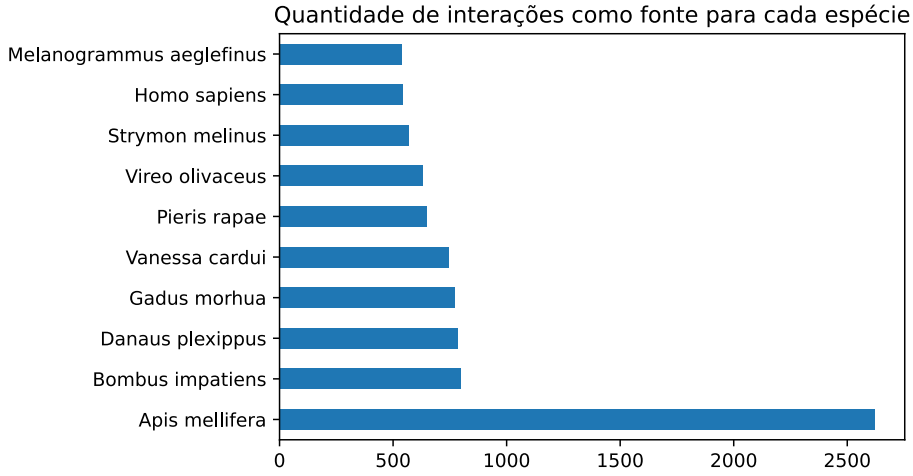

Figura 8: Gráfico das ocorrências de uma especie como fonte de interação

O gráfico revela que, diferente do que se imagina, o ser humano não se apresenta como maior fonte de interações, mas ocupa apenas a 9ª posição, A primeira posição é ocupada de forma disparada pela Apis mellifera, a espécie de abelha mais comum do mundo.

Para entender melhor como as abelhas conseguiram essa primeira posição de forma tão disparada, foram analisadas quais são as interações que têm as abelhas como fonte da ação. Para isso foi construído outro gráfico, este contendo a contagem das interações realizadas pela espécie Apis mellifera.

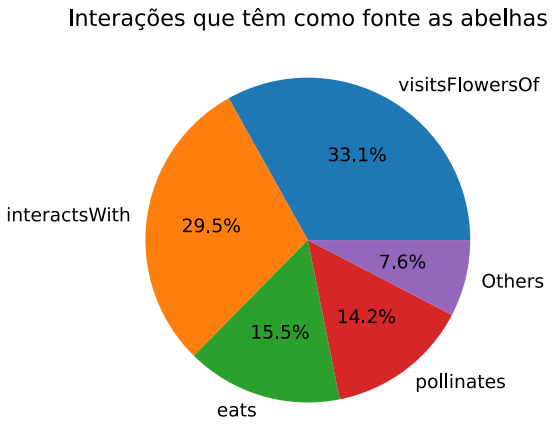

Figura 9: Gráfico da distribuição das interações da espécie Apis mellifera

Como mostram os dados, a interação mais realizada pelas abelhas é visitar flores, algo que condiz com o que se esperaria dessa espécie.

5.4 Ocorrência de uma espécie como alvo de uma interação

Nesta análise, de maneira análoga a anterior, foi verificado qual a espécie que apresenta a maior de quantidade de aparições como uma alvo de interação. Para isto foi construído um gráfico contendo as dez espécies que apresentam os maiores valores.

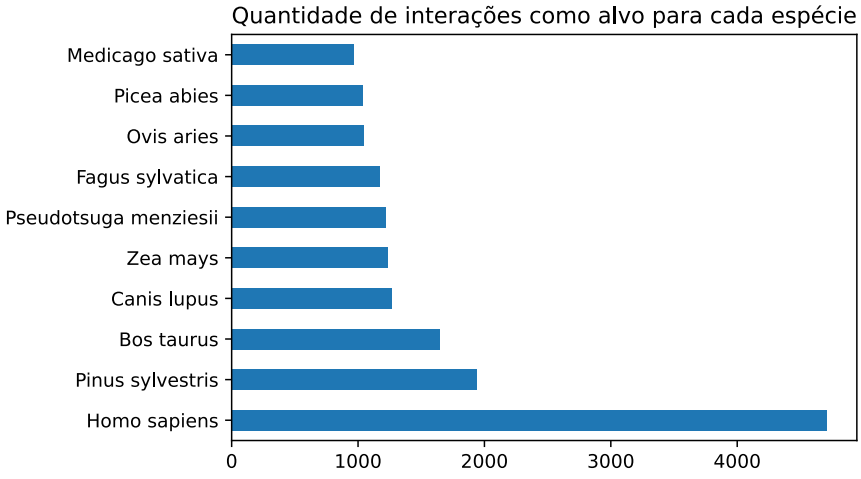

Figura 10: Gráfico das ocorrências de uma especie como alvo de interação

O gráfico revela que, surpeendentemente, o ser humano é o maior alvo de interações.

Para entender melhor como o ser humano alcançou esse posto de espécie que é o maior alvo de interações, foi feita a análise dos tipos das interações sofridas pelo ser humano.

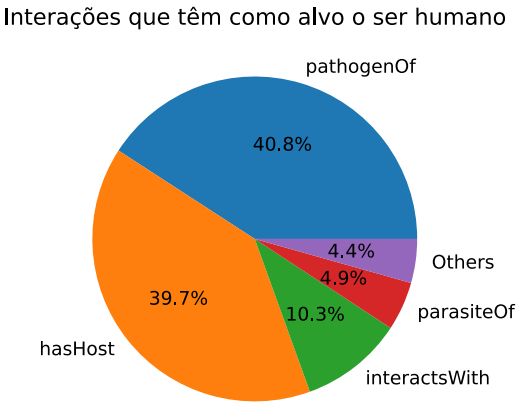

Figura 11: Gráfico da distribuição das interações sofridas pela espécie Homo sapiens

Como mostra o gráfico, a parte majoritária dessas interações parecem ser com vírus e

outros patógenos.

5.5 Espécies que mais se alimentam de outras

Uma das relações mais impactadas pela extinção de uma espécie é a de alimentação. Dentro desse cenário, com base nos dados de interação, foram analisadas as cinco espécies que mais se alimentam de outros.

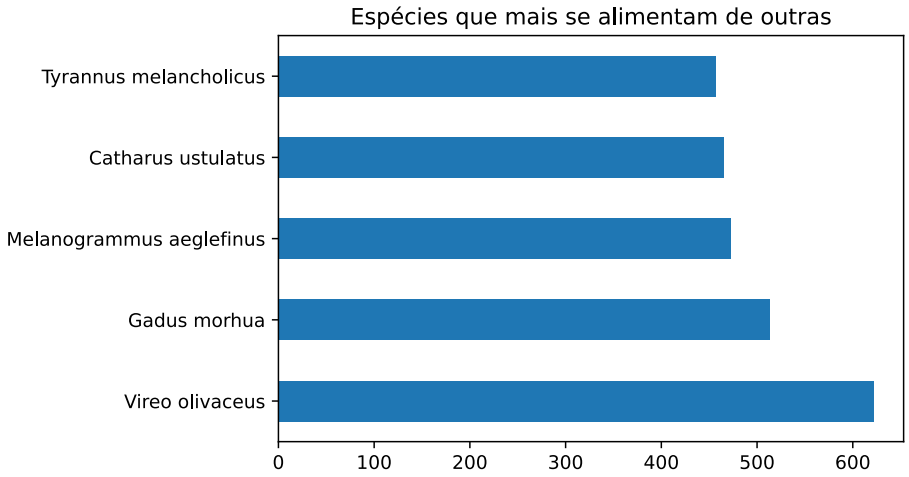

Figura 12: Gráfico das espécies que mais se alimentam de outras

A primeira posição é ocupada pelo Vireo olivaceus, um pássaro migratório, seguido de perto pelo Gadus morhua, um peixe.

5.6 Maiores predadores

Se alimentar de um maior número de espécies, não implica que a espécie seja uma grande predadora. É possível perceber isto ao analisar o gráfico obtido das cinco espécies que são maiores predadores.

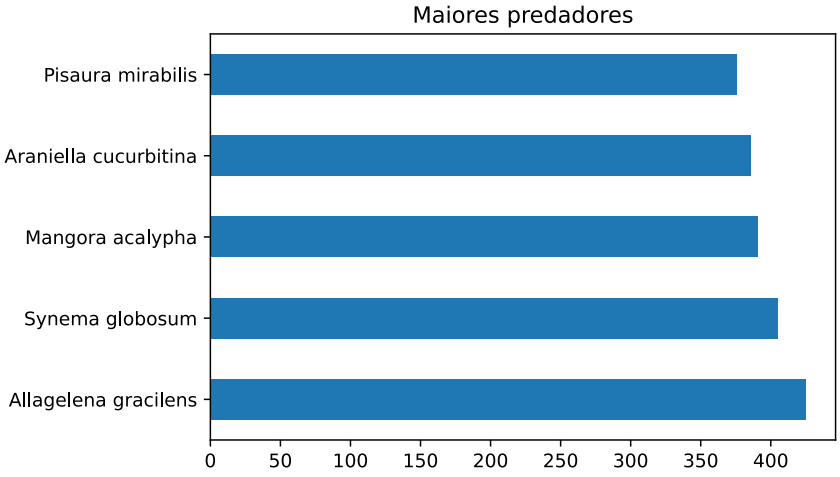

Figura 13: Gráfico das espécies que são maiores predadores

No gráfico, o maior predador encontrado na base foi o Allagelena gracilens, uma espécie de aranha encontrada na Europa, seguido de perto pela Synema Globosum, outro tipo de aranha.

5.7 Espécies que mais servem de alimento para outras

O mesmo foi feito para as espécies que mais servem de alimento. Foram analisadas as cinco espécies que mais servem de alimento.

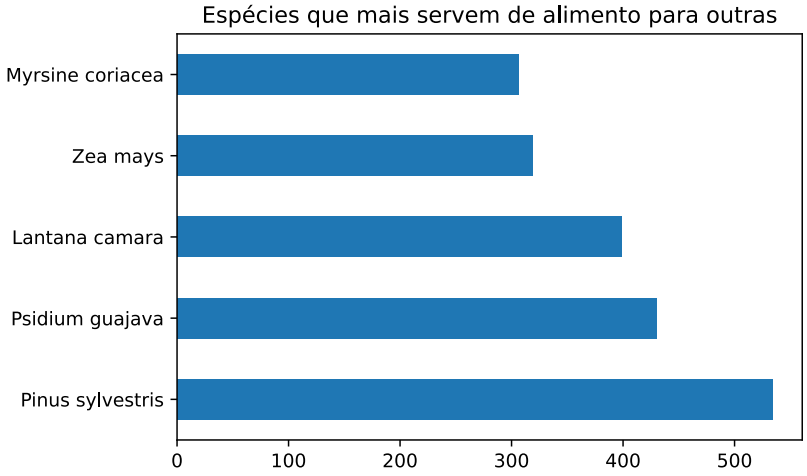

Figura 14: Gráfico das espécies que mais servem de alimento para outras

O primeiro lugar vai para o Pinus sylvestris, uma espécie de pinho silvestre, seguido pela Psidium guajava, uma outra planta. Esse resultado faz sentido, uma vez que as plantas são um tipo de alimento comum para animais.

5.8 Maiores presas

Já com relação às presas, o mesmo foi feito.

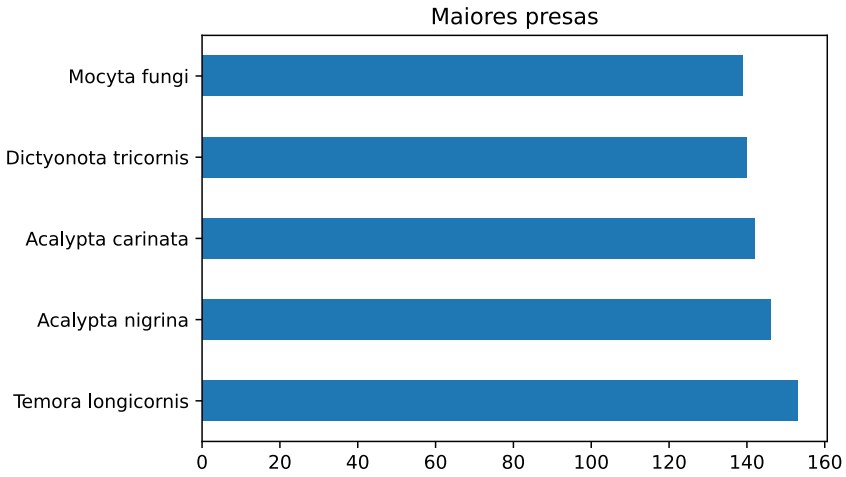

Figura 15: Gráfico das espécies que são maiores presas

O primeiro lugar fica para a Temora longicornis, um tipo de crustáceo, seguido de perto por Acalypta nigrina e Acalypta carinata, dois insetos. É possível notar que as primeiras posições são ocupadas por pequenos animais, que provavelmente ocupam a base da cadeia alimentar e, portanto, acabam sendo presas fáceis para outras espécies.

5.9 Interações entre as espécies em perigo crítico

Combinando os dados de ambas as bases, é possível observar a relação de espécies em perigo com as demais espécies. Nesse caso, foram observadas as interações que tinham como alvo espécies listadas como criticamente em perigo. A ideia é ter uma noção de como essas espécies, que já estão em alto perigo de extinção, podem impactar outras espécies.

Com esses dados das interações e da categoria das espécies alvo na lista da IUCN, foi criado um grafo mostrando a relação das espécies criticamente em perigo com as demais espécies.

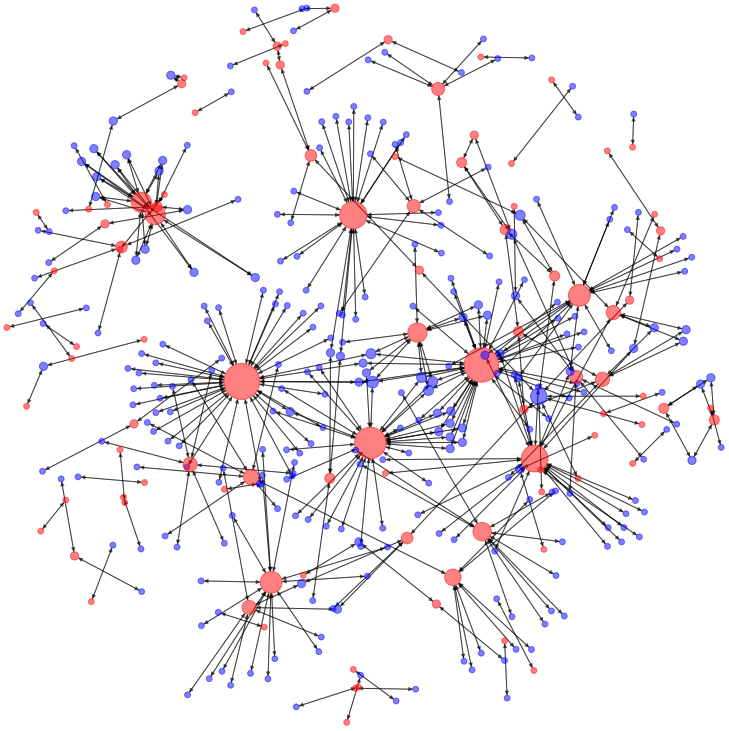

Figura 16: Grafo das interações entre as espécies em perigo crítico

É possível perceber que o grafo obtido não é completamente conectado, ou seja, há alguns casos de interações isoladas, onde o impacto parece ser menor. No entanto, vale destacar para estes casos que foram utilizadas somente as interações com espécies criticamente em perigo, então pode ser que haja mais interações para aqueles casos isolados que acabem não sendo exibidas nesse grafo. Por outro lado, é possível perceber também que há espécies que estão ligadas há mais de uma espécie em perigo. Assim, caso uma espécie em perigo seja extinta, ela pode até mesmo acelerar a extinção de uma outra espécie que já está em grande perigo.

5.10 Espécies com o maior grau de centralidade no grafo

Para ter uma ideia mais clara da influência de cada uma das espécies, foi calculado o grau de cada um dos nós, o que é baseado na quantidade de conexões que ele apresenta.

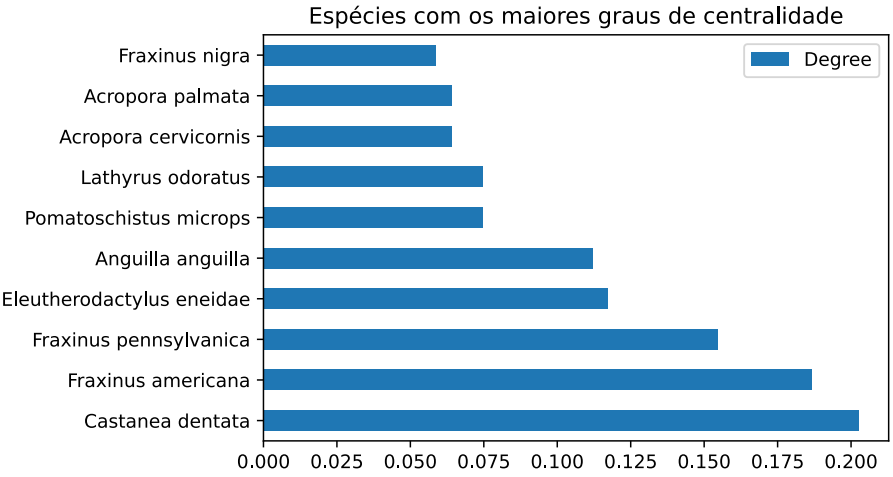

Figura 17: Gráfico das espécies com o maior grau de centralidade no grafo

Assim, os nós encontrados com mais conexões foram a Castanea dentata, uma planta de origem americana que produz castanhas, a Fraxinus americana, uma árvore também encontrada na América do Norte, e a Fraxinus Pennsylvanica. Coincidentemente ou não, as três primeiras espécies em perigo com maior influência nesse grafo gerado são plantas geralmente encontradas na América do Norte, o que pode indicar uma tendência a ser investigada.

6 Perguntas de pesquisa e explorações iniciais

O objetivo da equipe com o trabalho é tentar entender a relação entre espécies com risco de extinção e suas interações inter-espécie. Tendo em mente isso, foram levantadas algumas perguntas como: uma espécie estar em risco de extinção aumenta o risco das espécies que interagem diretamente com ela? É possível medir ou estimar o impacto que a extinção de uma espécie causará nos ambientes em que ela habita? Existe algum padrão ou tendência que levam grupos de espécies a serem mais vulneráveis? É possível medir os impactos que espécies invasoras tem nos ecossistemas em que elas são introduzidas?

Essas são questões bastante complexas e algumas delas um tanto subjetivas, no entanto os dados da IUCN e de interação entre espécies parecem ser razoáveis para responder a elas, pelo menos parcialmente.

7 Discussão e próximos passos

A exploração revelou algumas características e tendências importantes das bases de dados usada. O ponto mais importante é que as duas são extremamente ricas, com muito mais dados do que serão utilizados, não tendo falta de informações nos pontos que são importantes

para nossas análises. A base de dados das interações tem dois vieses importantes: o primeiro é a predominância de espécies da América do Norte e Europa; o outro é o favorecimento a espécies de interesse a nossa sociedade. Um bom exemplo dessas duas tendências é a espécie Gadus morhua, o bacalhau do atlântico, que aparece como uma das maiores fontes de interação. Similarmente, o Homo sapiens é o maior alvo, principalmente devido a patógenos. Uma boa explicação para esses vieses é a fonte desses dados, pesquisas que estudam interações entre espécies, que por motivos históricos e socio-econômicos ocorrem mais na América do Norte e Europa e, naturalmente, dão ênfase a assuntos que são interessantes para a sociedade. Já a Red List não parece apresentar essa mesma tendência, uma vez que seu objetivo é indexar todas as espécies existentes de forma global.

Tendo essas tendências em mente, as duas bases de dados juntas aparentam serem adequadas para responder nossas perguntas. Analisando a rede de interação de uma espécie, é possível comparar tendências populacionais e estimar a importância dela no seu ecossistema. A nossa exploração inicial mostrou que existem espécies em situação crítica com grau muito grande no grafo de interações, apontando para um caminho bem promissor. Com as informações geográficas é possível identificar tendencias regionais das espécies e analisar se a espécie é nativa ou exótica. Analisando o tipo de interação é possível descobrir se duas espécies competem por um mesmo recurso ou se participam de uma relação mutualística. Além disso, a Red List ainda possui descrições textuais extensas sobre cada espécie, seu habitat, as ameaças e várias outras informações que poderão ser usadas caso seja necessário mais embasamento para alguma conclusão.

Certamente, os dados poderiam ser mais completos. Nem todas as espécies ameaçadas possuem interações bem definidas na outra base de dados e a falta de séries históricas dificulta análises temporais mais complexas, mas ainda sim os dados disponíveis são suficientes para várias análises possíveis. Portanto, a equipe vai continuar com essas duas bases de dados, buscando explorar as questões propostas anteriormente. Não temos o objetivo de alcançar respostas definitivas no assunto, mas consideramos possível descobrir informações bastante interessantes.