Faculdade de Economia da Universidade do Porto

Tese de Mestrado de Análise de Dados e Sistemas de Apoio à Decisão

Estimação de Massa em Energia Eólica

Discente:

Maria José Gomes Pedroto

Orientador:

Prof. Dr. João Manuel Portela Gama

Setembro de 2013

Palavras-chave: Estimação de Massa, Algoritmos não Supervisionados de Identificação de Outliers, Half-Space Trees

Breve Nota Bibliográfica

Maria José Gomes Pedroto nasceu no Porto em Março de 1982.

Após ter frequentado o curso Tecnológico de Informática na Escola Secundária Fontes Pereira de Melo – Ensino Público ingressou em Setembro de 2000 na Licenciatura em Engenharia Informática e de Computação na Faculdade de Engenharia da Universidade do Porto. Para concluir a sua Licenciatura realizou um estágio na área de Serviços Móveis na empresa que à data geria o portal Telemoveis.com, tendo posteriormente trabalhado no desenvolvimento de aplicações informáticas em diferentes áreas (serviços móveis, seguradoras, transportes, serviços).

No ano de 2010 entrou na Faculdade de Economia da Universidade do Porto no Mestrado de Análise de Dados e Sistemas de Apoio à Decisão, encontrando-se a terminar a sua tese de mestrado na área de identificação de anomalias com recurso a métodos de estimação de massa.

Agradecimentos

A elaboração desta dissertação não teria sido possível sem a colaboração, apoio e amizade de algumas pessoas a quem gostaria de agradecer:

Ao Prof. Dr. João Portela Gama, pela motivação e contínuo incentivo na aprendizagem da temática de estimação de massa, transmitindo-me todos os aspetos mais importantes para a sua implementação. Por todo o apoio logístico prestado, quer através da disponibilização dos papers e documentos que constituem a base da presente dissertação, para além dos contatos estabelecidos com a empresa que forneceu a informação, a *Prewind Lda.*, na qualidade do Engenheiro João Sousa. Aproveito ainda para agradecer a disponibilidade, crítica e juízos de valor, para além de ser uma porta sempre aberta independentemente da altura escolhida, tanto na tese como nas cadeiras que leciona no mestrado.

A todo o corpo docente que constitui o Mestrado de Análise de Dados e Sistemas de Apoio à Decisão, por toda a excelência e qualidade que imprimem nos alunos.

À Angela, Vera, Eduarda, Rui, Sérgio e demais amigos que fiz durante a frequência do mestrado, pelo apoio, boa disposição e risos partilhados durante as aulas e fora delas.

À Cláudia, por me dar o estímulo e força inicial que necessitava para concluir a minha candidatura para o retorno à Faculdade.

À minha família, pelo apoio, motivação e força para me formar na pessoa que sou hoje. Gostaria especialmente de agradecer à minha Mãe. Apesar de fisicamente já não estares presente na minha vida, farás sempre parte dela.

Por fim gostaria de agradecer ao Paulo. Acho que sem ti nunca teria conseguido virar esta página.

Resumo

O objetivo deste trabalho passa pela aplicação de algoritmos de estimação de massa para a problemática de deteção de anomalias em *datasets* provenientes de sistemas SCADA – equipamentos eletrónicos acopolados em aerogeradores capazes de fornecer até 800 variáveis distintas.

Como resultado prático foi desenvolvida uma aplicação informática capaz de criar e manter estruturas dinâmicas de dados denominadas de *Half-Space Trees*, utilizadas para a implementação de algoritmos de Estimação de Massa, tanto na componente *offline* como *online*.

Os resultados de execução dos algoritmos implementados foram posteriormente comparados com os resultados de um algoritmo baseado em Densidade - *Local Outlier Factor*. Este algoritmo é considerado por muitos como um estado da arte das diferentes técnicas baseadas em densidade.

A implementação dos métodos considerados pertinentes permitiu obter um conhecimento aprofundado na temática de Estimação de Massa, um conceito aplicável em diferentes áreas, com especial relevância para Regressão, Estimação de Anomalias e Pesquisa de Informação.

De salientar que a metodologia de comparação dos algoritmos implementados pressupõem que os dados não estejam previamente classificados, sendo necessária a aplicação exclusiva de algoritmos não supervisionados.

Abstract

This work is about the application of mass estimation algorithms to the problem of detecting anomalies in datasets generated from SCADA systems - electronic equipment placed in wind turbines capable of supplying up to 800 separate variables.

As a practical result we developed a software application that is able to create and maintain dynamic data structures called Half-Space Trees, the basis to a successful implementation of Mass Estimation Algorithms, both in *offline* and *online* components.

The results of the implemented algorithms were then compared with the results of a density based algorithm - *Local Outlier Factor*. This algorithm is considered by many as a state of the art of the different density based techniques.

The implementation of the methods considered relevant allowed obtaining a deeper knowledge to Mass Estimation, a concept appliable in different areas, including Regression, Anomaly Detection and Information Retrieval.

Please note that the methodology used to compare the implemented algorithms assumes that the data hasn't been previously classified, which requires the application of only unsupervised algorithms.

Índice

Breve Nota Bibliográfica iii

Agradecimentos iv

Resumo v

Abstract vi

Índice vii

Lista de figuras x

Lista de tabelas xiv

Abreviaturas xv

Capítulo 1 1

Introdução 1
1.1. Contexto Geral 1
1.2. Energia Eólica em Números 2
1.3. Vantagens e Desvantagens do Uso da Energia Eólica 4
1.4. Objectivos da Dissertação 5
1.5. Estrutura da Dissertação 5

Capítulo 2 7

Energia Eólica 7
2.1. Sistema Mecânico associado à Energia Eólica 7
2.1. Análise da Performance Operacional de Parques Eólicos 8
2.2. Metodologias de Monitorização e Optimização de Produção Energética de Parques Eólicos 10
2.3. Análise de Curvas para Monitorização de Operação 11

Capítulo 3 15

Estado da Arte - Deteção de Outliers 15
3.1. Definição de Outliers 15
3.2. Tipos de Outliers 18
3.2.1. Outliers Globais ou Pontuais 19

3.2.2. Outliers Condicionais 19
3.2.3. Outliers Coletivos 20
3.3. Retorno de Algoritmos de Identificação de Outliers 20
3.4. Desafios na Deteção e Classificação de Outliers 21
3.5. Técnicas ou Metodologias de Deteção de Outliers 23
3.5.1. Técnicas Estatísticas 24
3.5.2. Técnicas ou Metodologias Baseadas em Profundidade (Depth-based Approaches) 24
3.5.3. Técnicas ou Metodologias baseadas em Distância (Distance-based Approaches) 26
3.5.4. Técnicas ou Metodologias baseadas em Densidade (Density-based Approaches) 27
3.5.5. Técnicas ou Metodologias baseadas em Agrupamento (Clustering-Based) 28
3.6. Ferramentas de Mineração de Dados para Deteção de Outliers 29
3.7. Estimação de Massa 29
3.7.1. Definição de Massa 30
3.7.2. Análise de Conjuntos de Dados (*Emsemble Analysis*) 31
3.7.3. Half-Space Trees e Estimação de Massa 31
3.7.4. Half-Space Trees e Deteção de Anomalias em Fluxos Contínuos de Dados 38

Capítulo 4 42

Análise Exploratória dos Dados 42
4.1. Caracterização das Variáveis em Estudo 42
4.2. Análise da Potência Gerada 43
4.3. Análise da Velocidade do Vento 44
4.4. Análise da Posição da Nacelle 46
4.5. Análise das Rotações por Minuto 47
4.6. Outliers Identificados com base nos Resíduos das Séries Temporais 47

Capítulo 5 49

Estimação de Massa 49
5.1. Especificação de Fases Comuns de Execução 49
5.1.1. Fase de Treino 50
5.1.2. Fase de Estimação de Outliers 52
5.2. Resultados Alcançados 53
5.2.1. Árvores de Cortes Aleatórios 53
5.2.2. Árvores de Cortes Alternados 57
5.2.3. Local Outlier Factor 59
5.2.4. Deteção de Anomalias em *DataStreams* 62
5.2.5. Comparação de Algoritmos 65

Capítulo 6 67

Conclusões 67

Referências 70

Anexos 73

Anexo 1 – Código R para Análise Exploratória Parcial 74

Anexo 2 – Estatísticas Descritivas das Variáveis Analisadas 77

Anexo 3 – Scoring segundo Análise da Série Temporal para as variáveis Velocidade do Vento e Potência Gerada 79

Anexo 4 – Análise do Scoring de Half-Space Trees de Cortes Aleatórios 82

Anexo 5 – Análise do Scoring de Half-Space Trees de Cortes Sequenciais 83

Anexo 6 – Análise do Scoring do LOF (Local Outlier Factor) 85

Anexo 7 – Análise do Scoring do Processamento de Fluxos Contínuos de Dados com um Modelo de Informação Vazio 88
Anexo 8 – Identificação de Anomalias nas Restantes Curvas de Energia 92

Lista de figuras

Figura 1 – 10 maiores Produtores Mundiais de Energia Éólica (Master Distancia Lda., 2012) 3

Figura 2 – Repartição da Energia Comercializada pela EDP Serviço Universal, por Tecnologia, durante o ano de 2012 3

Figura 3 – Identificação dos diferentes componentes de uma Turbina Horizontal 8

Figura 4 – Estado operacional de um parque eólico constituído por 40 aerogeradores (Harman e Raftery, 2003) 9

Figura 5 – Curva de Potência genérica de um aerogerador (Wind Power Program UK, 2009) 12

Figura 6 – Potência Gerada Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010 13

Figura 7 – Velocidade do Rotor Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010 13

Figura 8 – Ângulo do Blade Pitchº Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010 14

Figura 9 – Um exemplo simples de outliers num espaço bidimensional composto por duas variáveis: X e Y (Chandola *et al.*) 16

Figura 10 - Identificação dos diferentes níveis de valor na informação 17

Figura 11 – Exemplo de Outliers Globais (também denominados de Pontuais) num espaço bidimensional (Chandola *et al.*) 19

Figura 12 - Exemplo de um Outlier Coletivo – eletrocardiograma (Chandola *et al.*) 20

Figura 13 - Execução do IsoDepth num conjunto de dados bivariados no R 25

Figura 14 –Identificação de Outliers em dados com diferentes níveis de Densidade 27

Figura 15 – Pseudo-Código do LOF 28

Figura 16 - Half-Space Tree sem limite de massa ativo 33

Figura 17 – Pseudocódigo para Inicializar Espaço de Trabalho 33

Figura 18 – Pseudocódigo para Criar uma Half-Space Tree 34

Figura 19 - Exemplo de Half-Space Tree com limite de dimensão 35

Figura 20 – Função de Pontuação de um registo com base numa Floresta de Half-Space Trees 36

Figura 21 - Half-Space Tree com cortes alternados 36

Figura 22 - Half-Space Tree com cortes aleatórios 37

Figura 23 - Particionamento do Espaço de Dados em Células com Cortes Alternados 37

Figura 24 –Particionamento do Espaço de Dados em Células com Cortes Aleatórios 37

Figura 25 - Processamento de Fluxos Contínuos de Dados 38

Figura 26 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados ... 39

Figura 27 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados ... 40

Figura 28 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados ... 40

Figura 29 – Processamento do fluxo de dados 41

Figura 30 - Análise da Potência Gerada pelo Aerogerador 1 (janeiro, fevereiro, março 2010 - Parque X) 43

Figura 31 – Box-Plot da variável da Potência Gerada 44

Figura 32 - Série Temporal da Velocidade do Vento do Aerogerador 1 (janeiro, fevereiro e março de 2010 – Parque X) 44

Figura 33 - Diagrama de Box-Plot da variável de Velocidade do Vento 45

Figura 34 – Série Temporal da Posição da Nacelle do Aerogerador 1 (janeiro, fevereiro e março de 2010 – Parque X) 46

Figura 35 –Box-Plot da variável Posição da Nacelle 46

Figura 36 - Série Temporal da Rotação por minuto do Aerogerador 1 (janeiro, fevereiro e março de 2010 - Parque X) 47

Figura 37 – Box-Plot da variável Rotações por Minuto 47

Figura 38 - Identificação de Anomalias na Série Temporal da Potência Gerada 48

Figura 39 - Identificação de Outliers na Série Temporais da Velocidade do Vento 48

Figura 40 - Estimação de Outliers com recurso a Half-Space Trees e Criação de diferentes Conjuntos de Dados 50

Figura 41 - Identificação dos principais passos de Estimação de Massa com recurso a Half-Space Trees 51

Figura 42 - Séries Temporais associadas às variáveis escolhidas para análise 52

Figura 43 - Identificação da Estrutura de uma Half-Space Tree gerada pelo sistema 52

Figura 44 – Expressões que permitem calcular a pontuação para cada registo que faça parte do conjunto de teste 53

Figura 45 - Dispersão dos Dados Normalizados com identificação manual de 3 regiões de dados mais dispersos em relação à Curva de Potência original 54

Figura 46 – Ranking com Árvores de Cortes Aleatórios com 20 registos 55

Figura 47 - Ranking com Árvores de Cortes Aleatórios com 65 registos 55

Figura 48 - Ranking com Árvores de Cortes Aleatórios com 130 registos 55

Figura 49 - Ranking com Árvores de Cortes Aleatórios com 648 registos 55

Figura 50- Distribuição de Frequências do resultado de Scoring 56

Figura 51 - Distribuição de Frequências Acumulada dos grupos de Scoring obtidos com base nas Half-Space Trees de Cortes Aleatórios 56

Figura 52 –Distribuição de Frequências do resultado de Scoring 57

Figura 53 - Ranking para Árvores de Cortes Alternados (20 registos) 58

Figura 54 - Ranking para Árvores de Cortes Alternados (65 registos) 58

Figura 55 - Ranking para Árvores de Cortes Alternados (130 registos) 58

Figura 56 – Ranking Árvores de Cortes Alternados (648 registos) 58

Figura 57 –Distribuição de Frequências Acumulada dos grupos de Scoring obtidos com base nas Half-Space Trees de Cortes Sequenciais 59

Figura 58 - Diagrama de Distribuição de Frequências do LOF 60

Figura 59 – Distribuição de Frequências Acumulada dos grupos de Scorings criados com base no Local Outlier Factor 60

Figura 60 – Ranking com LOF (20 registos) 61

Figura 61 – Ranking com LOF (65 registos) 61

Figura 62 - Ranking com LOF (130 registos) 61

Figura 63 - Ranking com LOF (648 registos) 61

Figura 64 – Posição dos rankings no Modelo com DataStreams no 1ºgrupo 62

Figura 65 - Ranking com DataStreams e Modelo de Informação vazio (20 registos) 63

Figura 66 - Ranking com DataStreams e Modelo de Informação vazio (65 registos) 63

Figura 67 - Ranking com DataStreams e Modelo de Informação vazio (130 registos) 63

Figura 68 - Ranking com DataStreams e Modelo de Informação vazio (648 registos) 63

Figura 69 – Diagrama de Distribuição de Frequências com base no Processamento de Fluxos Contínuos de Dados 64

Figura 70 – Diagrama de Distribuição de Frequências Acumulada 64

Figura 71 - Distribuição do Scoring para a Velocidade do Vento 81

Figura 72 - Distribuição do Scoring para a Potência Gerada 81

Figura 73 – Distribuição de Scoring com Half-Space Trees de Cortes Aleatórios 82

Figura 74 – Distribuição de Scoring com Half-Space Trees de Cortes Sequenciais 84

Figura 75 - Interface do Local Outlier Factor 85

Figura 76 – Interface do Processamento de Fluxos Contínuos de Dados com base em Estimação de Massa 88

Figura 77 - Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa com Cortes Aleatórios – Curva do Rotor 92

Figura 78 – Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa de Cortes Aleatórios – Curva do Blade Pitch ... 92

Figura 79 – Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa de Cortes Aleatórios – Curva de Potência 93

Lista de tabelas

Tabela 1 - Métodos de Previsão de Vento e de Potência Gerada (Soman *et al.*, 2010) 10

Tabela 2 - Comparação de Métodos de Análise de Dados Univariados, Bivariados e Multivariados, adaptado de (Llango *et al.*, 2012) 21

Tabela 3 - Distinção de Técnicas de Deteção de Outliers baseadas em Machine Learning (Llango *et al.*, 2012)..... 22

Tabela 4 - Variáveis em Análise nos Datasets de Dados do Parque X..... 42

Tabela 5 – Tempo de Execução e Número de Anomalias dos Algoritmos Implementados 65

Tabela 6 – Estatísticas Descritivas da Velocidade do Vento 77

Tabela 7 – Estatísticas Descritivas da Potência Gerada..... 77

Tabela 8 – Estatísticas Descritivas das Rotações por Minuto 77

Tabela 9 – Estatísticas Descritivas da Posição da Nacelle..... 78

Tabela 10 – Scoring de Outliers Univariados da Série Temporal com base na Variável de Velocidade do Vento 79

Tabela 11 – Scoring de Outliers Univariados da Série Temporal com base na variável da Potência Gerada 80

Tabela 12 - Tabela de Distribuição de Frequências do Scoring obtido para o Algoritmo Aleatório de Half-Space Trees 82

Tabela 13 - Tabela de Distribuição de Frequências do Scoring obtido para o Algoritmo Sequencial de Half-Space Trees 83

Tabela 14 - Scorings obtidos com o LOF para os primeiros 64 grupos..... 87

Tabela 15 - Tabela de Distribuição de Frequências de Pontuações Atribuídas com base no processamento de Fluxos Contínuos de Dados (Estimação de Massa)..... 91

Abreviaturas

Lista de abreviaturas (ordenadas por ordem alfabética)

ANN	*Artificial Neural Networks* (Redes Neuronais Artificiais)
ARMA	*Modelo de Médias Móveis Auto-Regressivo*
BRIC	*Grupo de Países formado pelo Brasil, Rússia, Índia e China*
FEP	*Faculdade de Economia da Universidade do Porto*
GPS	*Global Positioning System*
LOF	*Local Outlier Factor*
MADSAD	*Mestrado em Análise de Dados e Sistemas de Apoio à Decisão*
NWP	*Numeric Weather Prediction*
PME	*Pequenas e Médias Empresas*
QREN	*Quadro de Referência Nacional*
SCADA	*Supervisory Control and Data Acquisition*
WWEA	*World Wind Energy Association*

Capítulo 1

Introdução

Na natureza, nada se cria, nada se perde, tudo se transforma

[Antoine Laurent de Lavoisier]

A Energia Eólica é uma fonte de energia renovável com um imenso potencial de crescimento (ENEOP, 2013). Por fonte de energia renovável deve-se entender um processo de geração contínua de energia, originária de fontes naturais e inesgotáveis.

Atualmente, as fontes de Energia Renovável são baseadas em: *BioCombustíveis*, *Biomassa*, *Energia Azul*, *Energia Geotérmica, Energia Hidraúlica, Hidreletricidade, Energia Solar, Energia Maremotriz, Energia das Ondas, Energia das Correntes Marítimas,* além da já referida *Energia Eólica*. Por outro lado, fontes de Energia Não Renovável são aquelas em que não é possível num curto espaço de tempo repor os recursos que se gastam, sendo necessários vários milhões de anos para que estes voltem a se formar. As principais fontes são baseadas em: *energia de fissão nuclear* e *combustíveis fósseis (petróleo, gás natural e carvão).*

1.1. Contexto Geral

Desde a *Idade Média* que um pouco por toda a Europa a energia do vento é utilizada para movimentar a engrenagem de moinhos de vento - edifícios responsáveis pela conversão de grãos de cereais em farinha (trigo, milho, centeio), para se obter, entre outros alimentos, o pão. Apesar dos avanços tecnológicos existentes, que já não fazem

depender de moinhos a obtenção de um dos principais ingredientes para a alimentação mundial, ainda hoje a energia gerada pelo vento é considerada uma mais-valia em diferentes áreas. Uma dessas áreas passa pela indústria marítima, onde, desde a *Antiguidade,* são conhecidos exemplos de como o vento movimenta embarcações através da força contínua que exerce nas velas.

Atualmente, tanto em terra como no mar, a força do vento é utilizada para mover aerogeradores – grandes sistemas eléctricos integrados no eixo de um catavento que convertem energia eólica em energia eléctrica.

1.2. Energia Eólica em Números

No campo de exploração de recursos energéticos, a Energia Eólica é vista como uma das fontes energéticas mais promissoras. É renovável, limpa, inesgotável, global e verde. Por outro lado, e contrariamente aos combustíveis convencionais, a energia do vento é uma fonte massiva de energia indígena, permanentemente disponível em qualquer parte do mundo (Brandão *et al.*, 2009). Assim, e com as actuais pressões mundiais a focarem o recurso a formas de energia verdes, aliado ao exponencial crescimento das necessidades energéticas da população mundial, esta é uma área em necessário crescimento.

Os valores de produção energética alcançados pelos sistemas baseados em Energia Eólica têm vindo a crescer desde 2007.

De acordo com a WWEA, em 2005 a Energia Eólica instalada no Continente Europeu alcançava níveis de potência da ordem dos 40,504 MW e era capaz de gerar anualmente 83 TWh de eletricidade (Brandão *et al.*, 2007). Neste campo, Portugal encontrava-se no 10º lugar dos países com maiores taxas de produção energética (Figura 1) baseada em Energia Eólica (Master Distancia Lda., 2012). A título de exemplo, dentre as diferentes fontes energéticas à disposição da EDP Serviço Universal, empresa responsável por uma fatia do mercado de fornecimento de energia em Portugal, tanto ao nível empresarial como doméstico, a Energia Eólica possui um papel preponderante com uma média de 40% do total de fornecimento energético, durante o ano de 2012 (Figura 2)..

POSIÇÃO	PAÍS	TOTAL DE CAPACIDADE INSTALADA [MW] PRIMEIRO SEMESTRE 2012
1	China	67.774
2	Estados Unidos	49.802
3	Alemanha	30.016
4	Espanha	22.087
5	Índia	17.351
6	Itália	7.280
7	França	7.182
8	Reino Unido	6.840
9	Canadá	5.511
10	Portugal	4.398

Figura 1 – 10 maiores Produtores Mundiais de Energia Éólica (Master Distancia Lda., 2012)

Atualmente, em Portugal, as Energias Renováveis contribuem com cerca de 3,2 milhões de euros para o PIB nacional e criam 50 mil postos de trabalho qualificado. Dessa forma promove-se indiretamente a competitividade, crescimento e independência económica nacional.

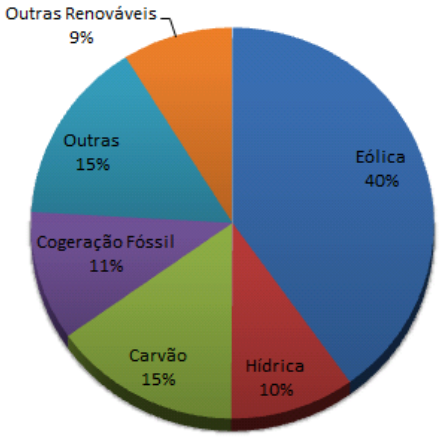

Figura 2 – Repartição da Energia Comercializada pela EDP Serviço Universal, por Tecnologia, durante o ano de 2012

Quando se analisa o retorno energético de um parque eólico é importante considerar diversos factores que influenciam a quantidade de energia gerada, de onde se salienta:

fatores diários (associados à hora do dia), sazonais, e ainda fatores associados à topografia e rugosidade do solo. Algumas das dificuldades associadas à gestão e manutenção dos parques eólicos passa pelas dificuldades de previsão da energia gerada (Kusiak *et al.*, 2009), cujos erros dificultam a definição de um processo de comercialização eficaz tanto a médio como a longo prazo.

1.3. Vantagens e Desvantagens do Uso da Energia Eólica

A aposta na criação e implementação de parques eólicos apresenta algumas vantagens face a outras fontes energéticas. Entre outras destacam-se:

- sistema gratuito em termos energéticos, não necessitando de combustíveis para apoiar o seu funcionamento;
- estrutura que não produz desperdícios nem aumenta os efeitos de estufa, que tanto têm danificado o meio ambiente;
- a possível reutilização das terras limitrofes para a prática da agricultura;
- possíveis atrações turísticas podendo aumentar as taxas de turismo numa dada região.

Por outro lado a criação de pequenos parques eólicos permite fornecer energia a populações em regiões remotas, diminuindo possíveis custos com o transporte da rede elétrica. Caso as necessidades de fornecimento sejam muito básicas, também poderá ser importante avaliar o recurso a painéis solares, especialmente em zonas do globo mais propícias a este investimento.

A aposta na Energia Eólica contribui para a autonomia energética dos países, com peso no aumento da independência política face a potências energéticas estrangeiras, diminuindo ainda riscos de aumento exacerbado do preço de combustíveis (Brandão *et al.*, 2009).

Assim como há vantagens, existem também algumas desvantagens face ao recurso de Energia Eólica, de onde se salienta:

- os níveis de ruído nas redondezas podem atingir valores bastante elevados, pelo que a região circundante não é favorável à localização de complexos habitacionais;

- pode haver interferência no movimento migratório de determinadas espécies de aves, sendo necessário um estudo ambiental antes da construção de um parque numa determinada localização;
- pelas alterações visuais que acarretam, os parques eólicos não são facilmente aceites pelas populações residentes nas áreas vizinhas;
- as zonas mais favoráveis à localização destes sistemas acarretam prejuízos imobiliários para os donos de terras próximas;
- existem incertezas de produção, pois o retorno energético pode variar consoante um conjunto variável de factores externos relacionados principalmente com o ecossistema de um parque eólico.

1.4. Objectivos da Dissertação

A temática de *Deteção de Outliers* é uma das áreas em constante crescimento na disciplina de *Datamining*. Esta técnica é utilizada para detetar e remover objetos anómalos em grandes volumes de dados. Estas anomalias surgem com alguma frequência através de falhas mecânicas, alterações no comportamento de sistemas complexos, existência de esquemas fraudulentos, intrusões em redes informáticas, erros humanos, entre outras situações.

Esta dissertação pretende relacionar a temática da Energia Eólica com a análise e investigação de anomalias em conjuntos de dados, mais exatamente com os esforços alcançados na área de Estimação de Massa. Considera-se que estas duas áreas partilham diferentes características: são complexas, abrangentes e têm temas interligados com outras disciplinas.

Com este trabalho pretende-se de uma forma prática, clara e objetiva fundamentar o recurso a técnicas de mineração de dados, nomeadamente na criação de modelos de informação que permitam a análise de anomalias.

1.5. Estrutura da Dissertação

Esta Dissertação encontra-se organizada em seis capítulos distintos.

No capítulo 1 é feita uma introdução ao tema da dissertação, sendo ainda feito um enquadramento da temática de Energia Eólica com a apresentação de um ponto de

situação ao nível nacional. É ainda elaborada uma clarificação dos objetivos de elaboração da presente tese.

No capítulo 2 é realizada uma breve introdução à temática de Energia Eólica e a todos os conceitos técnicos considerados mais importantes.

No capítulo 3 são apresentados os fundamentos teóricos associados ao estado da arte de *Deteção e Análise de Anomalias*. É ainda apresentada a temática relacionada com estimação de massa – a base do trabalho desenvolvido.

No capítulo 4 é realizada uma breve análise exploratória de parte dos dados disponibilizados.

No capítulo 5 é efetuada uma especificação dos métodos implementados no âmbito da presente dissertação. Os resultados das diferentes metodologias utilizadas são comparados tendo em conta as anomalias identificadas.

No capítulo 6 são definidas algumas áreas de desenvolvimento futuro, com vista à melhoria da aplicação desenvolvida. São ainda referidas as principais conclusões obtidas com a presente Dissertação.

Capítulo 2

Energia Eólica

A temática de Energia Eólica é complexa. Atualmente, os registos resultantes do funcionamento de cada aerogerador são recolhidos por sensores e enviados para um centro de controlo por sistemas SCADA - *Supervisory Control and Data Acquisition*. De seguida, os operadores do sistema têm de analisar estes dados para identificarem sintomas de quaisquer problemas com as turbinas. Tendo em atenção que cada aerogerador pode enviar até 800 sinais, compreende-se que esta seja uma tarefa complexa e que muitas vezes eventuais falhas sejam detetadas muito tarde (Brandão *et al.*, 2009).

De seguida será explicado o sistema mecânico associado aos aerogeradores – instrumentos responsáveis pela conversão da Energia Cinética do Vento em Energia Elétrica.

2.1. Sistema Mecânico associado à Energia Eólica

A Energia Eólica é obtida através da movimentação exercida nos aerogeradores pela força do vento.

Um aerogerador moderno de eixo horizontal (Figura 3) é constituído por uma torre, cuja altura pode variar entre os 50 e os 120 metros. No cimo dessa estrutura situa-se o rotor, mecanismo com 2 ou 3 pás e que tem uma dimensão que varia entre 25 e 45 metros; a *nacelle* que abriga o aerogerador; bem como os sistemas de controlo de toda a estrutura mecânica.

O correto funcionamento do aerogerador pressupõe que o vento seja de tal ordem que coloque em movimento as pás numa frequência de rotação na ordem das 10 a 25 voltas por minuto.

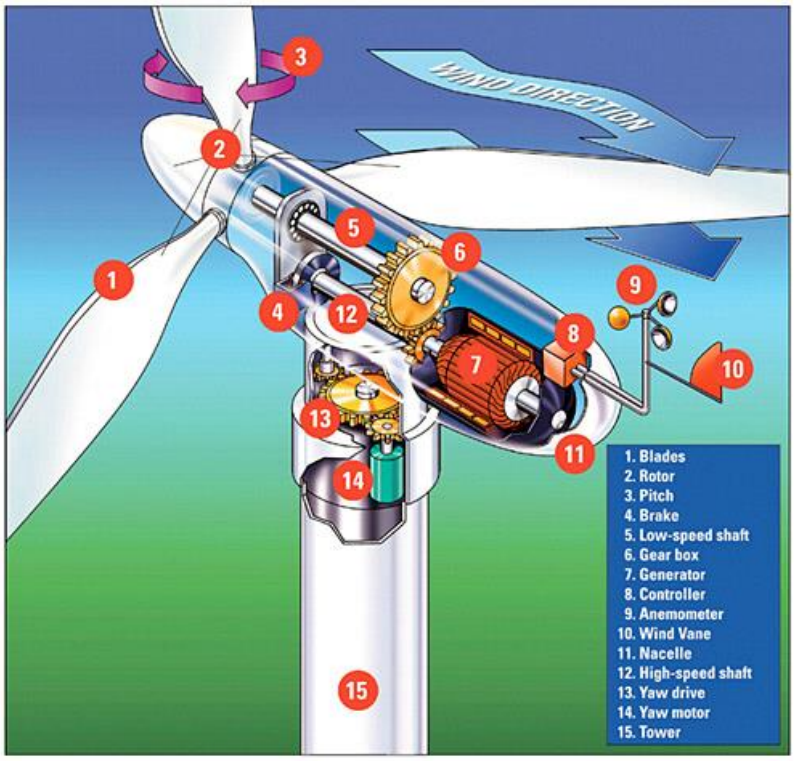

Figura 3 – Identificação dos diferentes componentes de uma Turbina Horizontal

O gerador que faz parte da *nacelle* transforma a energia mecânica do movimento das pás em energia elétrica. Um sistema rotativo acopulado à *nacelle* permite que esta se direcione sempre face ao vento, para assim maximizar a energia gerada.

A potência gerada pelas turbinas eólicas varia entre 2 e 3 MW. Estas normalmente situam-se em áreas ventosas, alternando entre a zona costeira e zonas montanhosas, onde o relevo fornece um efeito de aceleração à força do vento.

2.1. Análise da Performance Operacional de Parques Eólicos

Ao longo dos últimos anos foram definidos diferentes modelos para analisar o retorno produzido por parques eólicos. Estas técnicas tiveram de evoluir à medida que as características tecnológicas das próprias turbinas foram-se transformando. Dentre as diferentes técnicas de análise de dados SCADA salientam-se (Harman e Raftery, 2003; Lindahl e Harman, 2012):

- animação gráfica do estado operacional dos aerogeradores existentes num parque eólico em diferentes instantes temporais, com base na distribuição espacial dos aerogeradores e identificação do estado de cada máquina em instantes sucessivos (Figura 4);
- identificação e avaliação de padrões nos dados, identificando mudanças de índices de produção de cada aerogerador (normalmente atribuídos ao mau funcionamento do *pitch control*, danificação ou incrustação de gelo nas lâminas, redefinição dos ângulos das lâminas, melhoramentos aerodinâmicos ou restrições de funcionamento);
- avaliação do impacto de restrições de potência, com base no *pitch angle* das lâminas (em que se processa a correlação entre o ângulo das lâminas e a potência gerada pelo aerogerador);
- comparação de situações de restrições de potência vs. avaliação de situações sem quaisquer restrições;
- análise em pormenor da *Curva de Potência* de um parque eólico;
- previsão de velocidade de vento e da potência gerada;
- comparação de índices de produção energética tendo em conta as estações do ano;
- recurso à metodologia de *Change-Point Analysis* para identificação de alterações em séries temporais provenientes dos aerogeradores.

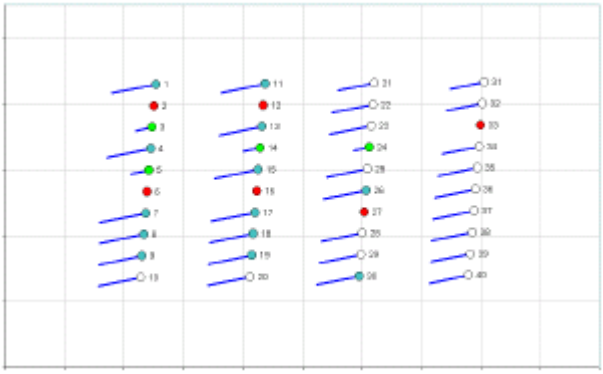

Figura 4 – Estado operacional de um parque eólico constituído por 40 aerogeradores (Harman e Raftery, 2003)

2.2. Metodologias de Monitorização e Optimização de Produção Energética de Parques Eólicos

Método de Previsão	Subclasse	Exemplos	Comentários
Métodos Persistentes (Naive Predictor)		*P(t+k) = P(t)*	- Método preciso para um horizonte temporal curto ou muito curto.
Métodos relacionados com modelos ambientais e físicos	*Numerical Weather Prediction (NWP)*	- *Global Forecasting System;* - MM5; - *Prediktor;* - HIRLAM.	- Uso de dados metereológicos (p.e. velocidade e direcção do vento, pressão do ar, temperatura); - Ideal para previsões de longo prazo.
Métodos Estatísticos	Redes Neuronais:	- Feed-Forward; - Recurrent; - Multilayer Perceptron; - Radial Basis Function; - ADALINE.	- Utilizados em previsões a curto prazo; - Fazem uso de estruturas híbridas, úteis para previsão de médio e longo prazo;
	Modelos de Séries Temporais:	- ARX; - ARMA e ARIMA; - *Grey Predictors;* - *Linear Predictions;* - *Exponential Smoothing.*	- Bons indíces de previsão a curto prazo; - Alguns modelos de séries temporais são melhores do que Redes Neuronais.
Metodologias Inovadoras		- Spatial Correlation; - Fuzzy Logic; - Wavelet Transform; - *Emsemble Predictions;* - Entropy based training.	- *Spatial Correlation* tem bons indíces para curto prazo; - Treino baseado em Entropia melhora a *performance* dos modelos.
Metodologias Híbridas		- NWP + NN; - ANN + Fuzzy logic = ANFIS; - Spatial Correlation + NN; - NWP + time series.	- ANFIS funciona bem para previsões de muito curto prazo; - Estruturas NWP + NN são muito precisas para previsões de médio e longo prazo.

Tabela 1 - Métodos de Previsão de Vento e de Potência Gerada (Soman *et al.*, 2010)

Em termos previsionais de potência gerada, os resultados obtidos dependem de fatores internos e de fatores externos. Algoritmos que devolvam bons resultados quando aplicados a uma gama de parques eólicos poderão não ter o mesmo resultado quando se considera outra gama.

Ao se avaliar a produção energética de um parque eólico é importante estudar diferentes métodos de previsão. Na Tabela 1 encontram-se sintetizadas as principais técnicas desenvolvidas até ao momento de onde se salienta o recurso a: Redes Neuronais, Modelação de Séries Temporais e Modelos Híbridos (Soman *et al.*, 2010). Existem referências a estas e outras metodologias em:

- Análise da Performance Operacional de Parques Eólicos (Harman e Raftery, 2003; Lindahl e Harman, 2012) ;
- Previsão de Potência e/ou Velocidade do Vento (Soman *et al.*, 2010);
- Monitorização de Parques Eólicos:
    - Monitorização com base em Métodos Estatísticos (Guo *et al.*, 2009);
    - Monitorização com base em Curvas de Performance (Kusiak *et al.*, 2009; Kusiak e Verma, 2013);
    - Monitorização de Falhas (Anaya-Lara *et al.*, 2006; Brandão *et al.*, 2009; Guo *et al.*, 2009; Aziz *et al.*, 2010; Daneshi-Far *et al.*, 2010).

2.3. Análise de Curvas para Monitorização de Operação

A curva de potência de um aerogerador demonstra as condições a que este sistema opera, assim como os seus índices de produtividade. Cada tipo de turbina distinta tem uma curva de potência certificada (Kusiak, 2010), uma operação realizada por parte dos fornecedores dos aerogeradores (Figura 5).

A análise subjacente a uma curva de potência real de um aerogerador permite:

- prever o impacto de condições atmosféricas adversas;
- monitorizar, detetar e prever falhas ou valores inesperados no sistema físico do aerogerador;
- apoiar a identificação de tempos de manutenção optimizados.

Assim, a análise da curva de potência demonstra como a turbina reage a uma determinada velocidade do vento. Algumas das razões para a existência de desvios nos dados obtidos são (Aguiar, 2009):

- turbulência;
- efeitos de esteira;
- complexidade do terreno;
- problemas nos controlo de potência;
- deterioração e danificação das pás rotativas;
- condições climatéricas e do meio;
- problemas nos programas de controlo;
- operação com limite de potência.

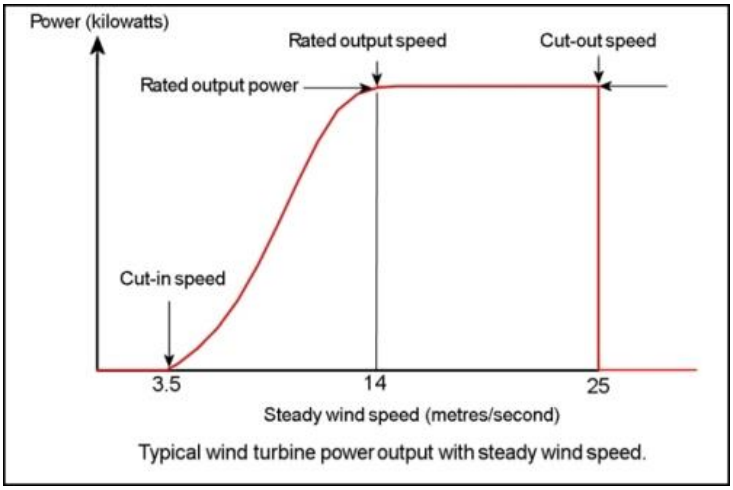

Figura 5 – Curva de Potência genérica de um aerogerador (Wind Power Program UK, 2009)

Em (Kusiak e Verma, 2013), os autores referem o recurso a três tipos curvas energéticas, capazes de auxiliar na monitorização do sistema associado aos aerogeradores.

Estas curvas são:

- Curva de Potência (tendo em conta a correlação entre as variáveis da Potência e da Velocidade do Vento) (Figura 6);
- Curva do Rotor (tendo em conta a correlação entre as variáveis da Velocidade do Rotor e da Velocidade do Vento) (Figura 7);
- Curva do Blade Pitch (tendo em conta a correlação entre as variáveis do ângulo do *Blade Pitch* e da Velocidade do Vento) (Figura 8).

Os diferentes diagramas estão representados de seguida.

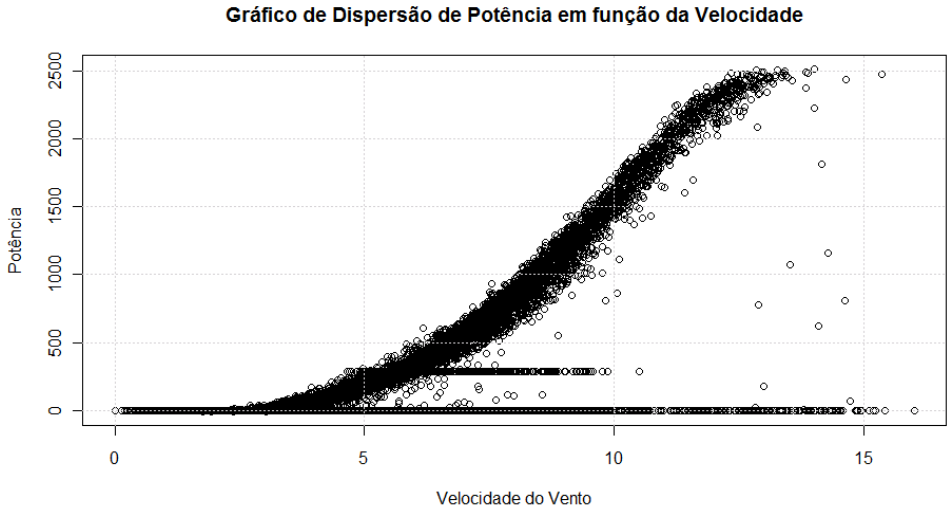

Figura 6 – Potência Gerada Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010

É possível identificar a existência de alguns registos dispersos, localizados em áreas isoladas dos gráficos (de baixa densidade).

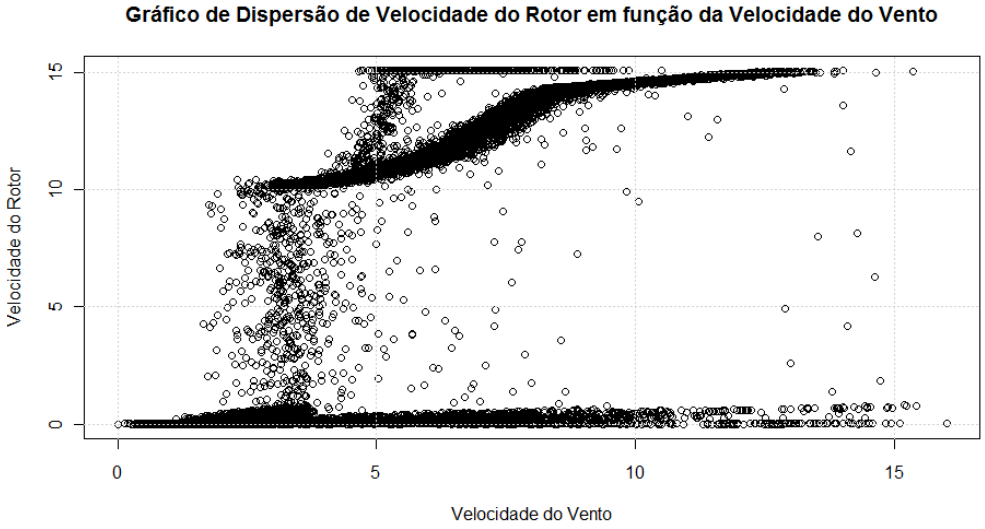

Figura 7 – Velocidade do Rotor Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010

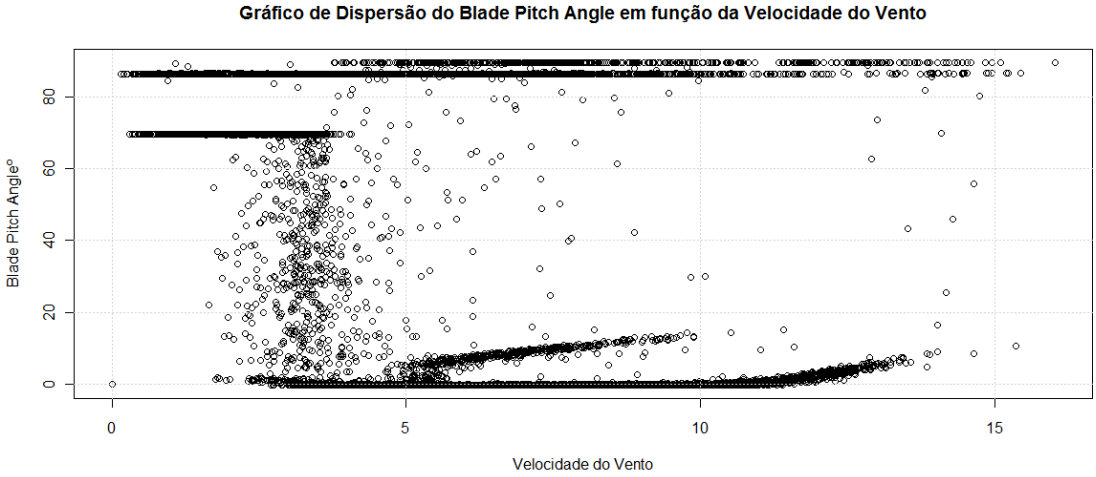

Figura 8 – Ângulo do Blade Pitchº Vs. Velocidade do Vento – Aerogerador 1, Parque X Meses de janeiro, fevereiro e março de 2010

Capítulo 3

Estado da Arte - Deteção de Outliers

One man's noise is another man's signal.

3.1. Definição de Outliers

Num conjunto de dados um *outlier* é um ponto significantemente diferente dos restantes registos observados, podendo também ser denominado por anormalidade, desvio, registo discordante, anomalia, novidade, evento, falha ou intruso (Gupta *et al.*, 2013).

A definição de *Outliers* tem sido alvo de diferentes caracterizações ao longo dos tempos. Segundo (Munõs-Garcia *et al.*, 1990), o estudo associado às observações anormais num conjunto de registos é *"uma das mais importantes etapas, em qualquer análise estatística de dados, pois permite estudar a qualidade das observações…"*. Já segundo (Barnett e Lewis, 1994) *"um outlier é uma observação* (ou um subconjunto de observações, denominado então de *outlier* coletivo) *que parece ser inconsistente quando comparado com os restantes dados"*.

Hawkins (Hawkins, 1980) definiu formalmente o conceito de uma anomalia como:

> *"Um outlier é uma observação que se desvia tanto das restantes observações de forma a criar suspeitas de que foi gerada por um mecanismo diferente"*.

A definição de Hawkins é considerada a mais abrangente e capaz de identificar as principais características associadas à análise de conjuntos de dados:

- conjuntos de dados são normalmente blocos de registos multivariados (ou multidimensionais);
- pode existir mais do que um mecanismo gerador de dados;
- anomalias podem ser criadas por mecanismos diferentes do que os que geram os dados normais.

Nas últimas décadas, o estudo de registos anormais tem sido realizado em diferentes áreas, com especial relevância para: dados multivariados e de alta dimensionalidade, dados incertos, fluxos contínuos de dados (*datastreams*), dados provenientes de redes informáticas e dados com um fator temporal agregado (séries temporais). De referir que registos *outliers* contêm informações úteis sobre características anormais de sistemas complexos que têm impacto no processo de geração de dados.

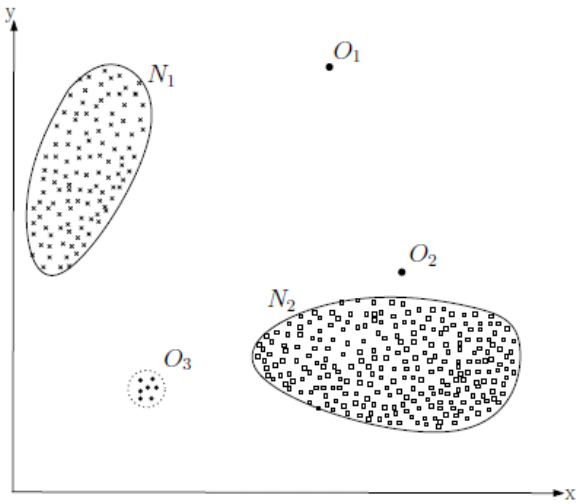

Figura 9 – Um exemplo simples de outliers num espaço bidimensional composto por duas variáveis: X e Y (Chandola *et al.*)

Na Figura 9 estão apresentados diferentes tipos de anomalias. Enquanto O1 e O2 representam dois registos *outliers*, O3 representa uma região de *outliers*. Já as regiões identificadas como N1 e N2 representam as regiões com dados ditos normais.

A preocupação subjacente ao estudo de observações anómalas é antiga e data das primeiras tentativas de análise de dados. Desde as primeiras experiências realizadas por Bernoulli (datadas de 1777) que se obtêm referências a tratamento de dados considerados discordantes. Dos registos obtidos da altura, verifica-se que a prática de

simplesmente se rejeitar essas observações era uma atividade comum e considerada usual. A discussão que envolvesse registos *outliers* centrava-se única e exclusivamente na justificação da sua rejeição do âmbito de análise.

Atualmente, as possíveis ações a executar quando se processam *outliers* sofreram alterações. No entanto denota-se que as opiniões não são claras e dependem muito do problema em discussão. Enquanto alguns cientistas defendem a sua rejeição, pois são registos "considerados inconsistentes quando avaliados com os restantes", outros discordam e defendem que devem ser tidos em consideração para as análises finais e assim contribuir com igual peso para as conclusões. Em ambos os casos é considerado que existe uma certa subjetividade sobre o tratamento dado a estes registos.

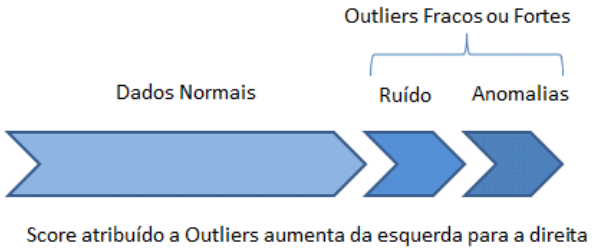

Figura 10 - Identificação dos diferentes níveis de valor na informação

Existem várias aplicações que utilizam a dinâmica de deteção de *outliers* nas mais diferentes áreas como forma de fornecerem novas funcionalidades aos seus utilizadores. As mais conhecidas são: deteção de fraudes, medicina, saúde pública, análise de ensaios clínicos, deteção de irregulariedades em eleições, intrusões em redes informáticas, previsões metereológicas, sistemas de informação geográfica, análise da condição física de atletas, entre outras.

No âmbito de deteção de *outliers*, e dependendo do tipo de aplicação informática em desenvolvimento, torna-se ainda importante distinguir entre ruído e anomalia (Figura 10), para além de caracterizar os diferentes cenários de deteção de registos anómalos em: Cenário Supervisionado, Cenário Semi-Supervisionado e Cenário Não Supervisionado.

Cenário Supervisionado

Designa-se por ***Cenário Supervisionado*** quando no conjunto de treino os registos normais e anómalos vêem caracterizados como tal. É importante referir que neste caso o

problema de classificação pode ser altamente desbalanceado. Este cenário apresenta como principais desvantagens:

- a necessidade da existência de dados pré-identificados como anomalias;
- a demora quando se tentam identificar novos tipos de eventos raros (Pokrajac *et al.*, 2007) pois estes podem não estar previamente classificados.

Cenário Semi-Supervisionado

No caso de ***Cenário Semi-Supervisionado*** apenas são fornecidos os registos considerados normais ou os registos considerados anómalos. A classe que não seja fornecida tem de ser identificada pelo algoritmo.

Cenário Não Supervisionado

Já quando se está na presença de um ***Cenário Não Supervisionado*** não existe um conjunto de treino corretamente caracterizado, sendo necessário processar os dados para encontrar os registos normais e os anormais (cenário mais utilizado, uma vez que não faz assumpções sobre os dados e sobre a classificação a atribuir aos dados).

3.2. Tipos de Outliers

Existem algumas denominações mais alargadas do conceito de *outliers*.

Em alguns casos específicos as anomalias necessitam de ser referenciadas não como registos individuais mas como agrupamentos de pontos (registos anómalos). A análise subjacente a conjuntos de anomalias é uma das áreas em desenvolvimento.

De seguida serão exploradas as diferentes denominações de *outliers*.

3.2.1. Outliers Globais ou Pontuais

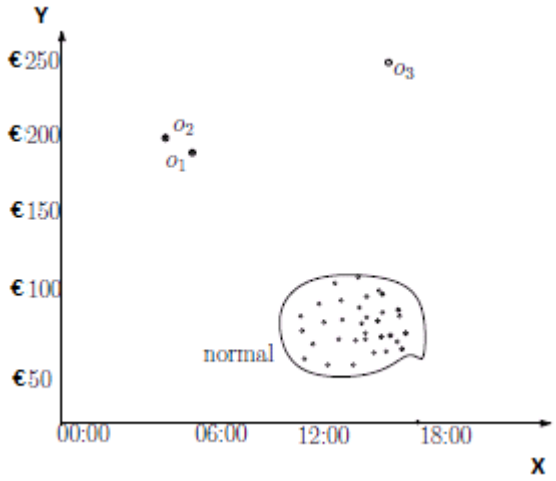

Figura 11 – Exemplo de Outliers Globais (também denominados de Pontuais) num espaço bidimensional (Chandola *et al.*)

Num determinado conjunto de dados, um registo é um *outlier* global se se desvia significantemente dos restantes registos. *Outliers* globais são também denominados de *Anomalias Pontuais* e são a forma mais simples de *Outliers* (Figura 11).

3.2.2. Outliers Condicionais

Existem registos que apenas são considerados *outliers* se tivermos em atenção a localidade e a região a que se referem. Por exemplo, se estivermos no Verão no Porto a temperatura de 10º é um *outlier*. Se estivermos no Inverno, este valor será considerado normal. Assim, a definição de *Outliers Condicionais* pode ser dada por:

> *"Num determinado conjunto de dados, um registo é um outlier condicional se se desviar significantemente dos restantes valores, em relação a um determinado contexto ou característica especial."*

Neste caso, os atributos ou variáveis podem ser categorizados em dois tipos distintos:

- *Atributo Contextual* – define o contexto em que o registo foi obtido. Por exemplo, coordenadas GPS dadas pelos atributos de latitude e longitude de uma determinada localização de terreno; ou o atributo de tempo quando se processa uma Série Temporal, uma vez que determina a posição do registo no âmbito da série em análise (caso específico da presente dissertação).

- *Atributo de Comportamento* – define as características do objecto.

Os *outliers* de contexto são mais usuais quando se efectua uma análise de séries temporais, uma vez que definem ou auxiliam a definir determinadas situações como a existência de sazonalidade.

3.2.3. Outliers Coletivos

Um subconjunto dos registos forma um *outlier coletivo* se os registos como um todo se desviarem dos restantes registos. No entanto, é possível que os registos como valores individuais não formem um *outlier* global. Estes *outliers* apenas têm consistência se surgirem numa sequência de dados ou se forem registos espaciais (Figura 12).

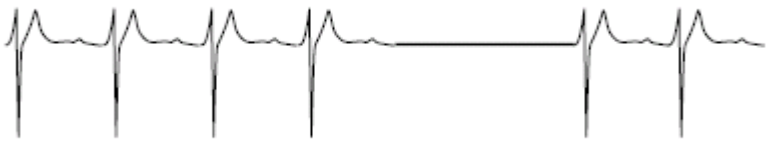

Figura 12 - Exemplo de um Outlier Coletivo – eletrocardiograma (Chandola *et al.*)

3.3. Retorno de Algoritmos de Identificação de Outliers

Usualmente, algoritmos de deteção de *outliers* têm 2 retornos diferentes:

- *Labelling* – corresponde a identificar cada registo como registo *outlier* ou normal através da criação de um novo atributo (atributo classe); neste caso o algoritmo comporta-se de forma idêntica a um algoritmo de classificação;
    - Vantagem – exato pois fornece uma lista de *outliers*;
    - Desvantagens – não fornece o grau com que um determinado registo é considerado *outlier*; não ordena os resultados consoante o grau de anomalia;
- *Scoring* – técnica que associa a cada registo de dados um valor que regista o grau com que esse registo é considerado anómalo;
    - Vantagem – fornece uma lista ordenada de *outliers*, muitas vezes associado a um limite de significância *(threshold)*;

- Desvantagem – a necessidade de definir o valor a ser utilizado como limite para o número de *outliers* a serem analisados é por vezes difícil de identificar, sendo necessária alguma experiência para avaliar os dados.

3.4. Desafios na Deteção e Classificação de Outliers

Como já foi mencionado, a preocupação sobre a identificação de *outliers* num conjunto de dados é antiga. Apesar de inicialmente se pensar que estes registos deveriam ser retirados do conjunto de dados em estudo é necessário ter alguns cuidados, pois uma avaliação cuidada pode permitir identificar as causas que levaram ao seu aparecimento.

	Dados Univariados	Dados Bivariados	Dados Multivariados
Definição:	Descreve um conjunto de registos em relação a um único atributo. Não prevê relações entre atributos.	Analisa dois atributos em simultâneo. Prevê a existência de ligações entre atributos.	Permite a análise entre duas ou mais variáveis num mesmo espaço dimensional.
Representação Gráfica:	Gráficos de Barras; Histogramas; Pie-Chart; Line Graph; Diagrama de Caule e Folhas; Box –Plot; Q-Q Plot.	Gráfico de Dispersão; Gráfico de Box-Plot Bidimensional (Tongkumchum, 2005); *Kernel Density Estimation*.	Profilling de Dados Multivariados; Transformação de Andrew's Fourier; Gráfico de Dispersão; Gráficos de Contorno.
Métodos:	Procedimento Simples e Sequencial; Procedimento *Inward* e *Outward*; Medidas Robustas Univariadas.	---	*Masking effect; Swamping effect;* Medidas Robustas Multivariadas.
Medidas Computacionais:	Medidas de Tendência Central; Medidas de Dispersão; Medidas de *Skewness*; *One-way ANOVA*; *Index Numbers*; Regressão e Correlação Simples.	Regressão Simples; Correlação Simples; *Two-way ANOVA*; Associação de Atributos.	Análise de *Clusters* e Análise de Componentes Principais; Análise Factorial; *Multi-ANOVA*; Análise Discriminante Múltipla; Regressão e Correlação Múltipla.

Tabela 2 - Comparação de Métodos de Análise de Dados Univariados, Bivariados e Multivariados, adaptado de (Llango *et al.*, 2012)

	Métodos Supervisionados	Métodos Semi Supervisionados	Métodos Não Supervisionados
Definição:	Requerem conhecimento da classe de dados normais e anormais. Deve-se construir um classificador.	Requer conhecimento da classe de dados normais. Utiliza modelo de classificação modificado para aprender o comportamento normal e detetar quaisquer desvios para o(s) comportamento(s) anómalo(s).	Assume que os objectos normais estão de alguma forma agrupados e que cada grupo tem características distintas. Um *outlier* encontra-se bastante afastado destes grupos de registos normais.
Vantagens:	Modelos que podem ser facilmente compreendidos. Bons resultados na deteção de vários tipos de anomalias.	O comportamento normal pode ser facilmente aprendido.	---
Desvantagens:	Requer labels para registos normais e anormais. Não consegue detetar anomalias que comecem a ser registadas. Requer alguns registos anómalos para estes serem detectados.	Requer a identificação para a classe normal. Tem indíce de falsos positivos alto – registos anteriormente considerados normais podem ser identificados como anómalos.	As técnicas não supervisionadas normalmente têm níveis de alarme falsos mais elevados, uma vez que as assumpções consideradas podem não se verificar em alguns dos casos de registos obtidos. Não deteta eficientemente *outliers* coletivos. Registos normais podem não partilhar características em comum, podendo os *outliers* partilhar indíces em comum numa área pequena.
Técnicas:	Redes Neuronais, Estatísticas Bayesianas, Modelos baseados em Regras, Aprendizagem baseada em Árvores de Decisão	*One Class SVM*, Modelo de *Markov*	*K-Means, LOF (Local Outlier Factor), CLAD, CBLOF, COF, DBSCAN, SNN*
Ferramentas: Weka, SPSS, SAS, Tanagra, SYSTAB, MATLAB, Minitab, Liseral, MEDCALC, R-packages			

Tabela 3 - Distinção de Técnicas de Deteção de Outliers baseadas em Machine Learning (Llango *et al.*, 2012)

Existem alguns métodos estatísticos que auxiliam na identificação de *outliers*:

- gráficos de Box-Plot;
- modelos de Discordância;

- teste de Dixon ou de Grubbs;
- Z-Scores.

Para deteção de *outliers* um dos pontos principais passa por identificar o tempo da ocorrência, que pode ou não ser reconhecido. Neste campo alguns desafios conhecidos são:

- necessidade de algoritmos eficientes e escaláveis;
- necessidade de proteção de dados privados (maioritariamente relacionado com compras online), não sendo possível registar todo o tipo de dados;
- algoritmos com bons indíces previsionais.

O processo de identificação de *outliers* é um processo não linear que depende da estrutura inerente aos registos, da sua correlação e da correspondente aglomeração. Os diferentes algoritmos variam consoante o tipo de dados em análise. Estes podem ser: registos nominais, registos ordinais, **registos intervalares**, registos proporcionais, registos binários, registos contínuos, **registos discretos**, registos transacionais, registos espaciais, registos espaço-temporais, dados sequenciais ou **dados de séries temporais**. Podem ainda variar consoante sejam dados univariados (com um único atributo), dados bivariados (com dois atributos) ou dados multivariados (se tivermos em atenção dois ou mais atributos).

Os diferentes mecanismos de identificação de *outliers* estão descritos na Tabela 2. De referir que os mecanismos visuais mais utilizados passam pela representação gráfica da aplicação de diferentes algoritmos, existindo um conjunto de medidas computacionais que permitem identificar melhor quais os métodos a utilizar.

3.5. Técnicas ou Metodologias de Deteção de Outliers

Existem diferentes técnicas de análise de *outliers*. Nesta área é importante distinguir entre:

- Técnicas Estatísticas
- Técnicas ou Metodologias baseadas em Profundidade (*depth-based*)
- Técnicas ou Metodologias baseadas em Distância (*distance-based*)
- Técnicas ou Metodologias baseadas em Densidade (*density-based*)
- Técnicas ou Metodologias baseadas em Agrupamento (*clustering-based*)

De referir que na literatura é usual criar-se outra ramificação para Técnicas ou Metodologias aplicáveis a grandes volumes de dados, uma vez que o tratamento deste tipo de registos pressupõe a aplicação de diferentes técnicas de programação, que evoluem ao longo do tempo (p.e. Programação Paralela).

3.5.1. Técnicas Estatísticas

Técnicas Estatísticas direcionadas para identificação de *outliers* tem como pontos-chave:

- associar à geração dos dados normais uma determinada distribuição estatística (por exemplo distribuição gaussiana, normal ou de *t-student*);
- obter as estatísticas descritivas robustas considerando o conjunto de dados (de referir que a *média* e o *desvio padrão* são altamente sensíveis à existência de *outliers*);
- identificar os *outliers* como os pontos que têm uma baixa probabilidade de serem gerados pela distribuição (p.e. desviam-se da média em mais de 3x).

Os testes estatísticos a serem utilizados dependem:

- do tipo de distribuição de dados;
- do número de atributos envolvidos;
- do número de distribuições envolvidas (pode-se tratar de distribuições mistura);
- da identificação da necessidade de utilização de modelos paramétricos ou de modelos não paramétricos.

A vantagem do uso de técnicas estatísticas passa pela modelação com recurso a distribuições bastante estudadas e conhecidas. A desvantagem passa por ser uma técnica mais utilizada em deteção univariada de *outliers*, podendo ser complexo o seu recurso para dados bivariados ou multivariados.

3.5.2. Técnicas ou Metodologias Baseadas em Profundidade (Depth-based Approaches)

Algoritmos baseados em Profundidade têm como conceito fundamental pesquisar por *outliers* nas zonas limitrofes do espaço de dados. Para tal é necessário organizar os

dados em camadas (*convex hull layers*), em que as anomalias são os objectos localizados nas camadas externas. Para o funcionamento destes algoritmos, a assumpção generalizada é de que os registos normais estão localizados no centro do espaço de dados.

Nesta área já existem alguns algoritmos implementados:

- ISODEPTH (Ruts e Rousseeuw 1996 – package depth no R);
- FDC (Johnson *et al.*, 1998);
- Ellipsoid (biblioteca MVE do R);
- Convex Peeling.

Os algoritmos ou metodologias baseadas em profundidade têm uma correlação com as metodologias clássicas baseadas em análise estatística, conseguindo no entanto manter-se independentes da distribuição de dados inerente aos registos. Este tipo de computação é eficiente em visualizações de espaço de dados a duas ou três dimensões.

As grandes vantagens passam por não ser necessário ajustar os dados a uma distribuição e não requerer o cálculo de distâncias entre registos. A grande desvantagem passa por serem ineficientes face a grandes volumes de dados, uma vez que as diferentes camadas podem ser difíceis de identificar.

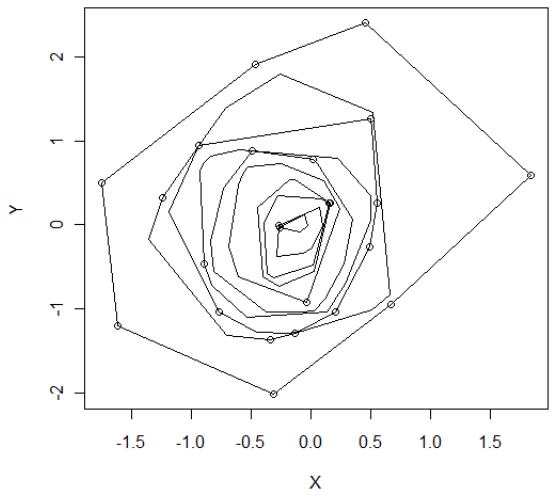

Figura 13 - Execução do IsoDepth num conjunto de dados bivariados no R

3.5.3. Técnicas ou Metodologias baseadas em Distância (Distance-based Approaches)

O modelo básico que deu origem às técnicas baseadas em distância foi criado por Knorr e Ng em 1997. Este modelo diz que: dado um raio ε e uma percentagem τ,um dado ponto $\wp$ é um *outlier* se no máximo τ % de todos os outros pontos tiverem uma distância a $\wp$ menor do que ε.

As abordagens relacionadas com o cálculo de distâncias têm uma ideia simples: ***julgar um registo baseado na distância(s) aos seus vizinhos***.

O conceito base deste tipo de algoritmos passa então pela noção de vizinhança de um ponto, conhecido tipicamente como a pesquisa dos *k* vizinhos mais próximos. As assunpções requerem que os registos normais tenham uma vizinhança densa e que os registos anómalos se localizem em regiões de densidade esparsa.

A vantagem passa por evitar a computação excessiva que pode ser necessária para encontrar uma distribuição estatística para os dados em análise e na subsequente seleção de testes de discordância. A desvantagem passa por analisar *outliers* locais em conjuntos de dados que podem sofrer de densidades diferentes, em diferentes horizontes temporais.

Neste caso existem três abordagens distintas que serviram de base à grande generalidade de algoritmos que entretanto surgiram:

- Algoritmo baseado em Indexação (Knorr e Ng., 1998);
- Algoritmo baseado em Dupla Iteração (Knorr e Ng., 1998);
- Algoritmo baseado em *Grid* (Knorr e Ng., 1998).

Com base nos três algoritmos base referidos anteriormente, foram criados:

- Nested-Loop (Ramaswamy *et al.*, 2000);
- Linearization (Angiulli e Pizzuti, 2002);
- ORCA (Bay e Schwabacher, 2003);
- RBRP (Ghoting *et al.*, 2006);
- ROF - *Resolution-based Outlier Factor* (Fan *et al.*, 2006).

3.5.4. Técnicas ou Metodologias baseadas em Densidade (Density-based Approaches)

O principal conceito associado a metodologias baseadas em densidade passa por estimar a distribuição de densidade dos dados e identificar os *outliers* como aqueles registos localizados em zonas de densidades mais baixas. Ou seja, a ideia geral passa por comparar a densidade à volta de um determinado ponto com a densidade dos seus pontos vizinhos. A densidade relativa de um ponto é computada como um *indíce* que fornece o grau de anomalia. Os diferentes algoritmos diferem na forma como a densidade é calculada.

As vantagens passam por poder detetar *outliers* que poderiam não ser detetados com critérios globais. A desvantagem passa pela dificuldade de identificar os parâmetros de limites superiores e inferiores, pois podem ser difíceis de descortinar.

Os principais algoritmos baseados em densidade são:

- LOF - *Local Outlier Factor* (Breunig *et al.*, 1999 e 2000) – considera densidade relativa de dados (Figura 15 – Pseudo-Código do);
- K-Distance;
- K-Distance Neighborhood;
- Reach Ability Distance;

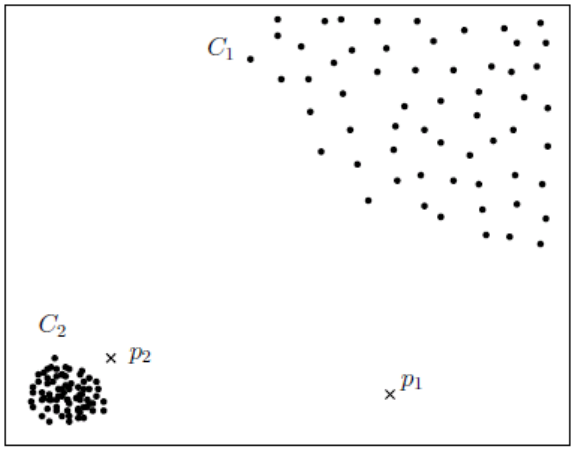

Figura 14 –Identificação de Outliers em dados com diferentes níveis de Densidade

3.5.5. Técnicas ou Metodologias baseadas em Agrupamento (Clustering-Based)

As metodologias baseadas em agrupamento identificam grupos de objetos altamente relacionados que podem funcionar como resumo de um conjunto de registos.

Um registo é um *outlier* se: não pertence a nenhum grupo; se existir uma grande distância para com outros grupos; ou se pertencer a um grupo pequeno ou esparso.

Estes algoritmos detetam *outliers* sem necessitar de dados pré-classificados (são não supervisionados). Quando a lista de anomalias é obtida é apenas necessário comparar os registos com os grupos sugeridos pelo método de agrupamento utilizado. Estes algoritmos são habitualmente rápidos e computacionalmente pouco custosos. A desvantagem é que o método de agrupamento utilizado pode não identificar a estrutura associada às anomalias.

Pseudocódigo (Local Outlier Factor)

1. Para cada registo q, calcular a distância $k - distance(q)$ como a distância ao $k - th$ vizinho mais próximo
2. Calcular as distâncias de acessibilidade para cada registo q em relação ao registo p como:

   $$reach - dist_k\,(q,p) = \max(d(q,p), k - distance(p))$$

   onde $d(q,p)$é a distância euclidiana de q a p.
3. Calcular densidade de acessibilidade local (lrd) do registo q como o inverso da média das distâncias de acessibilidade baseadas nos k vizinhos mais próximos do registo q.

   $$lrd(q) = \frac{1}{\sum_{p \in kNN(q)} reach - dist_k(q,p)/k}$$

4. Calcular LOF para cada registo q como o rácio entre a densidade de acessibilidade local de q's k vizinhos mais próximos e a densidade de acessidade local do registo q.

   $$LOF(q) = \frac{\frac{1}{k}\sum_{p \in kNN(q)} lrd(p)}{lrd(q)}$$

Figura 15 – Pseudo-Código do LOF

Nesta área já foram criados vários algoritmos:

- BIRCH;
- CLARANS;
- DBSCAN;
- GDBSCAN;
- OPTICS;
- PROCLUS;
- CBLOF.

3.6. Ferramentas de Mineração de Dados para Deteção de Outliers

Atualmente existem várias ferramentas com metodologias de identificação de *outliers* implementadas. Dentre as ferramentas mais conhecidas refere-se:

- R (package *outliers* (Komsta, 2011) entre outras);
- o MOA (*Massive Online Analysis* (Bifet *et al.*, 2010));
- o Tanagra (Rakotomalala, 2005);
- o ELKI (Kriegel, 2012);
- além do próprio MATLAB (Erdoğan, 2011).

As diferenças entre as diferentes ferramentas são principalmente: tipo de registos que permitem, tipo de algoritmos implementados e formatos de leitura de dados. De salientar que, caso as bibliotecas de algoritmos estejam disponibilizadas na internet, existe a possibilidade, após uma análise cuidada das estruturas necessárias, de se traduzir o código para outras linguagens de programação. Assim torna-se possível criar uma ferramenta adaptada para um problema específico.

3.7. Estimação de Massa

De seguida será explicada a dinâmica associada à estimação de massa – um mecanismo base de modelação de dados empregue para solucionar vários problemas relacionados com *machine learning*. É importante referir que esta é uma área recente em que se considera que estimação de massa pode ser uma medida de grande importância para identificação de anomalias – a génese da presente dissertação.

Nas ferramentas informáticas definidas anteriormente não foram encontradas bibliotecas de aplicação de algoritmos de estimação de massa.

3.7.1. Definição de Massa

Estimação de massa é um mecanismo recente, introduzido por (Ting *et al.*, 2010), e mais tarde sistematizado em (Ting *et al.*, 2011; Ting *et al.*, 2013). Atualmente possui referência em diferentes áreas de onde se salienta: deteção de anomalias em fluxos contínuos de dados, pesquisa de informação e regressão.

Na sua formulação mais básica, estimação de massa indica que para se obter a massa de um grupo de pontos:

> *"Apenas é necessário contabilizar os pontos existentes nessa região."*

Assim, independentemente do formato da área obtida, quaisquer dois grupos de pontos com idêntico número de registos lá pertencentes possuem o mesmo valor de massa.

Os modelos de informação criados com base em estimação de massa possuem dimensões bastante reduzidas, especialmente quando comparadas com as amostras necessárias para modelos baseados em densidade.

Assim, para criar um modelo de informação para estimação de anomalias é necessário especificar o volume de dados que fazem parte do conjunto de treino e do conjunto de teste, também denominado como conjunto de validação do modelo de dados criado. Genericamente, quando o conjunto de treino e de teste for definido como correspondendo à totalidade dos dados em análise, o resultado final passa por uma ordenação dos registos em que os de *pontuação* inferior são as anomalias. Esta ordenação poderá ser mais grosseira ou mais profunda, consoante o número de divisões do espaço de dados pedido pelo utilizador ou o número de árvores a considerar.

As principais vantagens associadas à criação de modelos baseados em estimação de massa são:

- tempos de processamento bastante reduzidos;
- gestão adequada da capacidade de memória das máquinas utilizadas para gerir os modelos;
- não necessitar de aperfeiçoamento nos parâmetros de entrada (no caso do *Local Outlier Factor* pode ser necessário aperfeiçoar a pesquisa dos *k* vizinhos mais próximos).

3.7.2. Análise de Conjuntos de Dados (*Emsemble Analysis*)

Análise de Conjuntos de Dados, atividade mais conhecida como *Emsemble Analysis* é um método utilizado na área de *Datamining* para reduzir a dependência do modelo criado para um determinado conjunto de dados. Dessa forma pretende-se aumentar a robustez de execução do método em diferentes pacotes de dados. Esta metodologia é utilizada em problemas de agrupamento ou de classificação de dados e tem a ideia geral de combinar diferentes modelos para criar um modelo geral mais robusto.

Na área de identificação de *outliers* a criação de um modelo com base em conjuntos de dados criados a partir de diferentes amostragens do conjunto de treino é uma temática que começa a surgir em diferentes trabalhos e livros, com especial ênfase para (Aggarwal, 2013). Esta necessidade verifica-se essencialmente no âmbito de dados com grandes dimensões (vários atributos para um mesmo registo). Assim, o método de estimação de massa proposto pressupõe a criação de diferentes conjuntos de treino, capazes de fornecer diferentes cortes sobre um mesmo espaço de dados, com base em diferentes amostragens aleatórias. Ou seja, um conjunto de árvores (floresta de items) é criado com base em diferentes amostragens do conjunto de treino pré-carregado no sistema. Todas as árvores terão uma dimensão idêntica.

3.7.3. Half-Space Trees e Estimação de Massa

Para estimação de massa foi proposto o recurso a uma estrutura de dados denominada de *Half-Space Tree (HS-Tree)*.

Num conjunto de dados de *D* atributos distintos, uma *Half-Space Tree* é construída iterativamente através da criação de um conjunto de nós filhos. Em cada nó pai é calculado aleatoriamente a dimensão de corte (qual o atributo a partir do qual o corte vai ser executado). Posteriormente é calculado o ponto de corte como o ponto médio existente para o espaço de dados da árvore. Existem referências de que o ponto de corte poderá ser aleatório, não tendo sido implementada nem testada esta característica. Poderá ser desenvolvida em âmbito de trabalho futuro.

Genericamente, as *Half-Space Trees* permitem criar uma hierarquia de partições independentes (células), e assim fornecer grupos de registos com massa idêntica. Com base nesta hierarquia é criado um índice para cada registo em que:

- a pontuação mais baixa identifica as possíveis anomalias;
- a pontuação mais elevada identifica os registos considerados normais.

Sendo *h* a profundidade de uma *Half-Space Tree*, após a sua criação esta possui 2^{h+1}-1 nós, caso todos os nós estejam a uma profundidade idêntica. Ou seja, se for uma árvore completamente populada de profundidade 15, esta terá 65335 nós.

Assim, é possível definir *Half-Space Trees* como estruturas de dados em que:

1) cada nó interno executa um corte médio do espaço de dados em dois subespaços de igual dimensão;

2) em cada nó externo o término de cortes é executado com base em restricções baseadas na profundidade atual do nó e no número máximo de registos de massa a contabilizar.

Todos os nós começam por registar a massa dos dados de treino nos seus próprios subespaços. Posteriormente calculam a pontuação do conjunto de teste com base nos dados de treino assimilados.

De referir que as principais vantagens de se utilizar *Half-Space Trees* para deteção de anomalias são:

- tempo de execução rápido;
- baixas necessidades computacionais em termos de memória (não necessita de cálculos de distância ou de densidade para cada registo de dados, facilitando a criação de estruturas de dados de gestão da informação).

De seguida serão explicados os métodos principais para criar *Half-Space Trees*.

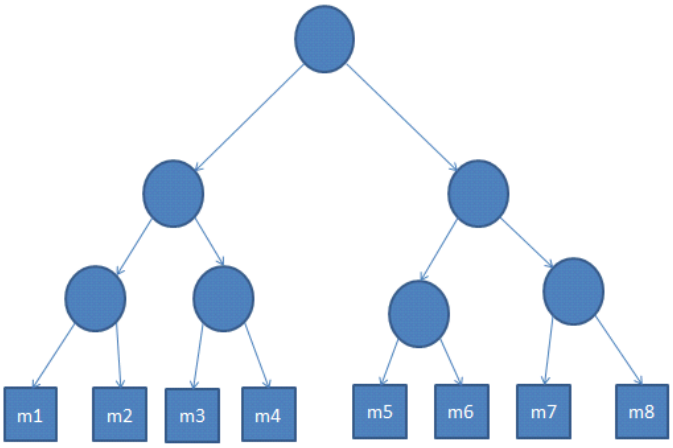

Figura 16 - Half-Space Tree sem limite de massa ativo

O procedimento de criação de uma *Half-Space Tree* inicia-se com a geração de um espaço de trabalho aleatório, tendo em conta os diferentes atributos do espaço de dados. A definição deste espaço de trabalho é fornecida com a função ***InicializarEspacoDados*** (Figura 17)**.**

Criação de uma Half-Space Tree

```
InicializarEspacoDados(Amostra S){
    Para cada atributo q na amostra S{
        encontrar zq como sendo o ponto aleatório no intervalo [Sminq, Smaxq]
        redefinir o atributo q como tendo novo intervalo [minq, maxq] = [zq-r, zq+r]
        sendo r = 2 * max(zq – Sminq, Smaxq – zq)
    }
    retornar o espaço de dados válido para cada atributo do espaço de dados
}
```

Figura 17 – Pseudocódigo para Inicializar Espaço de Trabalho

Construção de Half-Space Trees

A Figura 18 demonstra o procedimento para criar uma *Half-Space Tree*, em que cada nó interno é construído com a seleção aleatória da dimensão de corte *q* em dois espaços de dimensão idêntica (ponto de corte médio).

Entradas: min & max – *arrays* com o valor mínimo e máximo para cada dimensão do Espaço de Trabalho

```
            k – nível de profundidade corrente
Saídas: uma HS-Tree
ConstruirArvore(min, max, k)
{
    Se k == maxProfundidade então
        Retornar Nó (nulo)
    Senão
        Aleatoriamente selecionar dimensão q
        p = (maxq + minq) / 2
        {
            Construir dois nós (Esquerdo e Direito) com base num corte em dois espaços idênticos
        }
        temp = (maxq); maxq = p;
        Esquerdo = ConstruirArvore(min, max, k+1);
        maxq = temp; minq = p;
        Direito = ConstruirArvore(min, max, k+1);
        Retornar Nó(Esquerdo, Direito, AtributoCorte = q, PontoCorte = p)
}
```

Figura 18 – Pseudocódigo para Criar uma Half-Space Tree

Durante a criação da Half-Space Tree torna-se importante avaliar se foi atingido:

- o limite de profundidade;
- o limite de dimensão.

É com base no limite de profundidade e no limite de dimensão dos nós que é possível identificar a estrutura da árvore criada. Caso o limite de profundidade seja muito reduzido as aglomerações do espaço de dados podem ser muito grandes, aumentando o valor de pontuação de anomalia atribuída aos registos de teste; caso seja atribuído um limite de dimensão por nó pequeno as diferenças de pontuação entre os diferentes registos poderão ser muitas, havendo a criação de mais grupos para distinguir diferentes patamares de anomalia. Existe assim a necessidade de distinguir entre uma árvore com nós a uma dimensão idêntica e com massas diferentes (Figura 16) e árvores com limite de massa ativada (Figura 19).

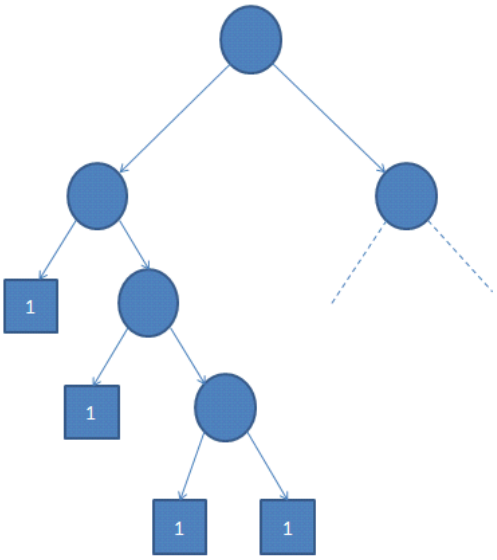

Figura 19 - Exemplo de Half-Space Tree com limite de dimensão

Função de Pontuação (Scoring)

Com base numa distribuição de massa não uniforme é possível estabelecer uma ordenação entre diferentes nós da árvore com base em:

$$m[i] * 2^i < m[j] * 2^j$$

É exatamente através da ordenação criada pela expressão anterior que se pode criar a função de pontuação de um determinado ponto *x* como:

$$s(x) = m[l] * 2^l$$

em que *l* designa a profundidade corrente do nó e $m[l]$ designa o número de pontos existentes atualmente no nó externo. É com base nesta expressão que são atribuídas as pontuações de anomalia aos diferentes registos do conjunto de teste.

Caso todos os nós externos estejam à mesma profundidade (Figura 16), então a função de pontuação passa a ser apenas em relação ao número de pontos existentes num determinado nó. Quando o limite de elementos por nó seja 1 (Figura 19), o *scoring* de anomalia depende apenas da altura *l* do nó até à raiz da árvore. Neste caso a profundidade do nó torna-se num *sinónimo* de massa do nó. Esta característica fornece a definição de dois tipos distintos de *Half-Space Trees* – as árvores baseadas em massa e as árvores baseadas em massa aumentada (*augmented mass estimation*).

Quando se analisa a função de pontuação para o conjunto de modelos agregados é importante referir que a estimativa de massa de um determinado ponto é dada pela

média de massas das diferentes regiões que contenham esse ponto, uma vez que a floresta de *Half-Space Trees* é criada com base em diferentes regiões de dados que se sobrepõem.

Resumidamente, a função de pontuação para a distribuição de massa de um ponto *x* é estimada usando *T* Half-Space Trees:

$$s_a(x) = m(T_a^h(x|D) * 2^{la}$$

$$m(x) = \frac{1}{T}\sum_{a=1}^{T} s_a(x)$$

Figura 20 – Função de Pontuação de um registo com base numa Floresta de Half-Space Trees

De notar que cada árvore distinta tem um espaço de dados distinto calculado com base na amostra de dados de treino encontrada para si.

Árvore de Cortes Aleatórios versus Árvore de Cortes Alternados

Cada nó existente numa *Half-Space Tree* pode ser um nó externo ou um nó com referência(s) para outro(s) nós pertencentes à estrutura de árvore.

Caso seja um nó com referência(s) para outro(s) nó(s) (neste caso um nó à direita e/ou um nó à esquerda) deve ser denominado de nó pai. Por sua vez, os nós que estejam referenciados à esquerda ou à direita do nó pai são denominados de nó(s) filho(s).

Cada nó pai tem uma referência para o atributo e o ponto de corte que origina o(s) nó(s) filho(s). É através deste atributo e ponto de corte que são identificadas as massas dos nós externos. A identificação dos nós de corte pode ser alternada ou aleatória.

Cada tipo de árvore possui implicações nos cortes que são atribuídos ao espaço de dados em análise.

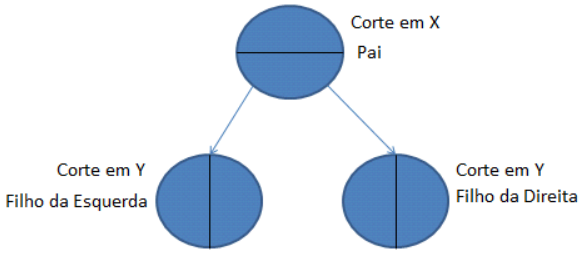

Figura 21 - Half-Space Tree com cortes alternados

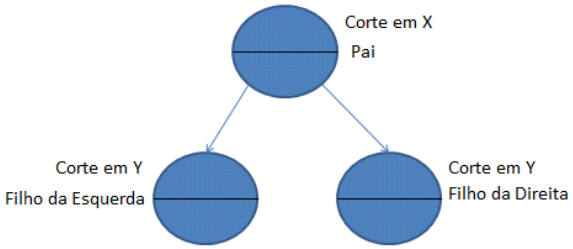

Figura 22 - Half-Space Tree com cortes aleatórios

Mediante o tipo de corte da árvore, o particionamento do espaço de dados sofre alterações. Quando o particionamento é alternado, cada corte dos filhos é o inverso do corte no pai, sendo criada uma *grid* com uma estrutura semelhante a uma *matriz* que poderá ter agregações a partir de determinado nível da árvore.

Quando os cortes são aleatórios, a estrutura de particionamento possui características irregulares. Para permitir uma melhor compreensão destes conceitos será explicado de seguida o corte aleatório e o corte sequencial tendo em conta dados bidimensionais.

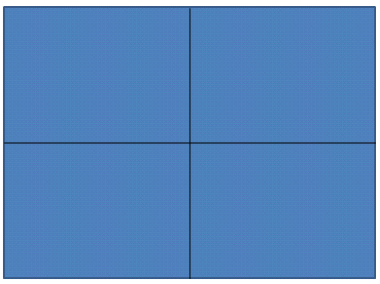

Figura 23 - Particionamento do Espaço de Dados em Células com Cortes Alternados

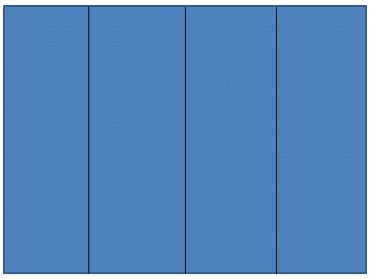

Figura 24 –Particionamento do Espaço de Dados em Células com Cortes Aleatórios

Assim, a Figura 21 e a Figura 22 exemplificam a estrutura de uma árvore de 1 nível em que são executados cortes alternados e cortes aleatórios. Já a Figura 23 e a Figura 24 correspondem ao particionamento do espaço de dados consoante os cortes das árvores definidos anteriormente.

3.7.4. Half-Space Trees e Deteção de Anomalias em Fluxos Contínuos de Dados

O trabalho desenvolvido até ao momento na área de processamento de fluxos contínuos de dados (*datastream mining*) é relativamente complexo. Este problema tem algumas características importantes quando em comparação com processamento global de dados.

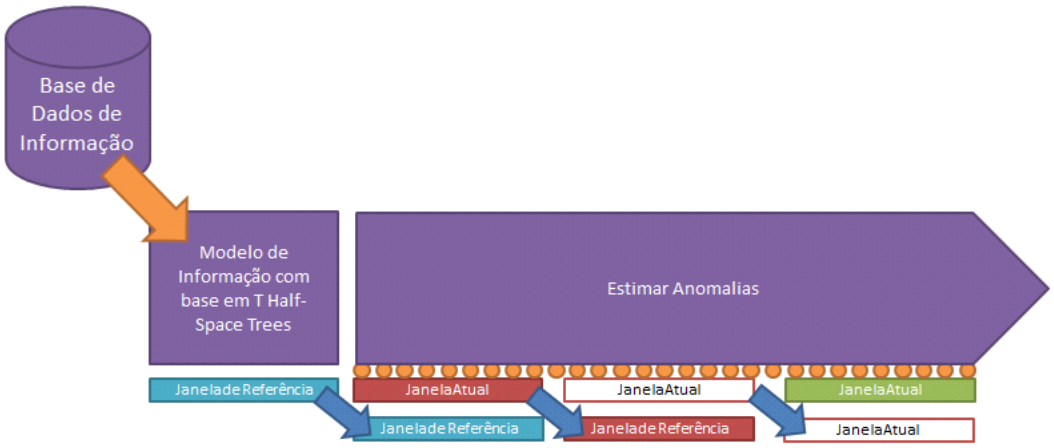

Figura 25 - Processamento de Fluxos Contínuos de Dados

Assim, em deteção de anomalias em fluxos contínuos de dados considera-se importante referir que normalmente estes modelos são considerados importantes já que existem (no âmbito de processamento de dados):

- restrições de memória, uma vez que não é possível alocar todos os registos de um determinado conjunto de dados em memória;
- os fluxos de dados podem ser altamente desbalanceados, e o conjunto de treino poderá não conter registos anómalos;
- os fluxos de dados evoluem ao longo do tempo, pelo que poderá ser necessário que o modelo de informação criado por um algoritmo de identificação de anomalias consiga se adaptar à aprendizagem de novos conceitos e assim manter altos níveis de avaliação.

Para processamento de fluxos contínuos de dados torna-se importante distinguir entre:

- tamanho (da janela) – parâmetro definido pelo utilizador e que define o número de registos processados antes de se atualizar o modelo de informação corrente;

- janela de referência – corresponde ao modelo de informação mantido até um determinado momento e que permitirá pontuar o nível de anomalia de um conjunto de registos a serem processados;
- janela atual – corresponde à chegada de um novo fluxo de registos de teste com um determinado tamanho - correspondente ao "*tamanho (da janela)"*. É importante referir que os contadores referentes à janela atual são colocados a zero ao se atualizar a janela de referência. Poderá ser interessante estudar-se o impacto de se manter os registos do histórico de dados já processado.

No caso de processamento de fluxos de dados as árvores são construídas sem quaisquer dados iniciais. Neste modelo também pode ser definido um conjunto de treino e um conjunto de teste.

Após a criação das árvores do modelo de informação são processados os dados de treino, processando-se de seguida o conjunto de teste. Para cada registo de teste é obtida a pontuação para o seu grau de anomalia.

Construção de Half-Space Trees para Processamento de Fluxos de Dados

```
Entradas: min & max – arrays com o valor mínimo e máximo para cada dimensão do WorkSpace
          k – nível de profundidade corrente
Saídas: uma HS-Tree
ConstruirArvore(min, max, k){
   Se k == maxProfundidade então
      Retornar Nó (r = 0; l = 0)
    Senão
      Aleatoriamente selecionar dimensão q
      p = (maxq + minq) / 2 {
          Construir dois nós (Esquerdo e Direito) com base num corte em dois espaços idênticos
      }
      temp = (maxq); maxq = p; Esquerdo = ConstruirArvore(min, max, k+1);
      maxq = temp; minq = p; Direito = ConstruirArvore(min, max, k+1);
      Retornar Nó(Esquerdo, Direito, AtributoCorte = q, PontoCorte = p, r = 0, l = 0)
}
```

Figura 26 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados

A construção de uma *Half-Space Tree* pressupõe a criação de dois registos nos nós da árvore para contabilizar as massas correspondentes à janela atual e à janela de

referência. É com base nestes novos campos que vão ser obtidas as pontuações de anomalia para cada registo.

Streaming das Half-Space Trees

```
Entradas: w – tamanho janela; t – número de árvores
Saídas: s – scoring de anomalia para cada registo pertencente ao conjunto de teste
FluxoArvores(w,t){
    Construir t Half-SpaceTrees
    Construir Modelo de Informação com base no Conjunto de Treino
    Count = 0;
    Enquanto continuar a receber dados{
        x = próximo registo;
        x.s = ;
        para cada árvore T no conjunto de Half-Space Trees{
                s = s + Score(x,T) {importante acumular os diferentes scores}
                AtualizarMassa(x,T.raiz, false);
        } Count ++;
        Se Count == w{
                AtualizarModelo: No.R = No.L;
                                 No.L = 0;
                                 Count = 0;
        }
    }
}
```

Figura 27 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados

```
Entradas: x – registo a ser processado; nó – um determinado nó da árvore; janelaReferencia – variável boleana a identificar se deve ser atualizada a janela de referência ou a janela atual
Saídas: nenhuma
AtualizarMassa(x, nó, janelaReferencia){
    janelaReferencia?  Nó.R ++ : nó.L ++;
    Se (nó.Profundidade < profundidadeMaxima) Então{
        Nó' é o próximo nó que x processa
        AtualizarMassa(x, Nó', janelaReferencia)}
}
```

Figura 28 – Pseudocódigo para Criar uma Half-Space Tree para processamento de Fluxos de Dados

De referir que para processamento de fluxos contínuos de dados apenas foi considerado o particionamento sequencial do espaço de dados, uma vez que os fluxos têm tendência a ser fornecidos por uma determinada ordem.

Atualização do Perfil de Massa com base nas janelas de Fluxos de Dados

De cada vez que é processado um registo existe uma atualização do perfil de massa de cada nó das Half-Space Trees que constituem o modelo de dados (Figura 28).

Com base no código para a construção do processamento de fluxos de dados (Figura 27) é possível verificar que:

- o processamento dos registos é iterativo;
- existe a atualização do modelo de informação com base nos valores da última janela atual (janela deslizante);
- as necessidades de memória são inferiores, especialmente quando comparado com o processamento *offline* de dados, uma vez que um volume de dados tem de ser mantido em memória.

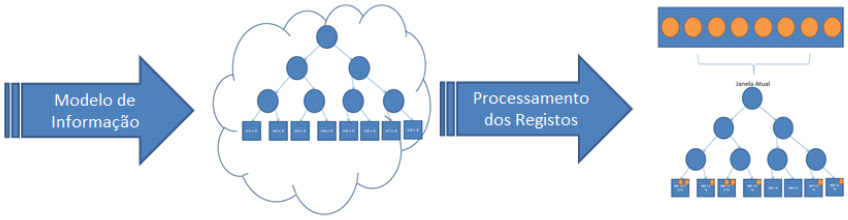

Figura 29 – Processamento do fluxo de dados

O processamento do fluxo de dados teve em conta duas fases distintas: a construção do modelo de informação e o processamento dos registos tendo em conta a árvore já criada (Figura 29).

Capítulo 4

Análise Exploratória dos Dados

Um sistema SCADA fornece um conjunto de registos associados a leituras de diferentes componentes. Estes indicadores podem fornecer indicação do estado de operação dos componentes eletrónicos em diferentes fases associadas ao processo de conversão de energia cinética em energia elétrica. Assim, entre outras tarefas, é possível otimizar atividades de manutenção destes sistemas.

4.1. Caracterização das Variáveis em Estudo

Para efeitos de aplicação de algoritmos de Identificação de Anomalias foram disponibilizados registos referentes a um parque eólico, doravante denominado de "***Parque X"***, composto por 20 aerogeradores distintos, no período compreendido entre janeiro e dezembro de 2010.

	Variável	Definição
1.	Lv.ActivePower	Potência Real do Sistema (kW)
2.	Wind.Speed	Velocidade do Vento (m/s)
3.	Yaw.NacellePosition	Posição da Nacelle (º)
4.	Rot.RpmMonitor	Velocidade de Rotação (r.p.m)

Tabela 4 - Variáveis em Análise nos Datasets de Dados do Parque X

As variáveis escolhidas para análise foram as identificadas como importantes para monitorização de aerogeradores (Kusiak, 2010; Kusiak e Verma, 2013) (ver Tabela 4).

De seguida, será elaborada uma análise suscinta das variáveis escolhidas para os diferentes tempos obtidos.

Para facilitar a análise comportamental das diferentes séries temporais, é importante proceder-se à sua representação gráfica e ao cálculo de algumas medidas de localização e dispersão.

As tabelas com as estatísticas descritivas básicas encontram-se no anexo 2.

4.2. Análise da Potência Gerada

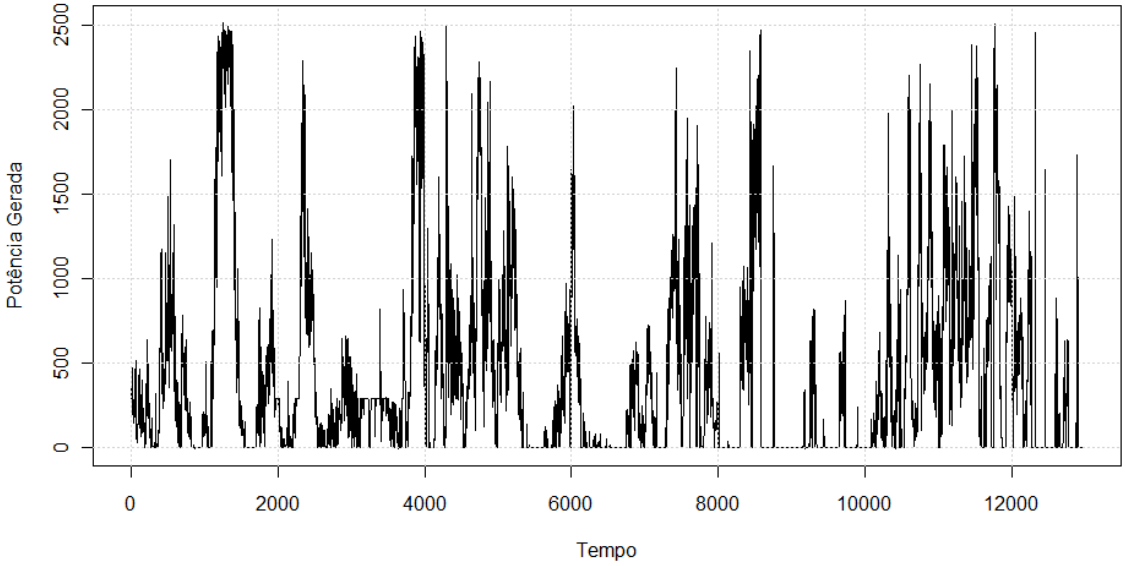

Figura 30 - Análise da Potência Gerada pelo Aerogerador 1 (janeiro, fevereiro, março 2010 - Parque X)

A análise de uma variável correspondendo a uma série temporal baseia-se então na representação do seu cronograma como relação entre o valor registado pelo sensor e o índice temporal em que foi obtido.

A análise do cronograma referente à Potência Gerada sugere:

- grande variabilidade ao longo do tempo, não existindo uma tendência óbvia;
- existência de períodos de potência controlada, nomeadamente perto dos 3000 minutos de operação;
- alturas de não disponibilidade, caracterizados por valores de potência próximos de 0 kW;

- grande dispersão dos dados (diferença entre a média (414.300 kW) e a mediana (198.600 kW) é grande – na ordem dos 215.7 kW)

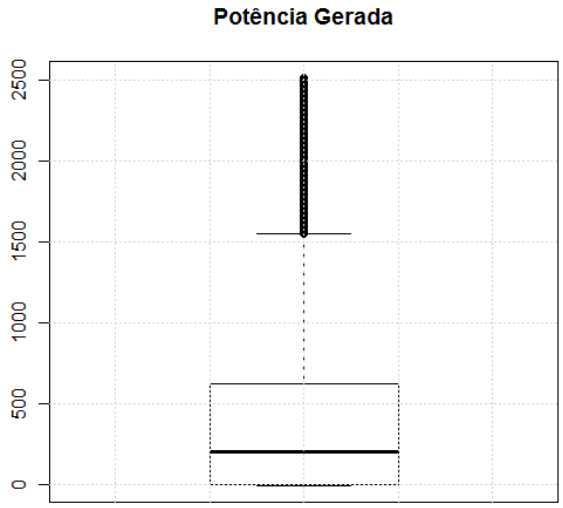

Figura 31 – Box-Plot da variável da Potência Gerada

O diagrama de Box-Plot (Figura 31) identifica a existência de anomalias univariadas (786 registos). A amplitude destes registos varia entre 1552 kW e 2516 kW.

4.3. Análise da Velocidade do Vento

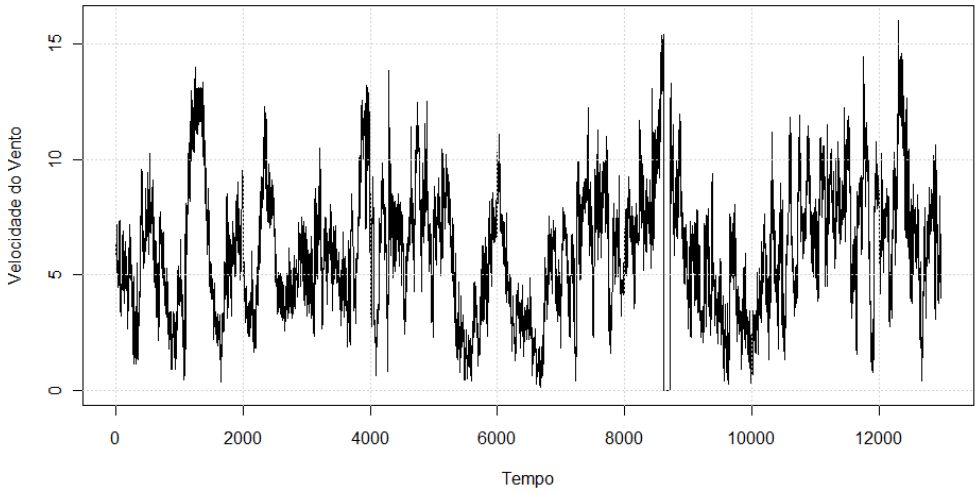

Figura 32 - Série Temporal da Velocidade do Vento do Aerogerador 1 (janeiro, fevereiro e março de 2010 – Parque X)

Com base no cronograma da série temporal associada à velocidade do vento é possível identificar como principais características a existência de:

- tendência linear;

- apresentação de alguns ciclos;
- média (5.962 m/s) e mediana (5.919 m/s) com valores muito semelhantes;
- existência de *outliers* univariados;
- grande variabilidade;
- tendência em torno de 5 m/s;
- mínimo obtido em 0 m/s e máximo em 16.02 m/s.

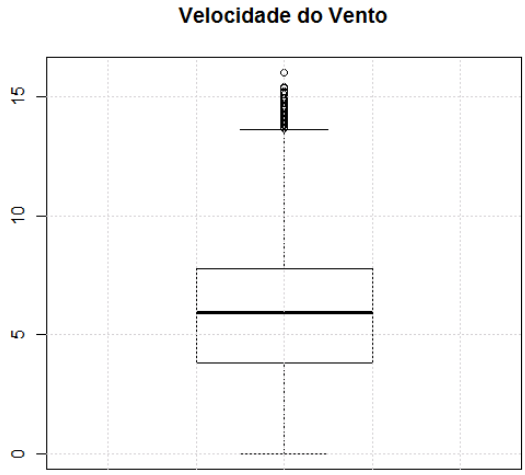

Figura 33 - Diagrama de Box-Plot da variável de Velocidade do Vento

Com base no diagrama de Box-Plot (ver Figura 33) é possível identificar a existência de registos anómalos univariados (56 registos).

A amplitude dos registos *outliers* varia entre 13.69 m/s e 16.02 m/s.

4.4. Análise da Posição da Nacelle

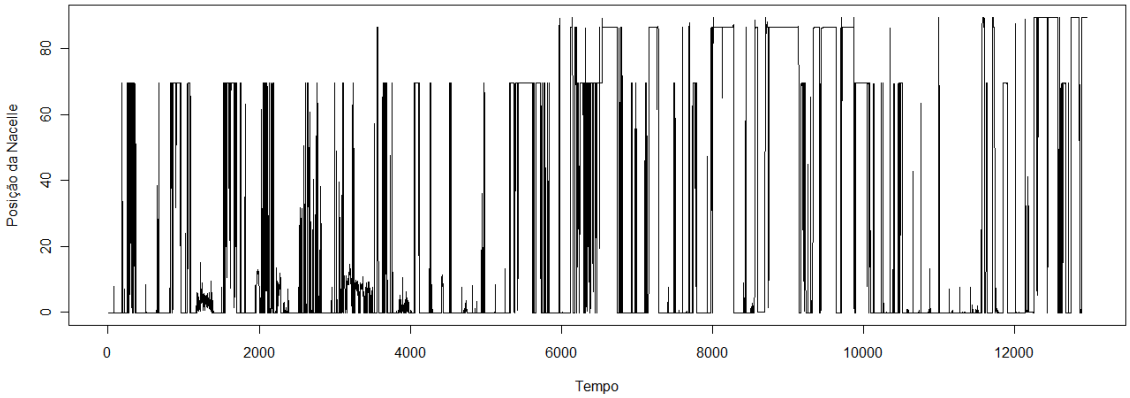

Figura 34 – Série Temporal da Posição da Nacelle do Aerogerador 1 (janeiro, fevereiro e março de 2010 – Parque X)

Com base na Figura 34 é possível identificar que os dados apresentam:

- grande variabilidade;
- períodos de tendência linear.

Com o diagrama de Box-Plot é possível identificar a não existência de outliers (ver Figura 35).

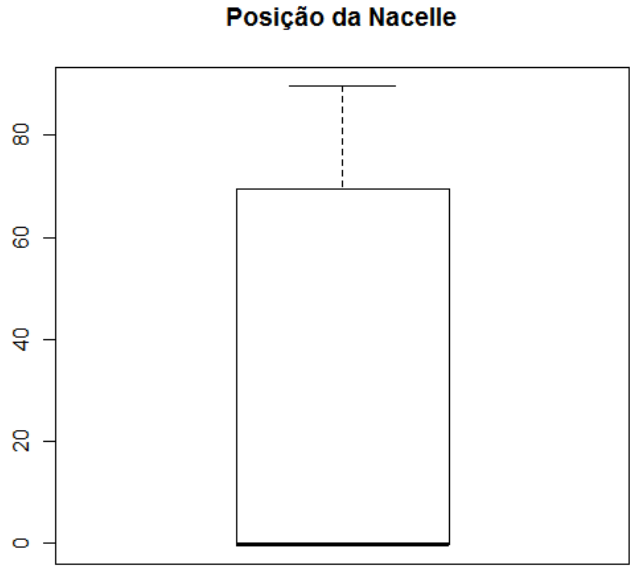

Figura 35 –Box-Plot da variável Posição da Nacelle

4.5. Análise das Rotações por Minuto

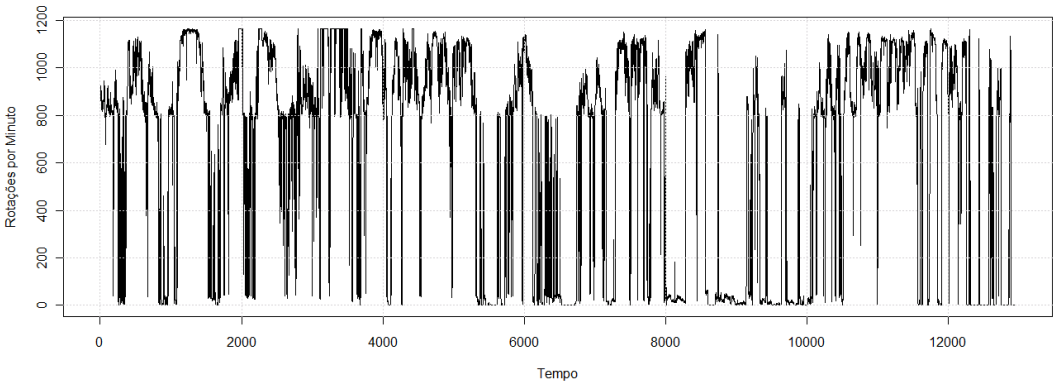

Figura 36 - Série Temporal da Rotação por minuto do Aerogerador 1 (janeiro, fevereiro e março de 2010 - Parque X)

As Rotações por Minuto do Aerogerador 1, identificadas com base no cronograma da Figura 36 registam uma grande variabilidade dos dados obtidos, para além de uma tendência em torno das 600 rotações por minuto.

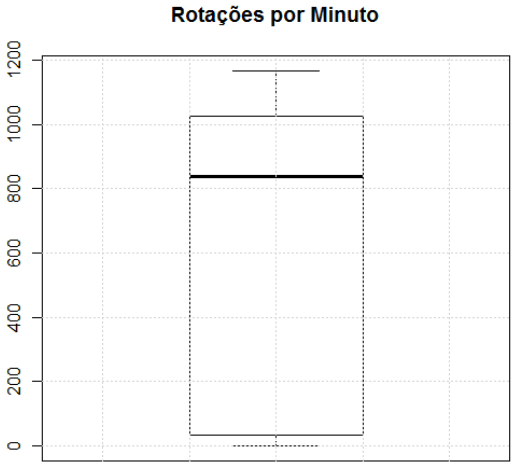

Figura 37 – Box-Plot da variável Rotações por Minuto

Com base no diagrama de Box-Plot é possível verificar a não existência de *outliers* (Figura 37).

4.6. Outliers Identificados com base nos Resíduos das Séries Temporais

Apesar de não estar relacionado com estimação de massa considera-se importante referir a possibilidade de estimar as *outliers* com base nos resíduos das séries temporais de cada uma das principais variáveis em análise.

Para tal é necessário:

- identificar as componentes de tendência e sazonalidade dos dados;
- subtraí-las à série inicial e identificar os resíduos.

O teste para identificação de *outliers* com base nos resíduos de séries temporais é idêntico ao teste para identificação de *outliers* nos diagramas de boxplot de dados univariados. Ou seja, assume-se que são anomalias os pontos acima ou abaixo dos intervalos de 1.5IQR (Amplitude Inter Quartil).

As séries temporais alvo de análise neste ponto serão as variáveis de Potência Gerada e Velocidade do Vento. O método utilizado e implementado na linguagem de programação R encontra-se no anexo 1. Para análise destas séries temporais foi considerado o modelo de *Loess*.

Nos dados em análise, é possível verificar a existência de 862 pontos identificados como anómalos para a série temporal da Potência Gerada, e 45 registos para a série temporal associada à Velocidade do Vento.

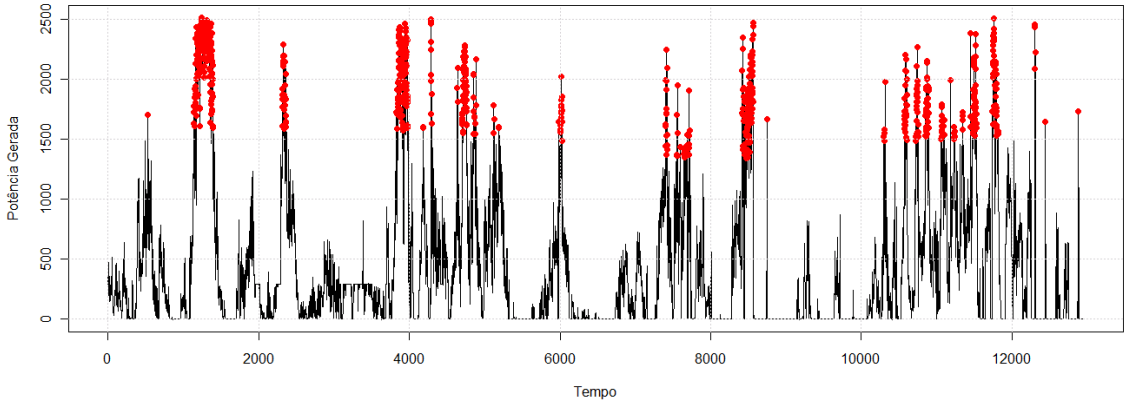

Figura 38 - Identificação de Anomalias na Série Temporal da Potência Gerada

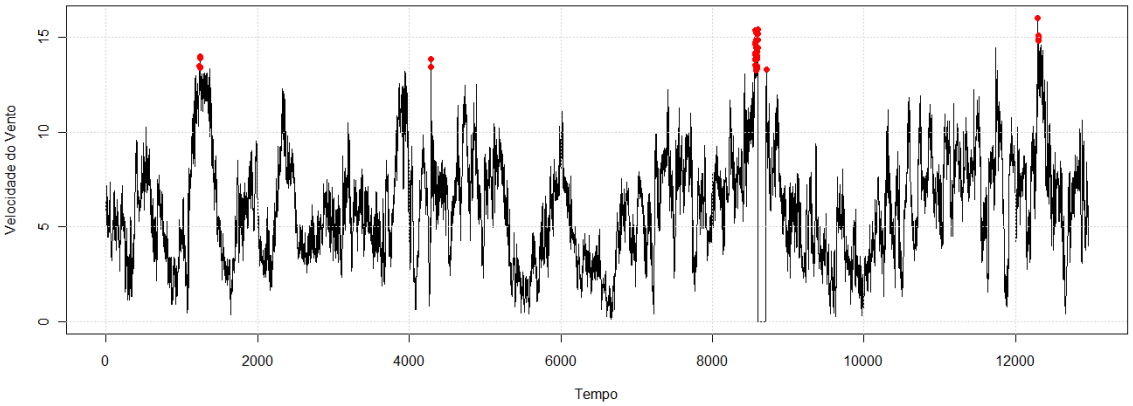

Figura 39 - Identificação de Outliers na Série Temporais da Velocidade do Vento

As pontuações obtidas estão registadas no anexo 3.

Capítulo 5

Estimação de Massa

Na presente dissertação foi implementada a execução de um estimador de massa, tendo em conta os atributos inseridos no sistema pelo utilizador. Para tal, foi desenvolvida uma aplicação informática com recurso ao ambiente de desenvolvimento DOTNET (.Net) *Windows Forms* na linguagem de programação CSHARP (C#).

O estimador de massa criado foi desenvolvido com base na criação das estruturas de dados *Half-Space Trees* nas duas dimensões: cortes alternados e cortes aleatórios (ver Capítulo 3.7 – Estimação de Massa).

A aplicação foi desenvolvida para permitir aplicar os conceitos de Estimação de Massa a outros *datasets*. De seguida será explicada a execução da leitura de ficheiros de dados e partição dos ficheiros em conjunto de treino e de testes.

5.1. Especificação de Fases Comuns de Execução

O processo inicia-se com a repartição dos dados num conjunto de treino e num conjunto de teste. Estes conjuntos podem ser aleatórios ou sequenciais, e são escolhidos com base em dimensões definidas pelo utilizador.

De seguida, são definidas as fases de ***Treino*** e de ***Estimação de Outliers*** que fazem parte da aplicação. Estas fases fazem parte dos componentes transversais aos diferentes algoritmos implementados.

5.1.1. Fase de Treino

Na fase de treino (Figura 41), também denominada de fase de criação do modelo de dados, são selecionados os seguintes parâmetros:

- Tamanho do Conjunto de Treino;
- Número de Florestas e número de Árvores em cada Floresta;
- Variável para o Eixo dos X;
- Variável para o Eixo dos Y;
- Profundidade Máxima de cada Árvore;
- Tamanho Máximo de registos por nó.

Estes parâmetros permitem que os algoritmos sejam aplicados a diferentes tipos e volumes de dados.

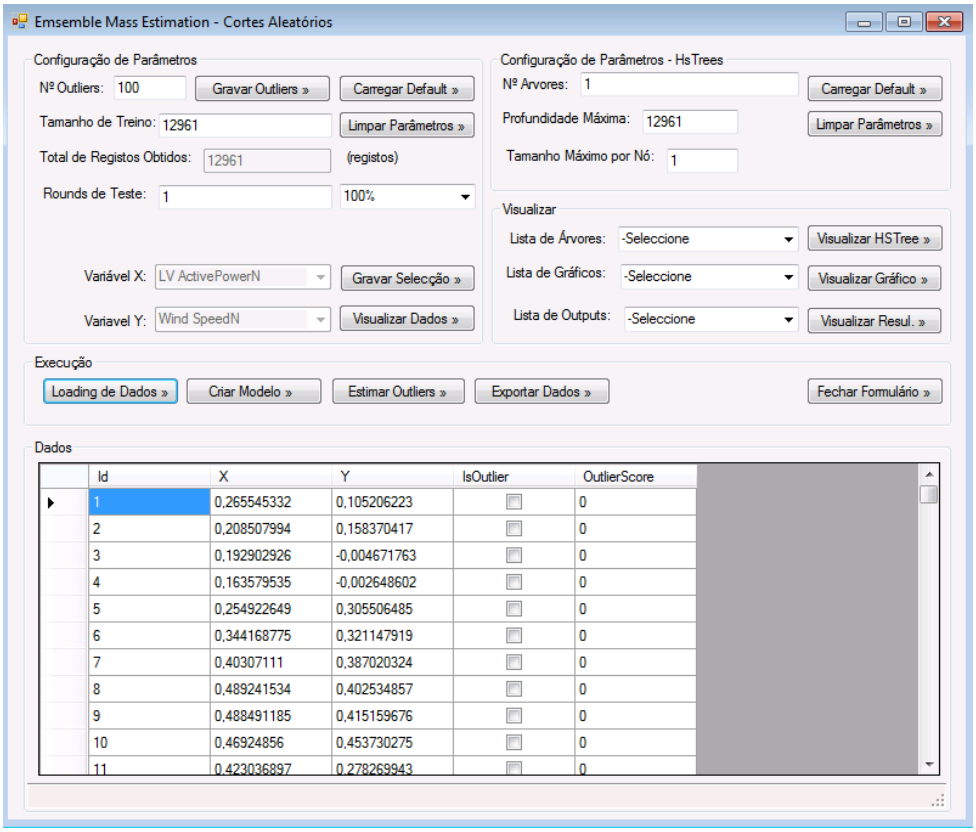

Figura 40 - Estimação de Outliers com recurso a Half-Space Trees e Criação de diferentes Conjuntos de Dados

Atualmente, o carregamento dos dados é efetuado com base em ficheiros *csv* (*comman separated values*).

Após se carregar os dados é possível visualizar os atributos definidos como principais no Capítulo 4. Após a seleção dos parâmetros de entrada é criado o conjunto de dados a ser analisado.

Antes de se executar o algoritmo selecionado é possível visualizar as séries temporais referentes aos campos X e Y (Figura 42) que permitirão criar a Curva de Energia correspondente.

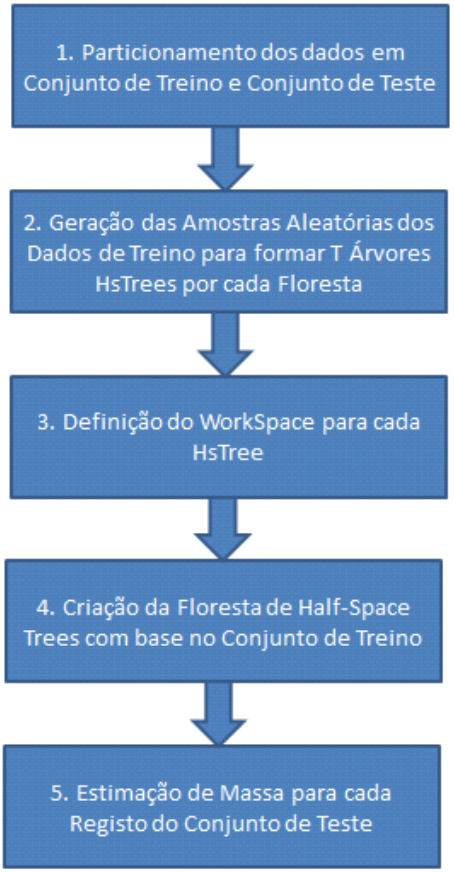

Figura 41 - Identificação dos principais passos de Estimação de Massa com recurso a Half-Space Trees

A execução do algoritmo pressupõe a criação de um modelo de informação baseado no conjunto de treino. Com o modelo de informação criado é possível visualizar-se cada árvore em separado (Figura 43).

Após a verificação da correta geração do modelo de treino, com base na hierarquia de árvores especificada pelo utilizador (tendo em conta cortes aleatórios ou sequenciais), é possível efetuar a estimação de anomalias.

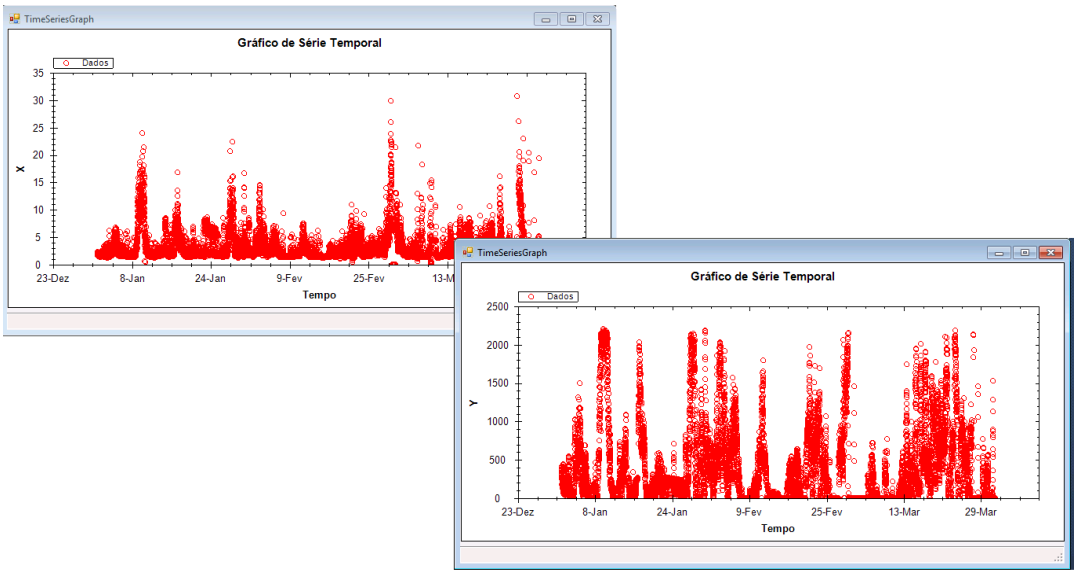

Figura 42 - Séries Temporais associadas às variáveis escolhidas para análise

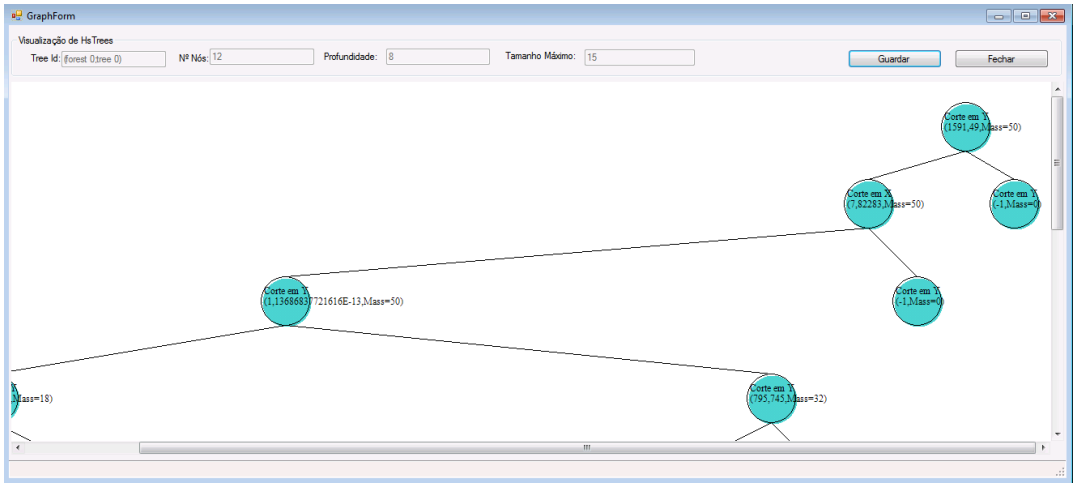

Figura 43 - Identificação da Estrutura de uma Half-Space Tree gerada pelo sistema

5.1.2. Fase de Estimação de Outliers

Na fase de *Estimação de Outliers*, as diferentes florestas e respetivas árvores, são sequencialmente percorridas para todos os registos que pertençam ao conjunto de teste.

Posteriormente é calculado o grau de anomalia de cada registo tendo em conta:

- se fazem parte do *Espaço de Trabalho* da Árvore T;
- se a árvore T foi criada com base em: limite de profundidade, limite de massa por nó externo ou em ambos;
- se a árvore T foi criada para ter idênticos espaços de agrupamento criados.

De referir que a função de anomalia já definida anteriormente é dada por:

$$s_a(x) = m(T_a^h(x|D) * 2^{la}$$

$$m(x) = \frac{1}{T}\sum_{a=1}^{T} s_a(x)$$

Figura 44 – Expressões que permitem calcular a pontuação para cada registo que faça parte do conjunto de teste

5.2. Resultados Alcançados

Os testes foram realizados após uma normalização dos dados e tendo em consideração os dados do "***Parque X"*** (Aerogerador 1) para o período compreendido entre janeiro e março de 2010.

Os testes para os algoritmos com base em Estimação de Massa pressupõem:

- 1 floresta;
- 1 árvore por cada floresta;
- não existência de limite de profundidade;
- 1 nó externo por cada registo de treino;
- conjunto de Treino e de Teste idênticos com 100% dos registos selecionados.

Por sua vez a execução do *LOF* para o mesmo horizonte temporal pressupõe uma pesquisa segundo os 100 vizinhos mais próximos.

Já o processamento segundo *Fluxos Contínuos de Dados* pressupõe:

- a criação de janelas de referência e janelas atuais com uma dimensão de 250 registos;
- altura máxima de árvore de 15 nós;
- existência de limite de pontos por nó externo sendo este limite definido como 25.

5.2.1. Árvores de Cortes Aleatórios

É possível visualizar nas Figuras 46 a 49 o resultado de execução do *ranking* de dados considerando um limite de significância na ordem dos 0.15%, 0.5%, 1% e 5% dos registos analisados. Assim, identifica-se que o algoritmo de modelação de dados com

base em Estimação de Massa consegue captar na perfeição a dinâmica associada à Curva de Potência.

Os gráficos identificados na Figura 46 e na Figura 47 identificam os *outliers* mais severos. Estas anomalias, identificadas manualmente na Figura 45 como "***Região A"***, localizam-se na zona mais afastada de toda a dinâmica da curva de potência. Nos gráficos subsequentes começam a ser identificadas as zonas com outros pontos mais afastados e caracterizados como pertencentes à "***Região B"*** e "***Região C***" da Figura 45.

As regiões identificadas como A, B e C caracterizam:

- Situações de Potência Controlada;
- Situações de Registos Dispersos em Função da Curva de Potência Gerada;
- Situações de Não Funcionamento do Aerogerador.

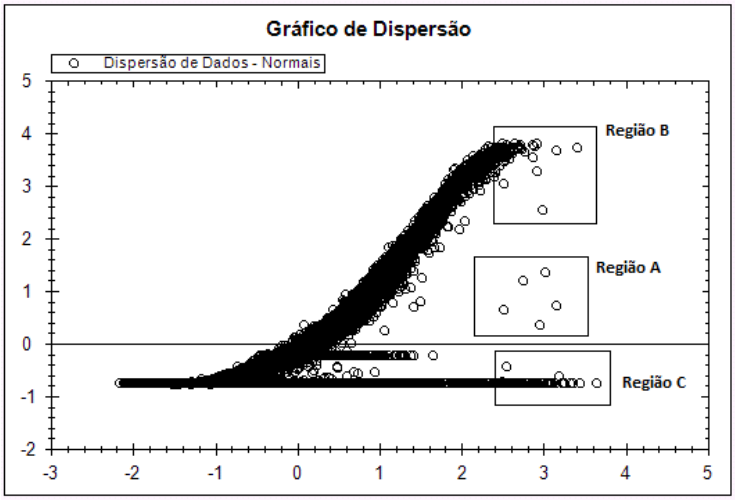

Figura 45 - Dispersão dos Dados Normalizados com identificação manual de 3 regiões de dados mais dispersos em relação à Curva de Potência original

Com base nos valores obtidos com a função de *ranking* dos dados é possível identificar a sua distribuição aproximadamente normal. Após a exportação dos resultados obtidos e a sua importação para o SPSS foi possível obter a sua distribuição de frequências (Figura 50). Com os resultados obtidos pelo diagrama de frequências e pela tabela de distribuição de frequências (Anexo 4) é possível categorizar os registos mediante diferentes "grupos" – estratificados com base na pontuação de anomalia originada pelo algoritmo.

O diagrama de distribuição de frequências permite que nos apercebamos que vários registos foram agrupados em torno de um mesmo *ranking*.

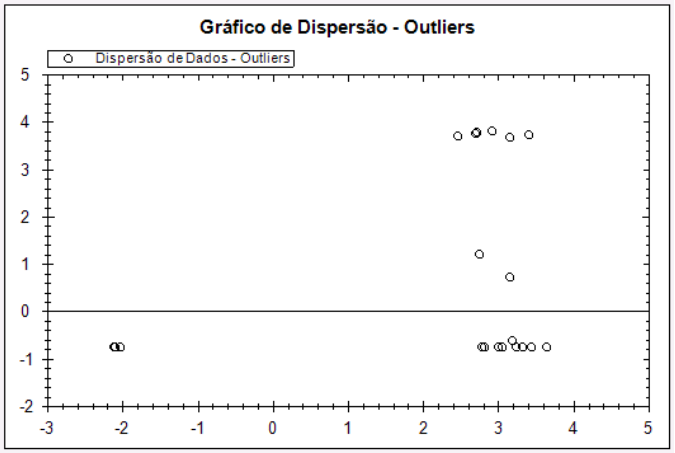

Figura 46 – Ranking com Árvores de Cortes Aleatórios com 20 registos

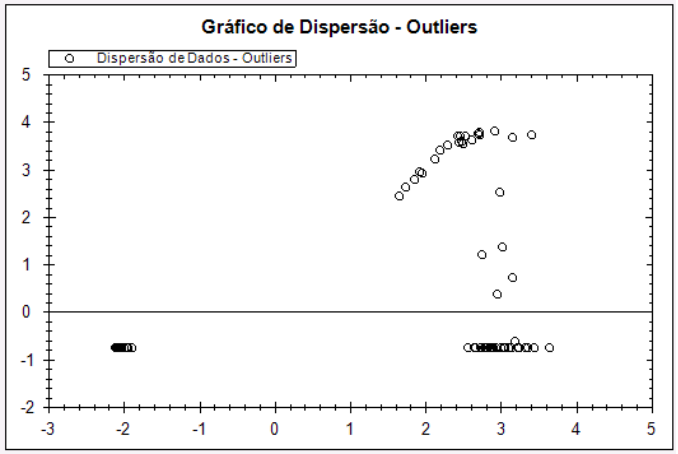

Figura 47 - Ranking com Árvores de Cortes Aleatórios com 65 registos

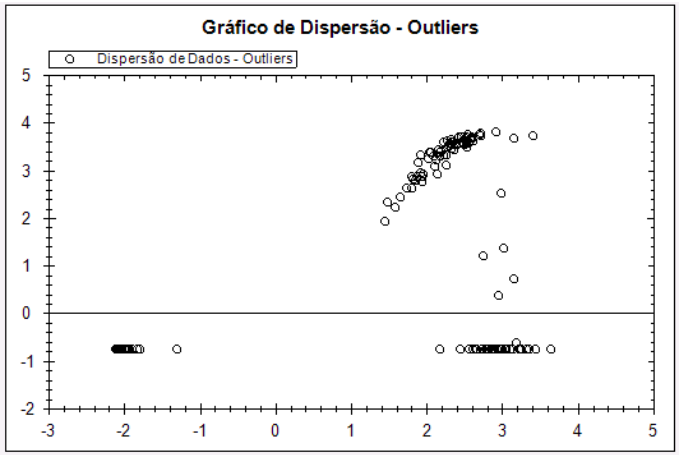

Figura 48 - Ranking com Árvores de Cortes Aleatórios com 130 registos

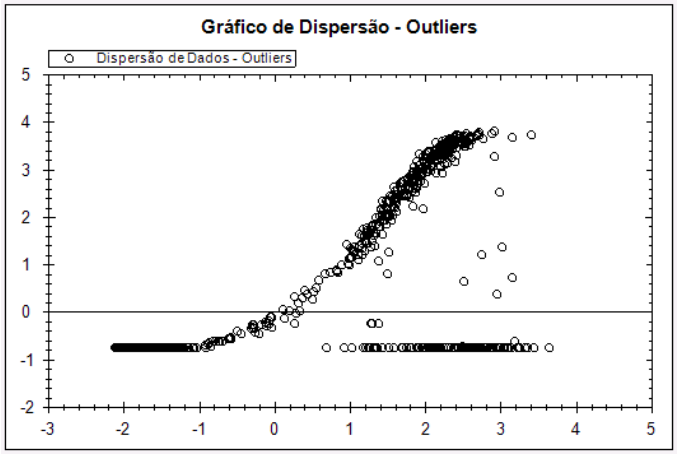

Figura 49 - Ranking com Árvores de Cortes Aleatórios com 648 registos

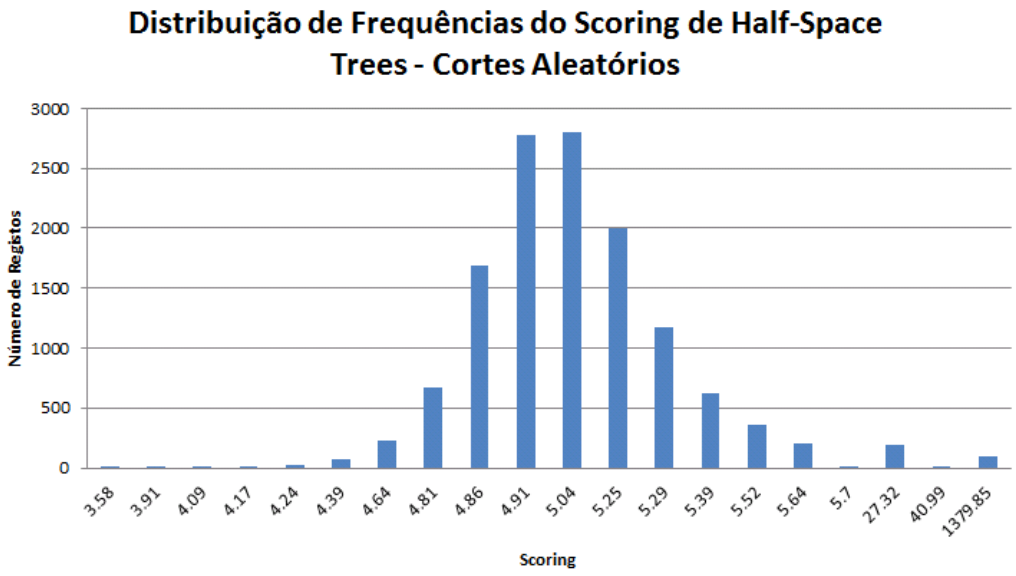

Figura 50- Distribuição de Frequências do resultado de Scoring

Genericamente é possível identificar que a função de pontuação atribui aos registos um nível com base na distância do nó externo até ao topo da árvore (nó raiz) e no número de registos (massa) no nó externo.

A distribuição de frequências acumulada (Figura 51) identifica como *threshold* o grupo de *outliers* de 4.81.

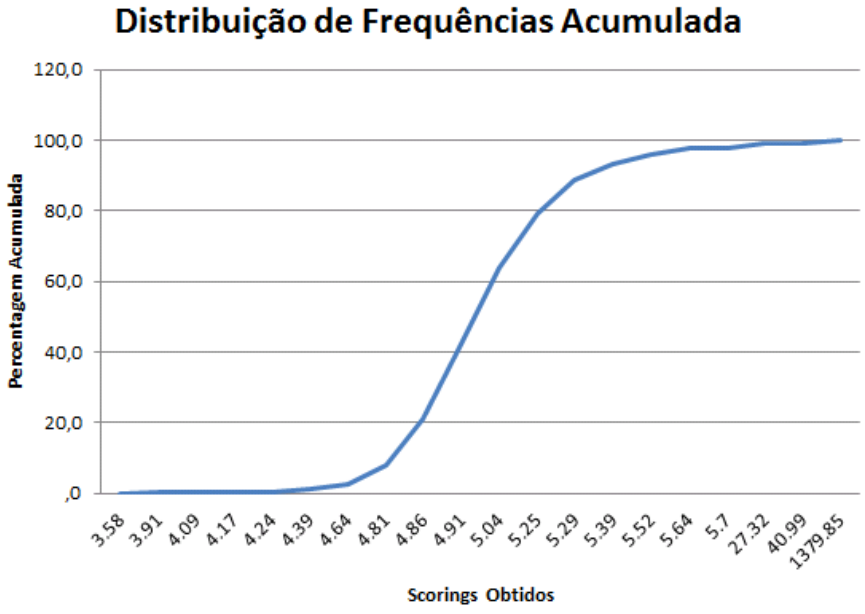

Figura 51 - Distribuição de Frequências Acumulada dos grupos de Scoring obtidos com base nas Half-Space Trees de Cortes Aleatórios

A partir deste momento, e selecionando os primeiros 1019 registos, é possível caracterizar a classe de anomalias e a classe de registos normais, tendo em conta as

diferentes variáveis à disposição do utilizador. Considera-se que esta análise não faz parte da presente Dissertação, uma vez que poderia ser um trabalho bastante extenso, mas importante com um estudo subjacente ao nível de significância considerado.

5.2.2. Árvores de Cortes Alternados

Uma comparação básica entre a Figura 50 e a Figura 52 permite verificar que uma árvore de cortes alternados consegue criar um *ranking* mais gradual e menos aglomerado, para um conjunto de parâmetros básicos. Dessa forma consegue distinguir melhor categorias ou níveis de anomalias.

A análise visual (ver Figura 53, 54, 55 e 56) com um limite de significância de 0.15%, 0.5%, 1% e 5% dos registos considerados anómalos, permite identificar as zonas caracterizadas na Figura 45 como região B e região C.

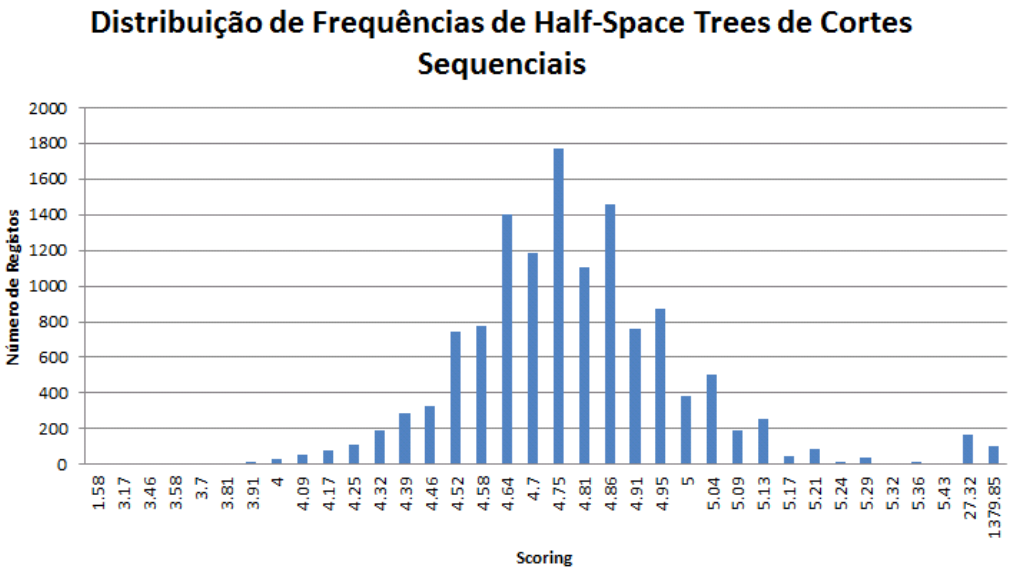

Figura 52 –Distribuição de Frequências do resultado de Scoring

O diagrama de distribuição de frequências das Half-Space Trees de Cortes Sequenciais (Figura 52) permite verificar que este modelo permite identificar mais grupos do que quando os cortes selecionados são aleatórios. Esta característica poderá estar relacionada com a forma como os cortes são executados, uma vez que cortes sequenciais obrigam a que as divisões sejam equitativas entre os diferentes atributos. De referir que as árvores de cortes aleatórios não obrigam sequer a que todos os atributos sejam tidos em consideração (existe a probabilidade de que num grupo de atributos selecionados pelo utilizador, nem todos os eixos sofram cortes).

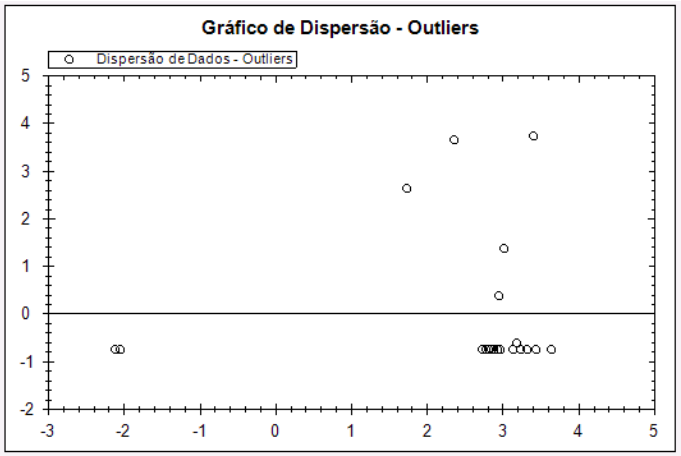

Figura 53 - Ranking para Árvores de Cortes Alternados (20 registos)

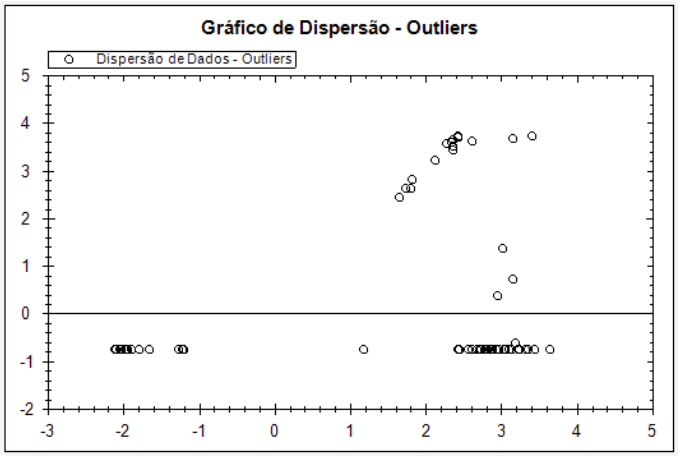

Figura 54 - Ranking para Árvores de Cortes Alternados (65 registos)

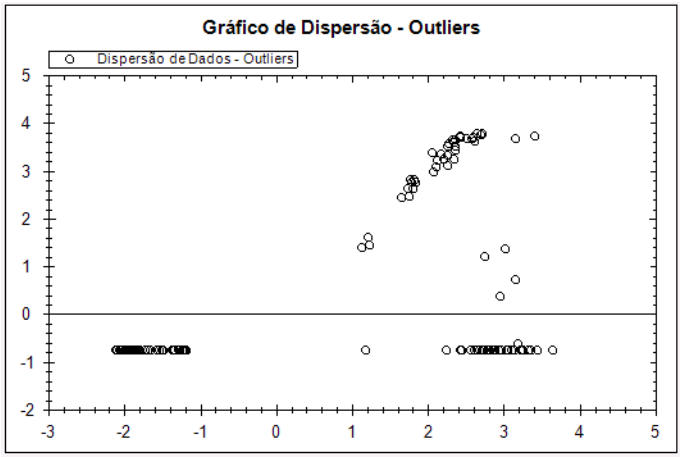

Figura 55 - Ranking para Árvores de Cortes Alternados (130 registos)

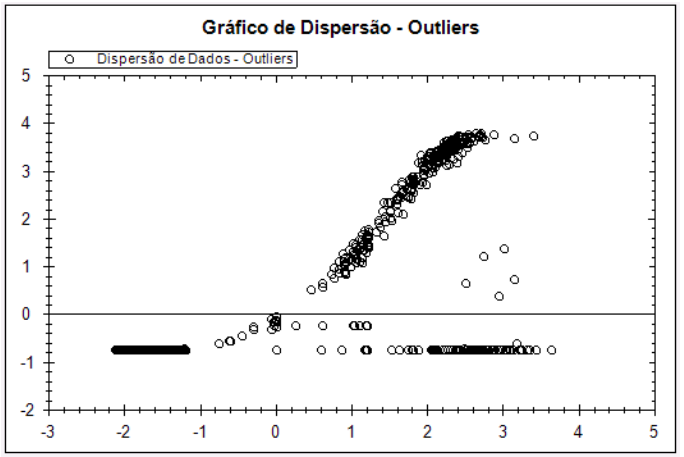

Figura 56 – Ranking Árvores de Cortes Alternados (648 registos)

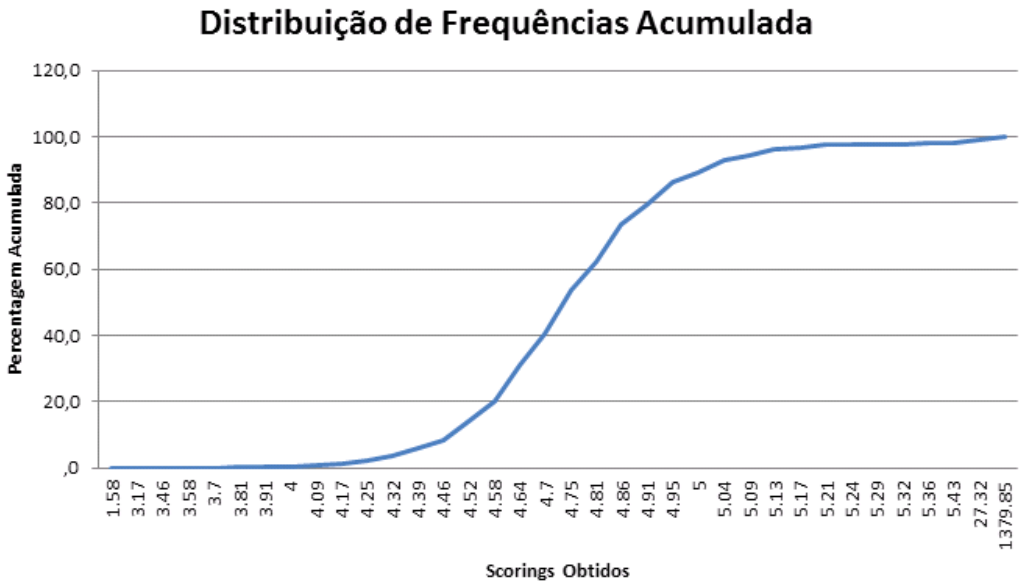

Figura 57 –Distribuição de Frequências Acumulada dos grupos de Scoring obtidos com base nas Half-Space Trees de Cortes Sequenciais

Por sua vez, e tendo em consideração a Figura 57 é possível identificar o ponto de corte no grupo 4.46. Assim, considera-se que o algoritmo de Cortes Sequenciais *Offline* sugere a existência de 1106 anomalias.

5.2.3. Local Outlier Factor

Além da implementação referente à estimação de massa, tanto *offline* como *online*, também foi desenvolvida a execução do *Local Outlier Factor* – um algoritmo "estado da arte" e pertencente à categoria dos métodos baseados em densidade (Figura 15). Como já foi dito anteriormente, a execução do LOF tem por base a seleção de 100 registos para o cálculo dos vizinhos mais próximos. Este é um algoritmo custoso uma vez que é necessário calcular a distância de todos os pontos, dois a dois.

O LOF pode-se tornar mais simples se se fornecer uma dimensão de Conjunto de Treino inferior à total dimensão de registos, uma vez que apenas é necessário calcular a distância dos registos de teste com todos os registos a serem analisados.

Com base no Diagrama de Distribuição de Frequências (Figura 58) e na Tabela de Distribuição de Frequências (anexo 6) é possível identificar a existência de mais grupos do que os definidos com base nos métodos de Estimação de Massa. Assim, o LOF fornece uma caracterização do grau de anomalia de um registo de forma mais gradual.

Por outro lado, com uma análise mais demorada à Figura 63 é possível verificar que para os parâmetros identificados, o LOF permite identificar algumas anomalias coletivas, existentes na zona limitrofe da curva de potência.

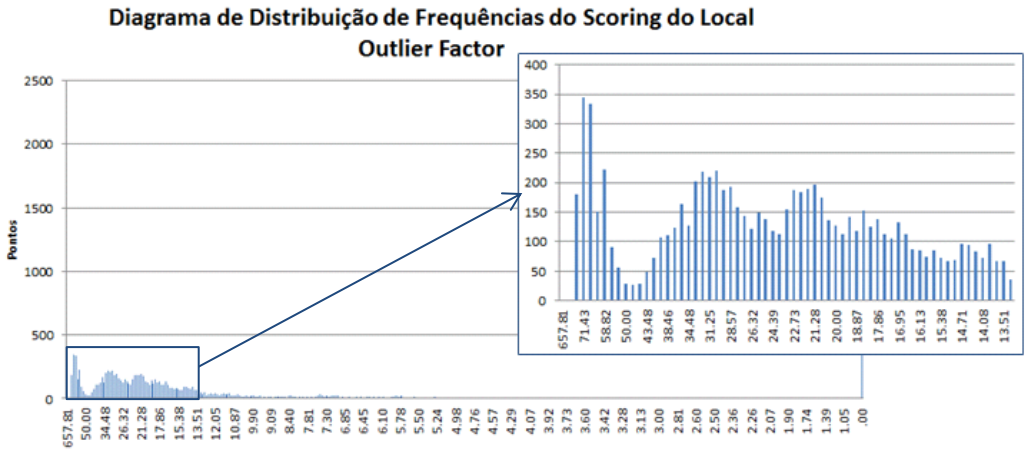

Figura 58 - Diagrama de Distribuição de Frequências do LOF

O diagrama de Distribuição de Frequências Acumulada (Figura 59) identifica o grupo 62.50 como o limite de significância para o número de anomalias detetadas. Assim, considera-se que o LOF com k = 100 vizinhos detetou 1012 registos como irregulares.

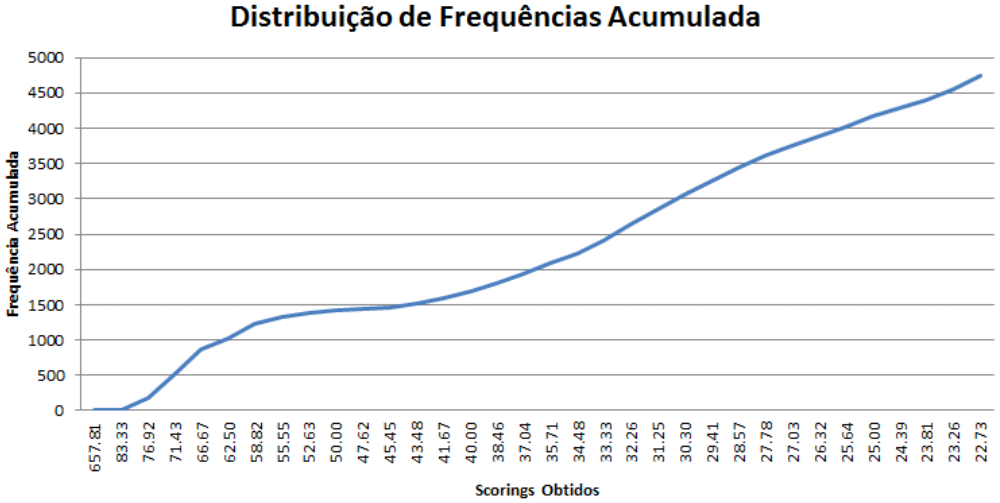

Figura 59 – Distribuição de Frequências Acumulada dos grupos de Scorings criados com base no Local Outlier Factor

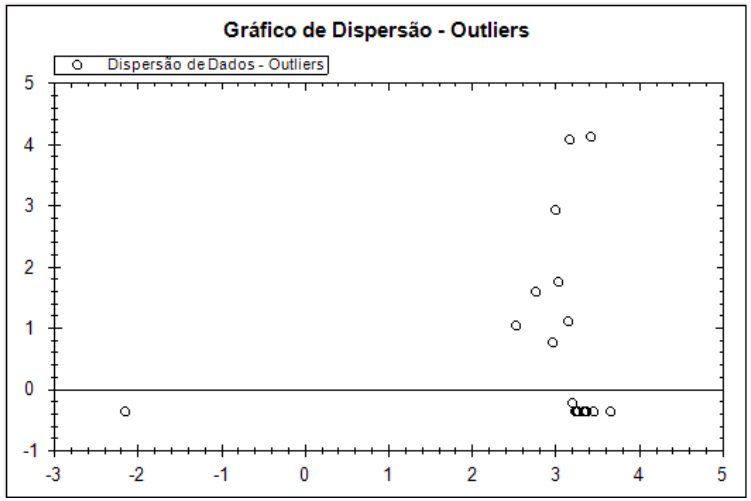

Figura 60 – Ranking com LOF (20 registos)

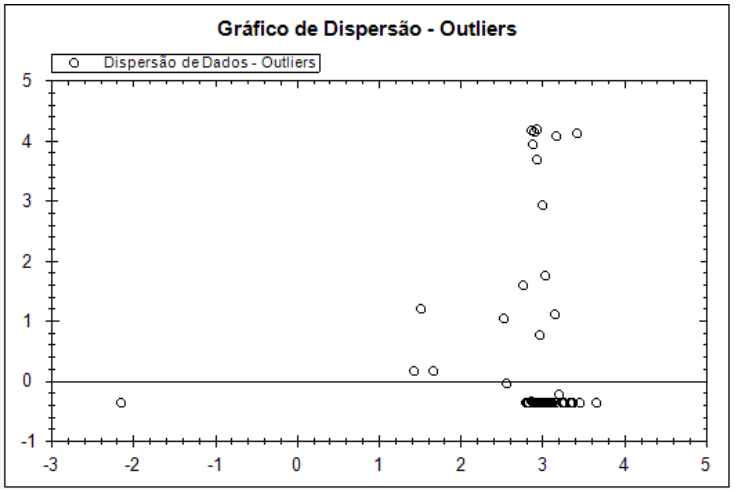

Figura 61 – Ranking com LOF (65 registos)

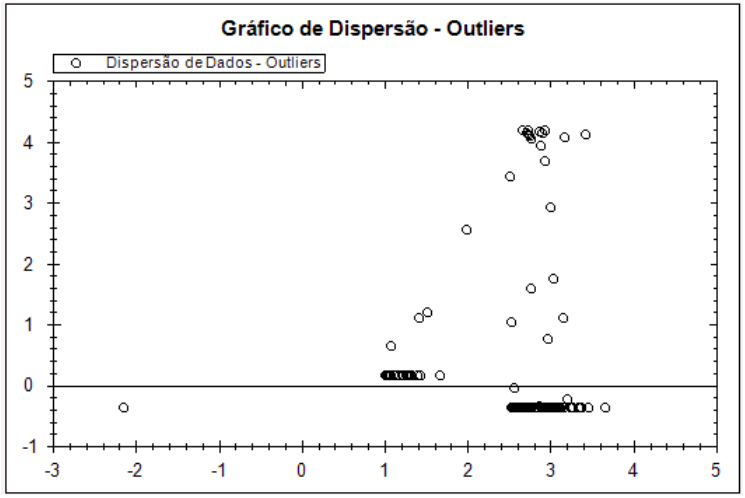

Figura 62 - Ranking com LOF (130 registos)

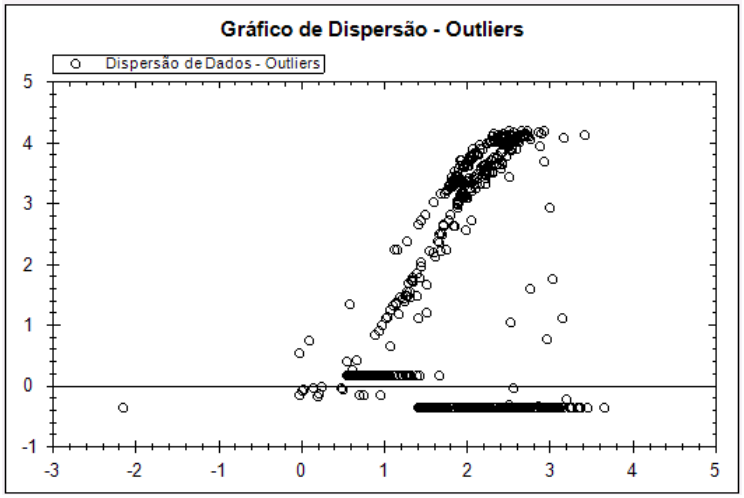

Figura 63 - Ranking com LOF (648 registos)

5.2.4. Deteção de Anomalias em *DataStreams*

Como foi referido em 3.7.4 – Half-Space Trees e Deteção de Anomalias em Fluxos Contínuos de Dados, existem já alguns estudos para a implementação do conceito de **Estimação de Massa** para a **Identificação de Anomalias em DataStreams**. Também neste caso foi implementada a execução de Estimação de Massa e registados os resultados de um teste autónomo. Os resultados encontram-se no anexo 7.

Os parâmetros escolhidos foram:

- criação de um Modelo de Informação sem registos;
- 1 floresta com 1 árvore de dados;
- 15 filhos como limite de profundidade da árvore (uma vez que o Modelo de Informação não contêm registos foi necessário definir um limite de profundidade);
- um tamanho de janela de 250 registos o que faz com que sejam processadas cerca de 28 janelas independentes;
- 25 registos como tamanho limite para os nós externos.

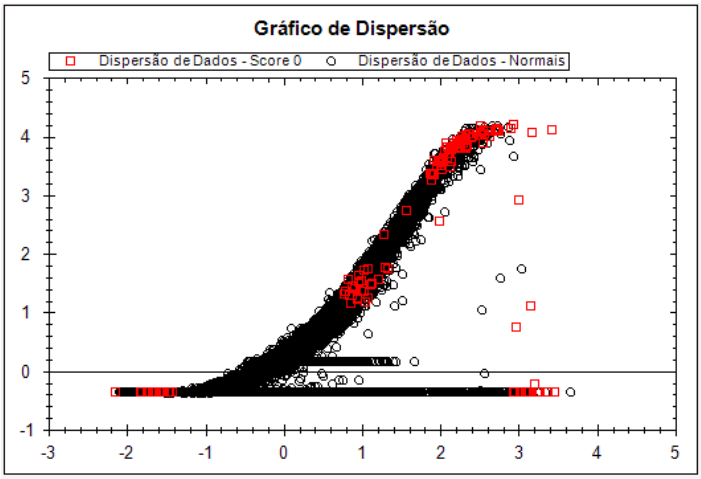

Figura 64 – Posição dos rankings no Modelo com DataStreams no 1ºgrupo

Os resultados de dispersão dos primeiros registos considerados anómalos, e identificados com base na (Figura 64) permite verificar que para um modelo incremental, sem limites de memória, rápido e com alta taxa de acerto, esta deverá ser uma das melhores opções a ter em consideração, apesar da alta taxa de erro também identificada (com base no registo de algumas anomalias no centro da curva de potência).

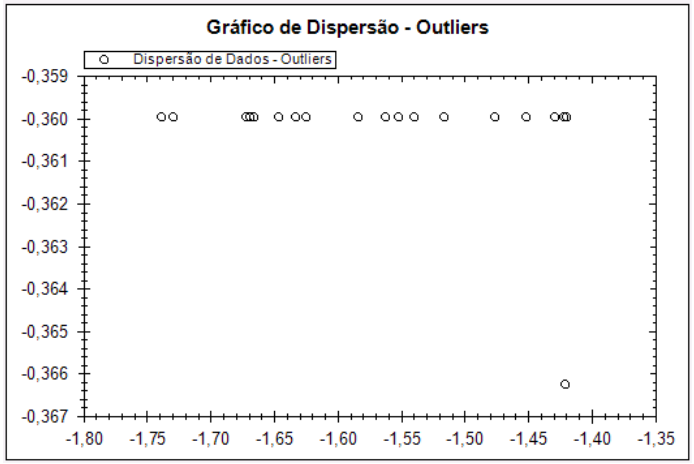

Figura 65 - Ranking com DataStreams e Modelo de Informação vazio (20 registos)

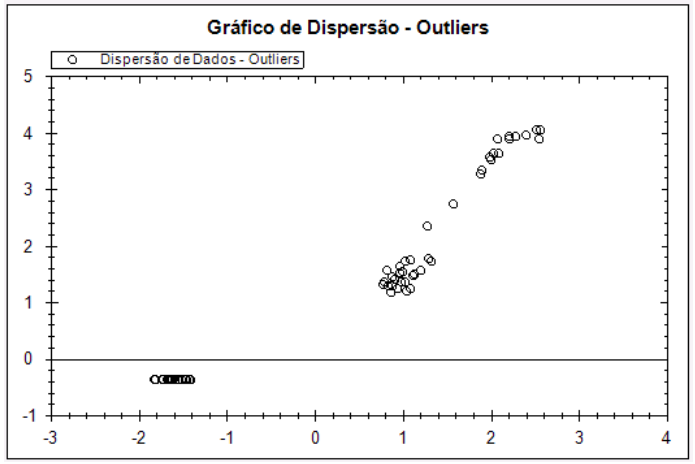

Figura 66 - Ranking com DataStreams e Modelo de Informação vazio (65 registos)

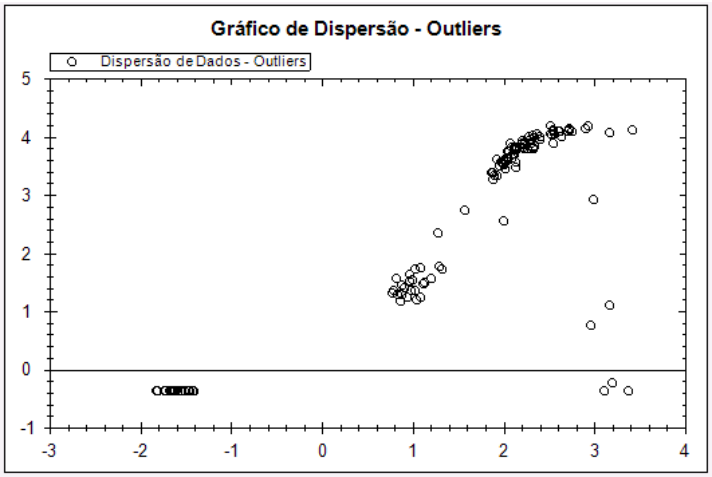

Figura 67 - Ranking com DataStreams e Modelo de Informação vazio (130 registos)

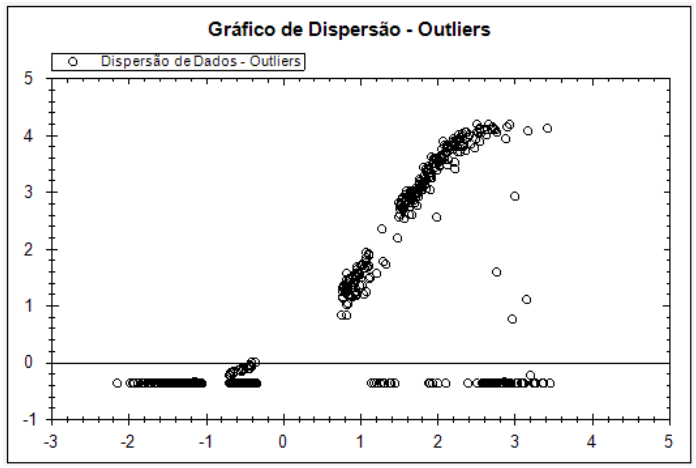

Figura 68 - Ranking com DataStreams e Modelo de Informação vazio (648 registos)

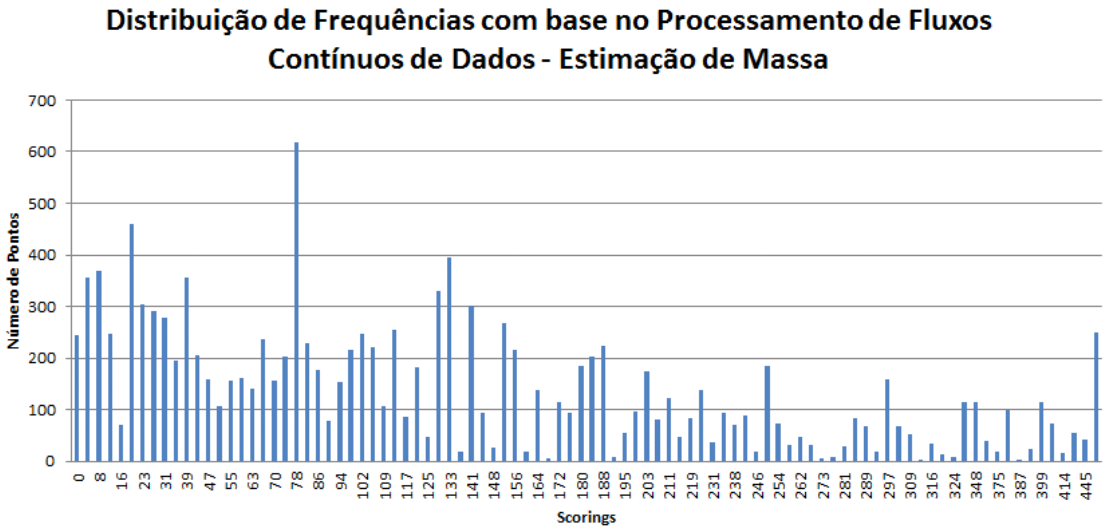

Figura 69 – Diagrama de Distribuição de Frequências com base no Processamento de Fluxos Contínuos de Dados

O diagrama de Distribuição de Frequências (ver Figura 69) sugere a criação de um sistema de pontuação ainda menos aglomerado.

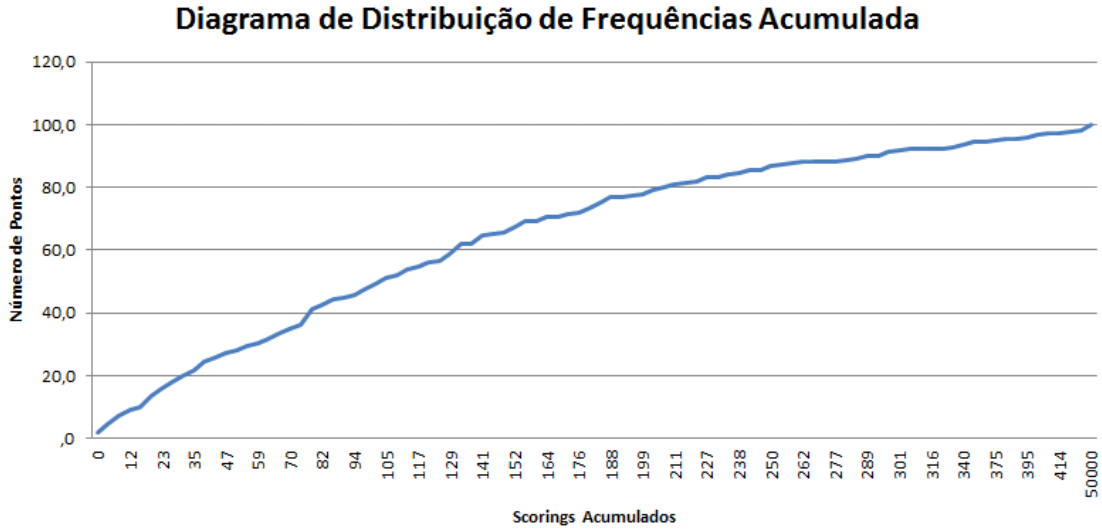

Figura 70 – Diagrama de Distribuição de Frequências Acumulada

Já o diagrama de Distribuição de Frequências Acumulada permite identificar o grupo 59 como limite de significância para os índices de distribuição de anomalias, tendo o algoritmo detetado a presença de 3963 registos anómalos. Este valor elevado teve impacto da escolha de uma janela deslizante para a passagem de informação entre o modelo atual e o modelo de referência.

O grupo criado com a pontuação de 50000 foi definido como sendo os registos pertencentes à primeira janela deslizante, e que portanto não tinham resultado de pontuação válido. Foram definidos com uma pontuação elevada para permitir que os

resultados deste grupo não sejam tidos em consideração para o limite de significância de anomalias encontradas.

5.2.5. Comparação de Algoritmos

A comparação de algoritmos terá como base a identificação do tempo de execução dos algoritmos. Neste campo é importante referir que os parâmetros de entrada escolhidos foram idênticos, para permitir uma plataforma de comparação.

Algoritmo	Tempo de Execução (Criação de Modelo e Estimação)	Número de Anomalias Identificadas
1. Estimação de Massa Half-Space Trees de Cortes Aleatórios	10 segundos	1019 registos
2. Estimação de Massa Half-Space Trees de Cortes Sequenciais	8 segundos	1106 registos
3. Local Outlier Factor	600 segundos	1012 registos
4. Estimação de Massa em Fluxos Contínuos de Dados	12 segundos	3963 registos

Tabela 5 – Tempo de Execução e Número de Anomalias dos Algoritmos Implementados

Com base nos tempos de execução (Tabela 5) e nos diferentes anexos com os outputs de execução dos algoritmos (anexo 4, anexo 5, anexo 6 e anexo 7) é possível identificar que:

- o LOF consegue consegue delinear melhor a estrutura inerente à Curva de Potência identificando a possível existência de anomalias coletivas (Figura 49, Figura 56, Figura 63, Figura 68);
- os tempos de execução demonstram a rapidez inerente aos algoritmos de Estimação de Massa;
- o modelo de fluxos contínuos permite a criação de um modelo incremental sem limites de memória, e capaz de identificar novas ocorrências;

- a existência de falsos positivos no modelo de fluxos contínuos, característica sugerida quando se compara o número de anomalias dos diferentes algoritmos, pressupõe a necessidade de implementação de um modelo híbrido com uma componente *offline* (em que volta a processar as anomalias identificadas na componente *online*) e a componente *online* (com recurso às Half-Space Trees); ou a simples mudança de janela deslizante para uma janela que acumule o conhecimento ao longo do tempo;
- a caracterização das diferentes curvas de energia mediante o limite de significância de 648 registos identifica a possibilidade de criação de modelos compactos, simples, ágeis e multivariados, com o recurso à temática de Estimação de Massa.

Capítulo 6

Conclusões

Na temática de identificação de anomalias existem diferentes caminhos que podem ser tidos em consideração. Mesmo quando se considera a temática de Estimação de Massa, as possibilidades de investigação são múltiplas.

Com a presente dissertação foram desenvolvidos esforços para aplicar a temática de estimação de massa, um tópico recente na área de *Datamining*, a conjuntos de registos provenientes de um parque eólico. Para efeitos de análise apenas foram tidos em consideração os registos de janeiro a março de 2010 para o aerogerador com o código WTG01 por restrições de memória da máquina de implementação (Processador de 64 bits, Windows 7 com 3 Gb de Memória RAM).

As principais conclusões obtidas com o presente trabalho foram que:

- o conceito de Estimação de Massa permite desenvolver algoritmos simples, eficazes e rápidos;
- algoritmos baseados em densidade (neste caso o LOF) conseguem se adaptar à estrutura dos dados;
- a execução do LOF é demorada, especialmente quando o volume de dados é grande (uma vez que necessita de muitos cálculos intermédios);
- os resultados obtidos com os diferentes algoritmos melhoram à medida que se trabalham diferentes parâmetros (no caso de estimação de massa quando se aumenta o valor referente ao número de árvores, no caso do LOF quando se aumenta o valor referente aos vizinhos mais próximos);

- à medida que os algoritmos criam grupos de *scoring* mais distintos, e com grupos de valores mais baixos, os resultados são mais fiáveis;
- pode-se tornar complicado identificar *outliers* coletivos nos algoritmos baseados em Estimação de Massa (o que sugere a possível necessidade de adaptação da função de pontuação).

Em termos de trabalho futuro para a aplicação criada é possível identificar como principais objetivos a possibilidade de:

- trabalhar com dados previamente classificados como anomalias;
- permitir a inserção de ficheiros com pontuações previamente calculadas, criando um sistema capaz de comparar diferentes tipos de algoritmos;
- criar um mapa do grau de anomalia dos registos com base na atribuição de uma cor capaz de variar consoante o índice de anomalia;
- aplicar a análise a outras variáveis enviadas pelos sistemas SCADA;
- descortinar a existência de *outliers* coletivos e *outliers* pontuais nos algoritmos de Estimação de Massa com base num estudo em relação à função de pontuação.

Já em termos de estudar a aplicabilidade dos métodos refere-se a necessidade de:

- avaliar a aplicabilidade de estimação de massa em outras atividades, em que normalmente se utilize estimação de densidade;
- detetar padrões com base em análise de séries temporais e assim diminuir o volume de dados em análise;
- avaliar a capacidade de robustez do sistema com base em diferentes *datasets* de dados, aplicando os conceitos a outro tipo de registos de monitorização (p.e. com base em registos referentes a equipamento rodoviário, aéreo ou ferroviário).

Após o desenvolvimento de um trabalho desta dimensão torna-se importante definir como possíveis planos, interessantes do ponto de vista de implementação:

- criar um modelo de informação em que cada floresta processa um conjunto de atributos distintos (relacionados com um tipo de problema em particular),

tentando assim identificar um tipo de anomalia em separado (com base no anexo 8 é possível verificar que as anomalias identificadas com apenas alguns atributos são possíveis anomalias detectadas noutras curvas de energia). Para tal sugere-se uma análise das possíveis falhas associadas a aerogeradores, com base no trabalho de (Brandão *et al.*, 2007; Brandão *et al.*, 2009), entre outros;

- criar um modelo de informação em que cada floresta processa os registos referentes a cada aerogerador, tentando identificar quebras de potência no parque eólico.

Referências

Aggarwal, C. C. (2013). Outlier Analysis.

Aguiar, A. R. d. S. M. (2009). "Utilização de Dados Operacionais para a Compreensão do Desempenho dos Aerogeradores." 20.

Anaya-Lara, O., N. Jenkins and J. R. McDonald (2006). Communications Requirements and Technology for Wind Farm Operation and Maintenance. First International Conference on Industrial and Information Systems. Sri Lanka, IEEE.

Aziz, M. A., H. Noura and A. Fardoun (2010). General Review of Fault Diagnosis in Wind Turbines. 18th Mediterranean Conference on Control & Automation Congress Palace Hotel, Marrakech, Morocco

Barnett, V. and T. Lewis (1994). Outliers in Statistical Data, John Wiley.

Bifet, A., G. Holmes and R. Kirkby (2010). "MOA: Massive Online Analysis."

Brandão, R. F. M., J. A. B. Carvalho and F. P. M. Barbosa (2007). "Fault Detection on Wind Generators."

Brandão, R. F. M., J. A. B. Carvalho and F. P. M. Barbosa (2009). Forced outage time analysis of a portuguese wind farm Universities Power Engineering Conference (UPEC), 2009 Proceedings of the 44th International Glasgow

Chandola, V., A. Banerjee and V. Kumar "Outlier Detection: A survey."

Daneshi-Far, Z., G. A. Capolino and H. Henao (2010). Review of Failures and Condition Monitoring in Wind Turbine Generators XIX International Conference on Electrical Machines - ICEM 2010. Rome, Italy, IEEE.

ENEOP. (2013). "O que é a Energia Eólica." from http://www.eneop.pt/canais.asp?id_canal=110.

Erdoğan, G. (2011). Outlier Detection Toolbox in MATLAB.

Guo, H., X. Yang, J. Xiang and S. Watson (2009). Wind Turbine Availability Analysis Based On Statistical Data. Sustainable Power Generation and Supply, 2009. SUPERGEN '09. International Conference on Nanjing

Gupta, M., J. Gao, C. C. Aggarwal and J. Han (2013). "Outlier Detection for Temporal Data: A Survey." IEEE Transactions on Knowledge and Data Engineering **25**(1).

Harman, K. D. and P. G. Raftery (2003). Analytical Tecniques for Understanding the Performance of Operational Wind Farms. European Wind Energy Conference, 2003.

Hawkins, D. M. (1980). Identification of Outliers, Chapman and Hall.

Komsta, L. (2011). "Package Outliers."

Kriegel, P. H.-P. (2012). "Environment for DeveLoping KDD-Applications Supported by Index-Structures."

Kusiak, A. (2010) "What You Don't Know About Power Curves " The University of Iowa, Wind Power Management Program, Iowa City, Iowa.

Kusiak, A. and A. Verma (2013). "Monitoring Wind Farms with Performance Power Curves." IEEE Transactions on Sustainable Energy **4**(1).

Kusiak, A., H. Zheng and Z. Song (2009). "Models for monitoring wind farm power." Renewable Energy **34**.

Kusiak, A., H. Zheng and Z. Song (2009). "On-line Monitoring of Power Curves." Renewable Energy.

Lda., M. D. (2012). "Portugal no 10ºlugar Mundial em capacidade Eólica instalada ", from http://www.cursosenergiasrenovaveis.com/index.php/2012/10/portugal-no-10-lugar-mundial-em-capacidade-eolica-instalada/.

Lindahl, S. and K. Harman (2012). Analytical Techniques for Performance Monitoring of Modern Wind Turbines. EWEA 2012. G. Hassan. Copenhagen.

Llango, V., R. Subramanian and V. Vasudevan (2012). "A Five Step Procedure for Outliers Analysis in Datamining." European Journal of Scientific Research **75**(3): 327-339.

Munõs-Garcia, J., J. L. Moreno-Rebollo and A. Pascual-Acosta (1990). "Outliers: a formal approach." 215-226.

Pokrajac, D., A. Lazarevic and L. J. Latecki (2007). Incremental Local Outlier Detection for Data Streams. IEEE Symposium on Computational Intelligence and Data Mining (CIDM).

Rakotomalala, R. (2005). TANAGRA : un logiciel gratuit pour l'enseignement et la recherche, Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005.

Soman, S. S., H. Zareipour and P. Maldal (2010). "**A Review of Wind Power and Wind Speed Forecasting Methods with Different Time Horizons**." IEEE.

Ting, K. M., J. T. S. Chuan and F. T. Liu (2010). "Mass versus Density: Which is a Better Ranking Measure for Anomaly Detection." Journal of Artificial Intelligence.

Ting, K. M., J. T. S. Chuan and F. T. Liu (2011). Fast Anomaly Detection for Streaming Data. Twenty-Second International Joint Conference on Artificial Intelligence.

Ting, K. M., G.-T. Zhou, F. T. Liu and S. C. Tan (2013). "Mass Estimation." Machine Learning.

Tongkumchum, P. (2005). "Two Dimensional Box-Plot."

UK, W. P. P. (2009). "Wind Turbine Power Output Variation With Steady Wind Speed."

Anexos

Nota: Na presente dissertação não se encontra disponibilizado o código fonte dos métodos de estimação de anomalias implementados.

Prevê-se a sua disponibilização na Internet num futuro próximo.

Anexo 1 – Código R para Análise Exploratória Parcial

```
tsoutliers <- function(x,plot=TRUE, labx, laby)
{
   x <- as.ts(x)
   tt <- 1:length(x)
   resid <- residuals(loess(x ~ tt))
   resid.q <- quantile(resid,prob=c(0.25,0.75))
   iqr <- diff(resid.q)
   limits <- resid.q + 1.5*iqr*c(-1,1)
   score <- abs(pmin((resid-limits[1])/iqr,0) + pmax((resid - limits[2])/iqr,0))
   if(plot)
   {
      plot(x, xlab = labx, ylab= laby)
      x2 <- ts(rep(NA,length(x)))
      x2[score>0] <- x[score>0]
      tsp(x2) <- tsp(x)
      points(x2,pch=19,col="red")
      grid()
      return(invisible(score))
   }
   else
      return(score)
}
x1 <- read.table("D:\\D\\WTG01_10minchanged.csv", header=TRUE, sep = ';')
names(x1)
summary(x1)
```

```
#1ª variável
plot.ts(x1$Wind.Speed, ylab="Velocidade do Vento", xlab="Tempo")
grid()
out1 <- boxplot(x1$Wind.Speed, main="Velocidade do Vento")
```

```
grid()
plot(out1$out)
grid()
summary(out1$out)
score1 <- tsoutliers(x1$Wind.Speed, TRUE, "Tempo", "Velocidade do Vento")
plot(score1)
table(score1)
```

```
#2ºgráfico
plot.ts(x1$LV.ActivePower, ylab="Potência Gerada", xlab="Tempo")
grid()
out2 <- boxplot(x1$LV.ActivePower, main="Potência Gerada")
grid()
plot(out2$out)
grid()
summary(out2$out)
score2 <- tsoutliers(x1$LV.ActivePower, TRUE, "Tempo", "Potência Gerada")
```

```
#3ºgráfico
plot.ts(x1$Gen.RpmMonitor, ylab="Rotações por Minuto", xlab="Tempo")
grid()
out3 <- boxplot(x1$Gen.RpmMonitor, main="Rotações por Minuto")
grid()
plot(out3$out)
grid()
summary(out3$out)
tsoutliers(x1$Gen.RpmMonitor, TRUE, "Tempo", "Rotações por Minuto")
```

```
#4ºgráfico
plot.ts(x1$Pitchangle.1..Red., ylab="Posição da Nacelle", xlab="Tempo")
grid()
out4 <- boxplot(x1$Gen.RpmMonitor, main="Posição da Nacelle")
grid()
plot(out4$out)
```

```
grid()
summary(out4$out)
tsoutliers(x1$Pitchangle.1..Red., TRUE, "Tempo", "Posição Nacelle")
```

```
#5ºgráfico
plot(x1$LV.ActivePower ~ x1$Wind.Speed, ylab="Potência Gerada", xlab="Velocidade do Vento")
grid()
```

```
#6ºgráfico
plot(x1$Gen.RpmMonitor ~ x1$Wind.Speed, ylab="Rotações por Minuto", xlab="Velocidade do Vento")
grid()
```

```
#7ºgráfico
plot(x1$Pitchangle.1..Red. ~ x1$Wind.Speed, ylab="Posição da Nacelle", xlab="Velocidade do Vento")
grid()
```

Anexo 2 – Estatísticas Descritivas das Variáveis Analisadas

	Estatística	Valor
1.	Mínimo	0 (m/s)
2.	1ºQu.	3.817 (m/s)
3.	Mediana	5.919 (m/s)
4.	Média	5.962 (m/s)
5.	3ºQu.	7.761 (m/s)
6.	Máximo	16.020 (m/s)

Tabela 6 – Estatísticas Descritivas da Velocidade do Vento

	Estatística	Valor
1.	Mínimo	-7.153 kW
2.	1ºQu.	0 kW
3.	Mediana	198.600 kW
4.	Média	414.300 kW
5.	3ºQu.	620.400 kW
6.	Máximo	2516.000 kW

Tabela 7 – Estatísticas Descritivas da Potência Gerada

	Estatística	Valor
1.	Mínimo	0 (r.p.m)
2.	1ºQu.	32.91 (r.p.m.)
3.	Mediana	837.50 (r.p.m.)
4.	Média	655.50 (r.p.m.)
5.	3ºQu.	1026.00 (r.p.m.)
6.	Máximo	1167.00 (r.p.m.)

Tabela 8 – Estatísticas Descritivas das Rotações por Minuto

	Estatística	Valor
1.	Mínimo	-0.3503 (º)
2.	1ºQu.	-0.2828 (º)
3.	Mediana	-0.2231 (º)
4.	Média	25.0900 (º)
5.	3ºQu.	69.5100 (º)
6.	Máximo	89.6100(º)

Tabela 9 – Estatísticas Descritivas da Posição da Nacelle

Anexo 3 – Scoring segundo Análise da Série Temporal para as variáveis Velocidade do Vento e Potência Gerada

ID	Score
12299	0,00113
1244	0,003471
8578	0,00875
1247	0,009972
8713	0,01103
1233	0,025217
8580	0,026088
8588	0,02717
12297	0,030785
4285	0,042951
8584	0,06182
12298	0,070381
8562	0,077783
1240	0,128838
4286	0,148929
8569	0,1502
8583	0,159111
1241	0,159725
8560	0,160009
8589	0,170826
8571	0,174706
8592	0,182265
8593	0,183457
8576	0,204053
8586	0,20955662
8568	0,22379731
8563	0,24436014
8591	0,25441968
8585	0,2774571
12288	0,30759806
8595	0,31242401
8574	0,32064585
8573	0,32071639
8577	0,32257373
8575	0,33666249
8567	0,36339333
8561	0,36690737
8570	0,38916954
8597	0,41973512
8582	0,42835913
8587	0,44486429
8594	0,50518145
8572	0,51685001
8565	0,5509602
8596	0,56716998

Tabela 10 – Scoring de Outliers Univariados da Série Temporal com base na Variável de Velocidade do Vento

ID	Score
6027	0,002159
2340	0,004753
11057	0,006037
11233	0,006515
8471	0,010553
7417	0,011403
11232	0,01231
10723	0,012492
7655	0,014175
8499	0,015491
2341	0,015558
3918	0,017385
11795	0,020174
10725	0,022983
2366	0,023947
7561	0,024396
7550	0,025777
11796	0,026278
3842	0,028996
10603	0,029264
4716	0,029449
2368	0,031919
4877	0,032717
11231	0,032747
2322	0,033596
4858	0,034656
7545	0,035305
10307	0,035394
11235	0,037937
11504	0,03892
2317	0,040652
11819	0,043956
1390	0,04652
3972	0,050661
11069	0,051028
3920	0,053973
5122	0,054446
11799	0,056064
11804	0,0572
4718	0,061062
8455	0,062082
7645	0,06627837
1394	0,06639325
11487	0,06692844
10862	0,06814738
10590	0,06859595
4184	0,06927557
10754	0,06972988
11478	0,07209885
1392	0,07236394
11099	0,07372456
10718	0,07748094
1224	0,07959569
10853	0,08008424
4183	0,08034495
7423	0,08070801
3978	0,08137181
3950	0,08227557
3914	0,08459594
1155	0,08481663
7662	0,0853302
8514	0,08578185
11230	0,08661718
11812	0,0875725
11486	0,09264587
2367	0,09611024
3971	0,09615072
7651	0,09644128
11805	0,09710086
10870	0,09823904
2364	0,09848442
10299	0,10019688
11506	0,10752888
7673	0,1096888
11234	0,11087529
11061	0,11275186
10733	0,11486537
8482	0,12063656
11507	0,12106816
3912	0,12410971
7402	0,12616949
6011	0,12664852

Tabela 11 – Scoring de Outliers Univariados da Série Temporal com base na variável da Potência Gerada

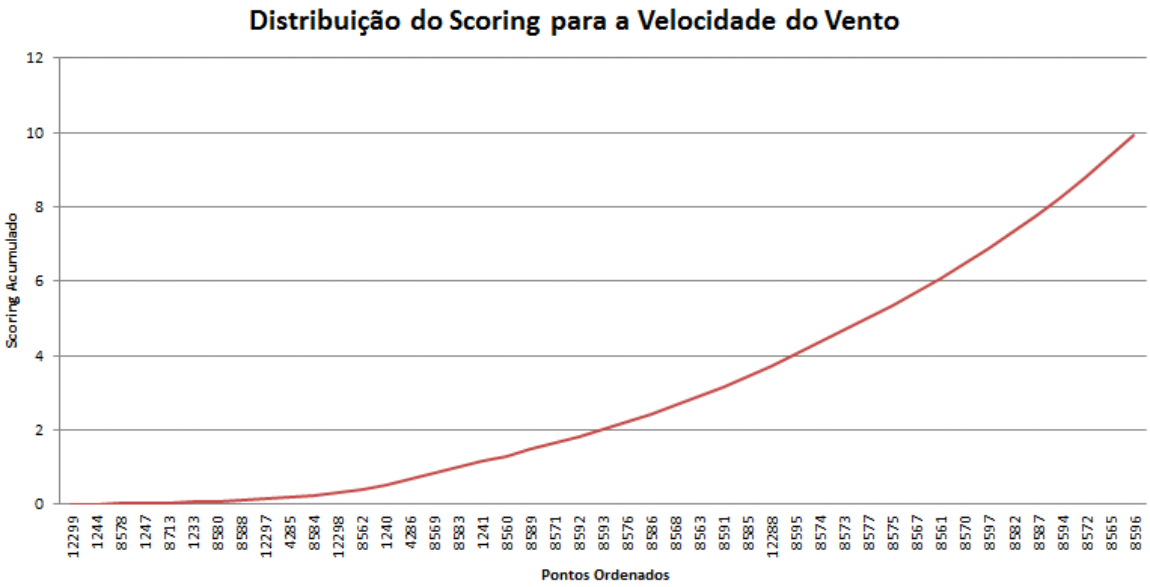

Figura 71 - Distribuição do Scoring para a Velocidade do Vento

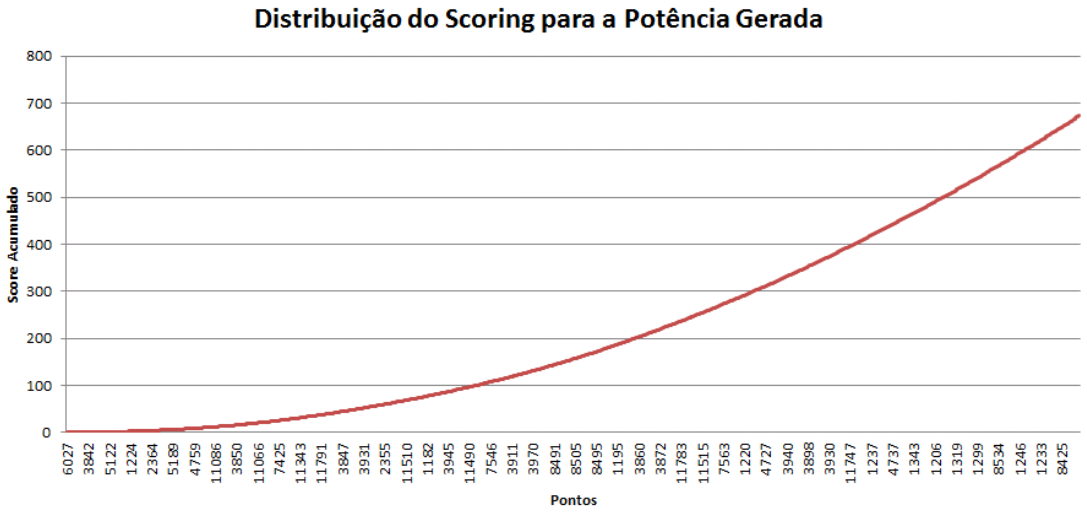

Figura 72 - Distribuição do Scoring para a Potência Gerada

Anexo 4 – Análise do Scoring de Half-Space Trees de Cortes Aleatórios

	Frequência	Percentagem	Percentagem Válida	Percentagem Acumulada
3.58	1	,0	,0	,0
3.91	2	,0	,0	,0
4.09	8	,1	,1	,1
4.17	11	,1	,1	,2
4.24	21	,2	,2	,3
4.39	77	,6	,6	,9
4.64	228	1,8	1,8	2,7
4.81	671	5,2	5,2	7,9
4.86	1693	13,1	13,1	20,9
4.91	2782	21,5	21,5	42,4
5.04	2798	21,6	21,6	64,0
5.25	1998	15,4	15,4	79,4
5.29	1175	9,1	9,1	88,5
5.39	619	4,8	4,8	93,2
5.52	364	2,8	2,8	96,0
5.64	205	1,6	1,6	97,6
5.7	13	,1	,1	97,7
27.32	190	1,5	1,5	99,2
40.99	3	,0	,0	99,2
1379.85	101	,8	,8	100,0
Total	12960	100,0	100,0	

Tabela 12 - Tabela de Distribuição de Frequências do Scoring obtido para o Algoritmo Aleatório de Half-Space Trees

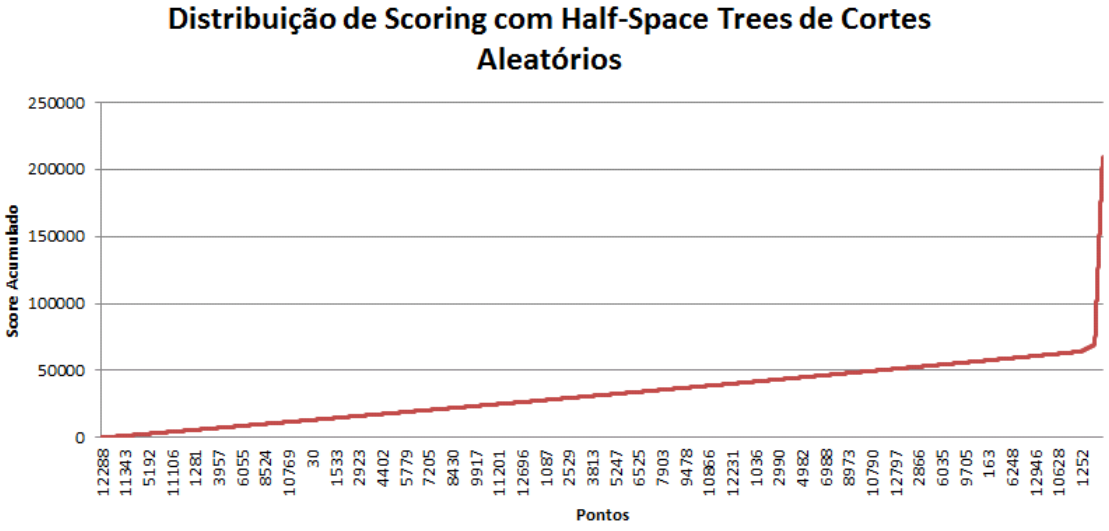

Figura 73 – Distribuição de Scoring com Half-Space Trees de Cortes Aleatórios

Anexo 5 – Análise do Scoring de Half-Space Trees de Cortes Sequenciais

	Frequência	Percentagem	Percentagem Válida	Percentagem Acumulada
1.58	1	,0	,0	,0
3.17	1	,0	,0	,0
3.46	1	,0	,0	,0
3.58	1	,0	,0	,0
3.7	2	,0	,0	,0
3.81	7	,1	,1	,1
3.91	14	,1	,1	,2
4	27	,2	,2	,4
4.09	53	,4	,4	,8
4.17	79	,6	,6	1,4
4.25	114	,9	,9	2,3
4.32	188	1,5	1,5	3,8
4.39	289	2,2	2,2	6,0
4.46	329	2,5	2,5	8,5
4.52	740	5,7	5,7	14,2
4.58	775	6,0	6,0	20,2
4.64	1405	10,8	10,8	31,1
4.7	1189	9,2	9,2	40,2
4.75	1768	13,6	13,6	53,9
4.81	1101	8,5	8,5	62,4
4.86	1459	11,3	11,3	73,6
4.91	756	5,8	5,8	79,5
4.95	876	6,8	6,8	86,2
5	380	2,9	2,9	89,2
5.04	504	3,9	3,9	93,0
5.09	187	1,4	1,4	94,5
5.13	251	1,9	1,9	96,4
5.17	42	,3	,3	96,8
5.21	86	,7	,7	97,4
5.24	12	,1	,1	97,5
5.29	36	,3	,3	97,8
5.32	4	,0	,0	97,8
5.36	10	,1	,1	97,9
5.43	4	,0	,0	97,9
27.32	168	1,3	1,3	99,2
1379.85	101	,8	,8	100,0
Total	12960	100,0	100,0	

Tabela 13 - Tabela de Distribuição de Frequências do Scoring obtido para o Algoritmo Sequencial de Half-Space Trees

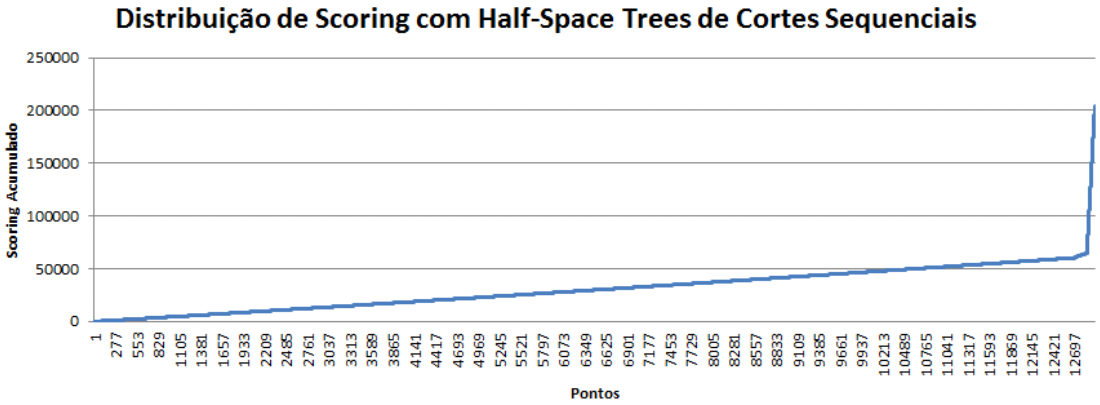

Figura 74 – Distribuição de Scoring com Half-Space Trees de Cortes Sequenciais

Anexo 6 – Análise do Scoring do LOF (Local Outlier Factor)

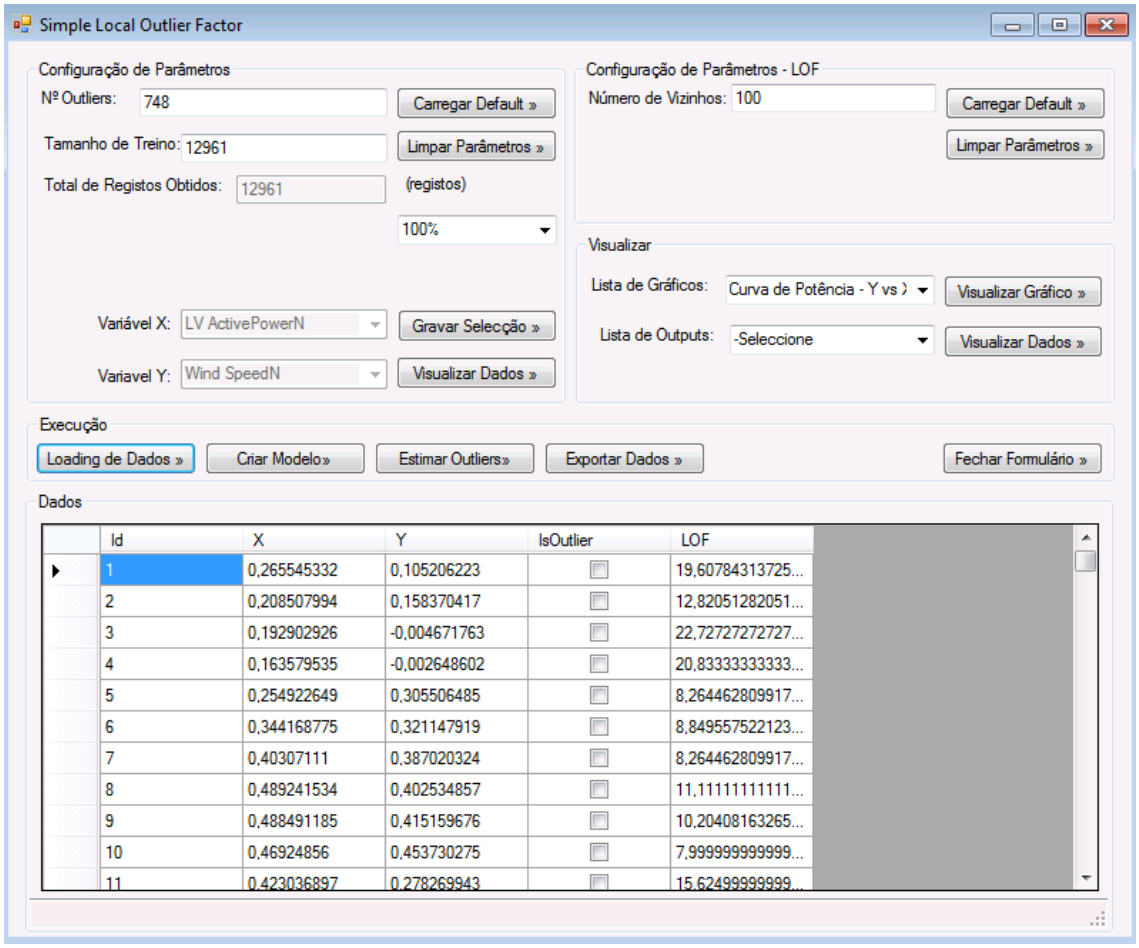

Figura 75 - Interface do Local Outlier Factor

Scoring LOF				
	Frequência	Percentagem	Percentagem Válida	Percentagem Acumulada
13.51	67	,5	,5	37,4
13.70	67	,5	,5	37,9
13.89	96	,7	,7	38,7
14.08	73	,6	,6	39,3
14.29	83	,6	,6	39,9
14.49	94	,7	,7	40,6
14.71	96	,7	,7	41,4
14.93	69	,5	,5	41,9
15.15	67	,5	,5	42,4
15.38	72	,6	,6	43,0
15.63	85	,7	,7	43,7
15.87	75	,6	,6	44,2
16.13	86	,7	,7	44,9
16.39	87	,7	,7	45,6
16.67	112	,9	,9	46,5
16.95	132	1,0	1,0	47,5
17.24	105	,8	,8	48,3
17.54	112	,9	,9	49,2
17.86	138	1,1	1,1	50,2
18.18	125	1,0	1,0	51,2
18.52	153	1,2	1,2	52,4
18.87	119	,9	,9	53,3
19.23	142	1,1	1,1	54,4
19.61	113	,9	,9	55,3
20.00	128	1,0	1,0	56,3
20.41	136	1,0	1,1	57,4
20.83	175	1,4	1,4	58,7
21.28	197	1,5	1,5	60,3
21.74	189	1,5	1,5	61,7
22.22	184	1,4	1,4	63,2
22.73	187	1,4	1,5	64,6
23.26	155	1,2	1,2	65,8
23.81	112	,9	,9	66,7
24.39	119	,9	,9	67,6
25.00	138	1,1	1,1	68,7
25.64	150	1,2	1,2	69,9
26.32	122	,9	,9	70,8
27.03	143	1,1	1,1	71,9

27.78	159	1,2	1,2	73,2
28.57	193	1,5	1,5	74,7
29.41	188	1,5	1,5	76,1
30.30	221	1,7	1,7	77,8
31.25	209	1,6	1,6	79,5
32.26	219	1,7	1,7	81,2
33.33	202	1,6	1,6	82,7
34.48	128	1,0	1,0	83,7
35.71	164	1,3	1,3	85,0
37.04	123	,9	1,0	86,0
38.46	111	,9	,9	86,8
40.00	107	,8	,8	87,7
41.67	72	,6	,6	88,2
43.48	49	,4	,4	88,6
45.45	28	,2	,2	88,8
47.62	26	,2	,2	89,0
50.00	29	,2	,2	89,2
52.63	57	,4	,4	89,7
55.55	91	,7	,7	90,4
58.82	223	1,7	1,7	92,1
62.50	151	1,2	1,2	93,3
66.67	333	2,6	2,6	95,9
71.43	345	2,7	2,7	98,6
76.92	181	1,4	1,4	100,0
83.33	1	,0	,0	100,0
657.81	1	,0	,0	100,0
Total	12859	99,2	100,0	

Tabela 14 - Scorings obtidos com o LOF para os primeiros 64 grupos

Anexo 7 – Análise do Scoring do Processamento de Fluxos Contínuos de Dados com um Modelo de Informação Vazio

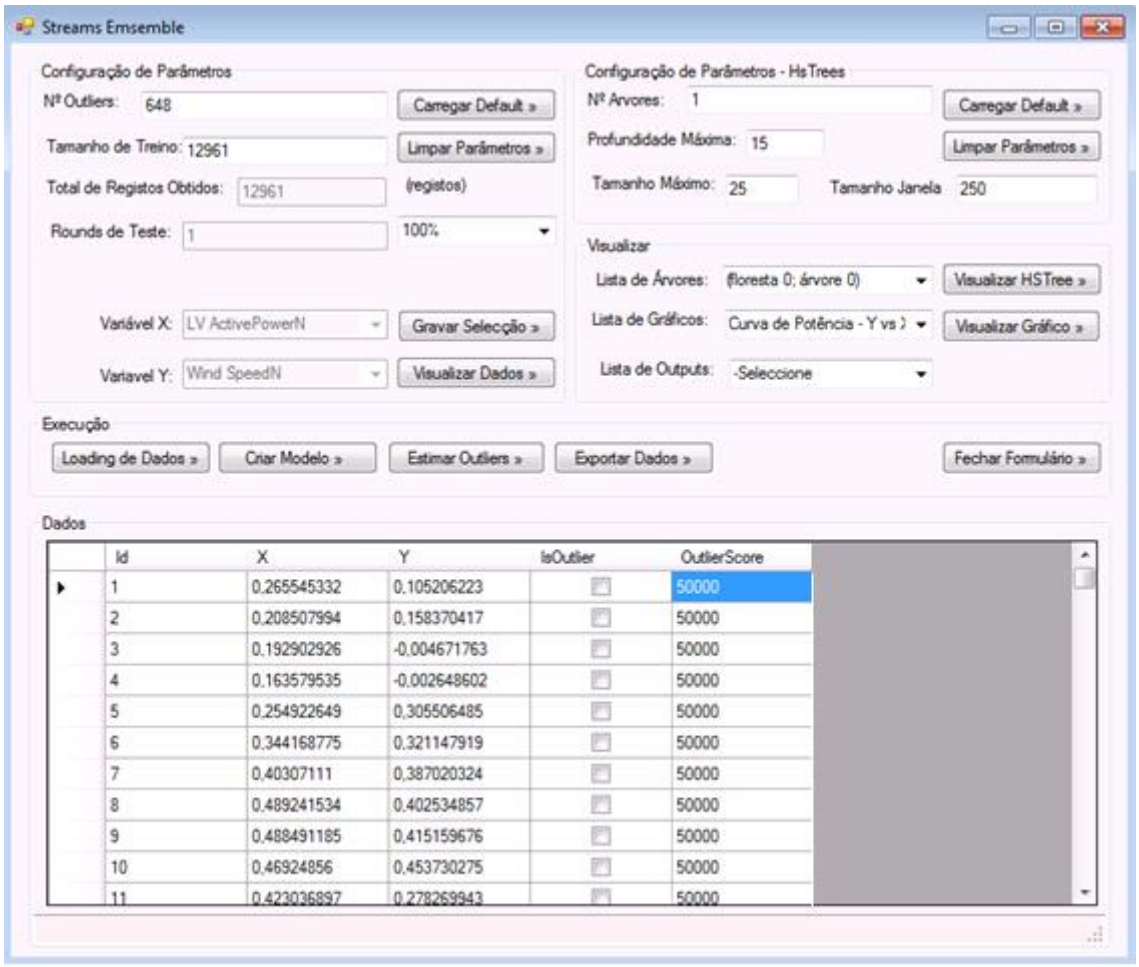

Figura 76 – Interface do Processamento de Fluxos Contínuos de Dados com base em Estimação de Massa

Scoring Fluxos Contínuos de Dados

	Frequência	Percentagem	Percentagem Válida	Percentagem Acumulada
0	244	1,9	1,9	1,9
4	357	2,8	2,8	4,6
8	368	2,8	2,8	7,5
12	246	1,9	1,9	9,4
16	70	,5	,5	9,9
20	461	3,6	3,6	13,5
23	305	2,4	2,4	15,8
27	292	2,3	2,3	18,1
31	279	2,2	2,2	20,2
35	196	1,5	1,5	21,7
39	355	2,7	2,7	24,5
43	205	1,6	1,6	26,1
47	159	1,2	1,2	27,3
51	108	,8	,8	28,1
55	157	1,2	1,2	29,3
59	161	1,2	1,2	30,6
63	140	1,1	1,1	31,7
66	237	1,8	1,8	33,5
70	156	1,2	1,2	34,7
74	202	1,6	1,6	36,3
78	619	4,8	4,8	41,0
82	230	1,8	1,8	42,8
86	178	1,4	1,4	44,2
90	78	,6	,6	44,8
94	155	1,2	1,2	46,0
98	216	1,7	1,7	47,6
102	246	1,9	1,9	49,5
105	220	1,7	1,7	51,2
109	108	,8	,8	52,1
113	255	2,0	2,0	54,0
117	86	,7	,7	54,7
121	182	1,4	1,4	56,1
125	48	,4	,4	56,5
129	329	2,5	2,5	59,0
133	396	3,1	3,1	62,1
137	20	,2	,2	62,2
141	302	2,3	2,3	64,6
145	94	,7	,7	65,3

148	28	,2	,2	65,5
152	269	2,1	2,1	67,6
156	216	1,7	1,7	69,2
160	18	,1	,1	69,4
164	138	1,1	1,1	70,4
168	5	,0	,0	70,5
172	115	,9	,9	71,4
176	93	,7	,7	72,1
180	185	1,4	1,4	73,5
184	202	1,6	1,6	75,1
188	225	1,7	1,7	76,8
191	8	,1	,1	76,9
195	54	,4	,4	77,3
199	96	,7	,7	78,0
203	175	1,4	1,4	79,4
207	82	,6	,6	80,0
211	124	1,0	1,0	81,0
215	47	,4	,4	81,3
219	83	,6	,6	82,0
227	138	1,1	1,1	83,0
231	36	,3	,3	83,3
234	94	,7	,7	84,0
238	72	,6	,6	84,6
242	90	,7	,7	85,3
246	18	,1	,1	85,4
250	185	1,4	1,4	86,9
254	73	,6	,6	87,4
258	33	,3	,3	87,7
262	48	,4	,4	88,0
270	33	,3	,3	88,3
273	6	,0	,0	88,3
277	8	,1	,1	88,4
281	29	,2	,2	88,6
285	84	,6	,6	89,3
289	67	,5	,5	89,8
293	18	,1	,1	89,9
297	158	1,2	1,2	91,1
301	69	,5	,5	91,7
309	52	,4	,4	92,1
313	2	,0	,0	92,1
316	34	,3	,3	92,4

320	14	,1	,1	92,5
324	9	,1	,1	92,5
340	116	,9	,9	93,4
348	116	,9	,9	94,3
367	39	,3	,3	94,6
375	20	,2	,2	94,8
379	100	,8	,8	95,6
387	1	,0	,0	95,6
395	24	,2	,2	95,7
399	114	,9	,9	96,6
402	73	,6	,6	97,2
414	17	,1	,1	97,3
434	54	,4	,4	97,7
445	42	,3	,3	98,1
50000	251	1,9	1,9	100,0
Total	12960	100,0	100,0	

Tabela 15 - Tabela de Distribuição de Frequências de Pontuações Atribuídas com base no processamento de Fluxos Contínuos de Dados (Estimação de Massa)

Anexo 8 – Identificação de Anomalias nas Restantes Curvas de Energia

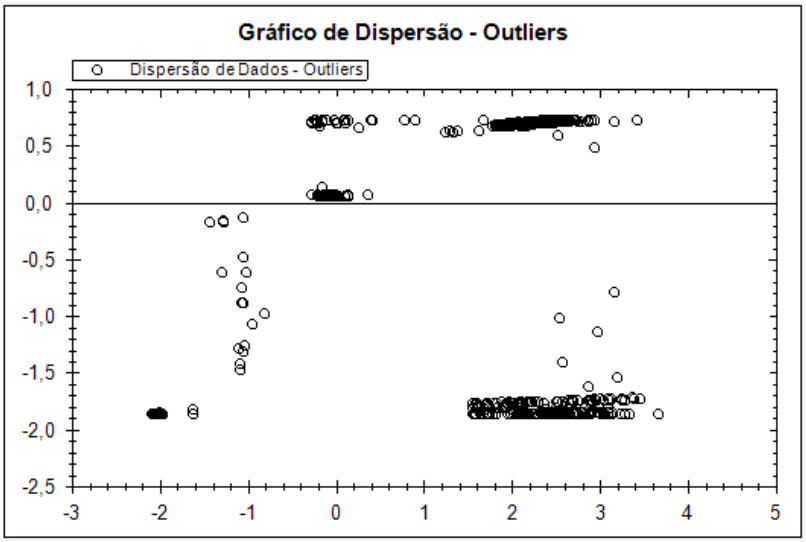

Figura 77 - Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa com Cortes Aleatórios – Curva do Rotor

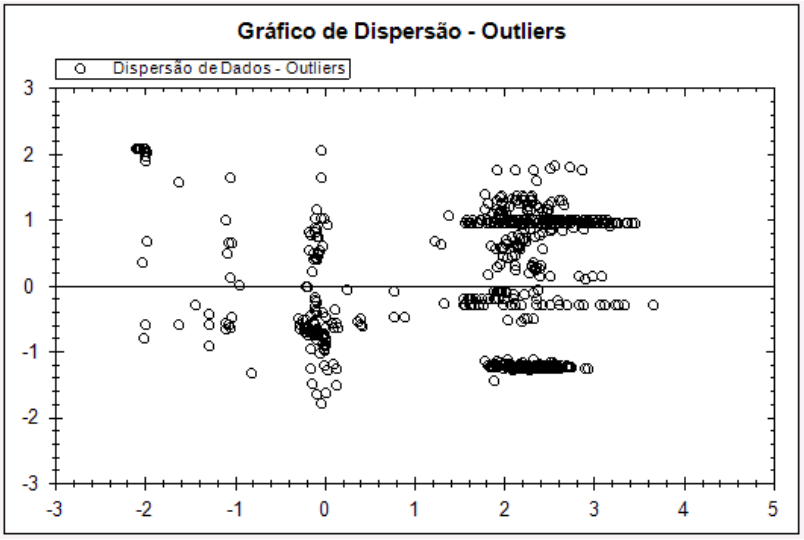

Figura 78 – Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa de Cortes Aleatórios – Curva do Blade Pitch

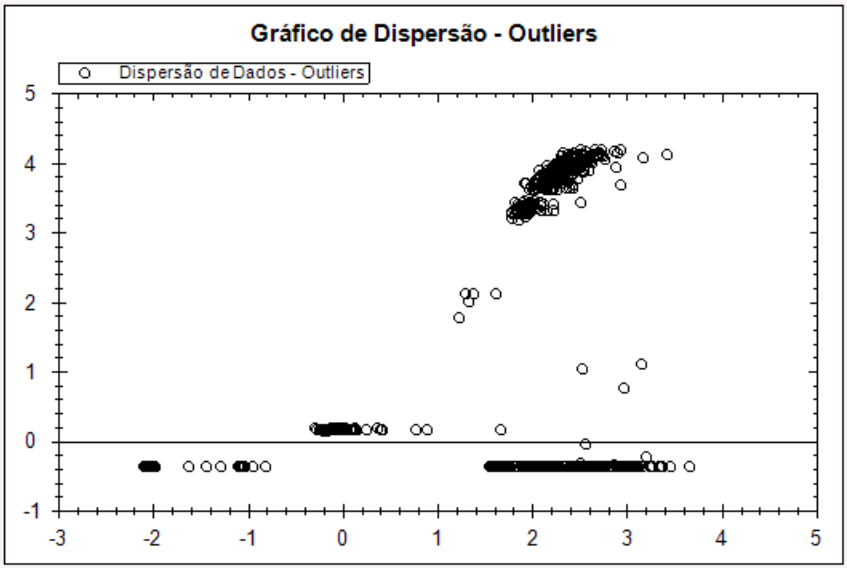

Figura 79 – Anomalias com um limite de significância de 648 registos com a seleção de 1 floresta com 100 árvores distintas – Estimação de Massa de Cortes Aleatórios – Curva de Potência