3
Ambiente experimental

Dado que a regressão PLS é originária da área de quimiometria, vários conjuntos de dados, no total nove, praticamente todos oriundos da prática quimiométrica com espectros NIR[1], foram usados para avaliar os algoritmos apresentados nesta tese. Por este motivo, foi criado o *conjunto NIR*, contendo a descrição e preparação de cada um destes.

Por outro lado, alguns experimentos necessitaram o uso de séries temporais, e da mesma forma que para o conjunto NIR, foi criado o *conjunto Santa Fé* [53]. Ambos são descritos neste capítulo.

PUC-Rio - Certificação Digital Nº 9916205/CA

3.1
Conjunto NIR

Nove conjuntos de dados[2] compõem o conjunto NIR, são eles: Wheat, Light gas oil, Combustible, Metal ions, Corn, Wet grass, Dry grass, Meat e Polymer. A tabela 3.1 resume as principais dimensões destes conjuntos. Observe que o número de amostras é relativamente pequeno se comparado ao número de variáveis.

Wheat

O primeiro conjunto foi fornecido por Kalivas [25]. O conjunto contém o espectro NIR de 100 amostras de trigo juntamente com a quantidade de proteína e umidade correspondente. As amostras foram obtidas usando refletância difusa como $log(1/R)$ de 1100 a 2500nm em intervalos de 2nm. Dos 100 espectros, 70 foram usados para treinamento e os 30 restantes para teste do modelo construído. Além disto, os espectros foram reduzidos a 141 variáveis usando uma a cada 5 das originais.

[1] Near InfraRed.

[2] Os nomes foram mantidos em inglês para que pudessem ser encontrados nas referências.

Tabela 3.1: Conjuntos de dados

Conjunto	N. Amostras	Variáveis Indep.	Variáveis Dep.
Wheat	100	141	2
Light gas oil	114	572	4
Combustible	30	363	3
Metal ions	130	176	3
Corn	80	700	4
Wet grass	282	1050	3
Dry grass	141	1050	3
Meat	215	100	3
Polymer	61	10	4

Light gas oil

Como segundo conjunto de dados, usamos o conjunto *light gas oil* disponível na universidade de Dalhousie [7]. Este conjunto é usado para a calibração de gasóleo leve com relação ao conteúdo em hidrocarbonetos. É composto de 115 amostras vindas de três sub-conjuntos para as quais um espectro em 572 canais foi obtido. Como matrizes de treinamento e validação usamos as primeiras 70 amostras e as 44 restantes respectivamente, juntamente com as concentrações de quatro componentes. Como recomendado pelos autores, a última amostra (115) foi desconsiderada por ser espúria. As concentrações de quatro componentes relacionados com o conteúdo em hidrocarbonetos, foram usadas como variáveis dependentes.

Combustible

Para o terceiro conjunto foram usadas 30 amostras de combustível, para as quais o espectro NIR em 3632 canais foi medido. As amostras foram reduzidas de forma a conterem apenas 363 medidas, usando uma resposta a cada 10. Para treinamento 21 amostras foram usadas (70% do conjunto) e para teste as 9 restantes. Para as variáveis dependentes, as concentrações de três componentes relacionados com o desempenho do combustível, foram usadas.

Metal ions

Este conjunto, também disponível na universidade de Dalhousie [8], foi obtido através de um experimento envolvendo a mistura de três com-

PUC-Rio - Certificação Digital Nº 9916205/CA

ponentes de ions de metal (Co(II), Cr(III), Ni(II)). O espectro foi medido no intervalo de 300-650nm. Medidas foram realizadas em intervalos de 2nm com um tempo de integração de 1s. Para treinamento, 92 amostras foram usadas e para teste as 38 restantes. Para as variáveis dependentes, foram usadas as absorbâncias de Cr, Ni e Co respectivamente.

Corn

Para o quinto conjunto [12], o espectro NIR de amostras de milho foram usadas gerando um total de 80 amostras. A faixa para o comprimento de onda é de 1100-2498nm com medições em intervalos de 2nm, resultando num total de 700 leituras. Para as variáveis dependentes, a quantidade de água, óleo, proteína e amido de cada amostra foi foram empregadas. 70% das amostras foram usadas para treino e o restante para teste.

Wet grass e Dry grass

Estes dois conjuntos foram obtidos na competição do IDRC98 [29], uma conferência realizada em Cambersburg, Pennsylvania. As amostras são provenientes de grama cultivada em solos que receberam quatro níveis de fertilização (0, 50, 250 e 500 ppm de nitrogênio). As concentrações, ou variáveis dependentes, foram obtidas com um LECO CNS-2000, um analisador de Carbono, Nitrogênio e Enxofre. Este instrumento é projetado para medir o conteúdo de carbono, nitrogênio e enxofre em uma grande variedade de compostos orgânicos. O carbono e enxofre são medidos por detecção de radiação infra-vermelha e o nitrogênio é determinado por condutividade. O primeiro conjunto de dados corresponde a amostras úmidas analisadas logo após a colheita, enquanto que o segundo conjunto é composto de amostras secas e moídas. Para cada treinamento são usadas 196 amostras do conjunto *Wet grass* e 96 do *Dry grass*. Para teste, 86 e 45 amostras foram respectivamente usadas.

Meat

Este conjunto é fornecido por Tecator [51]. A tarefa é predizer o conteúdo de gordura de uma amostra de carne com base no seu espectro NIR de absorbância. Os dados foram obtidos com o analisador de comida Tecator trabalhando na região de 850 a 1050nm. Cada amostra é composta por um

PUC-Rio - Certificação Digital Nº 9916205/CA

espectro de 100 canais juntamente com o conteúdo de umidade, gordura e proteínas, medido em porcentagem. 172 amostras foram utilizadas para treino, e as 43 restantes para teste.

Polymer

O último conjunto utilizado foi obtido de testes realizados na fabricação de um polímero. Os dados foram disfarçados por uma transformação linear, de forma a que todas as variáveis ficassem no intervalo de 0,1 a 0,9. Devido à natureza proprietária dos dados, as dez variáveis independentes são medidas de variáveis controladas do processo de fabricação (por exemplo temperaturas, taxas de alimentação, entre outras) e as quatro variáveis dependentes são medidas da produção. Das 61 amostras disponíveis, 42 foram usadas para treinamento e 19 para teste.

PUC-Rio - Certificação Digital Nº 9916205/CA

3.2 Conjunto Santa Fé

O Instituto Santa Fé organizou em 1992 o *Santa Fe Time Series Prediction and Analysis Competition*[53], para reunir e comparar os métodos mais sofisticados de predição de Series temporais. Para esta finalidade foram disponibilizados 6 series[48], de A a F, provenientes de experimentos ou observações em diversas áreas, como astro-física ou mercado financeiro. Nesta Tese, foi empregadas a séries D, descrita a seguir.

3.2.1 Serie D, Sintética

A série D foi numéricamente gerada para a competição, contendo dados sintéticos relativos às equações de movimento de uma partícula. 100.000 pontos foram fornecidos para os competidores, sendo pedidos os 500 seguintes. A figura 3.1 ilustra seu comportamento.

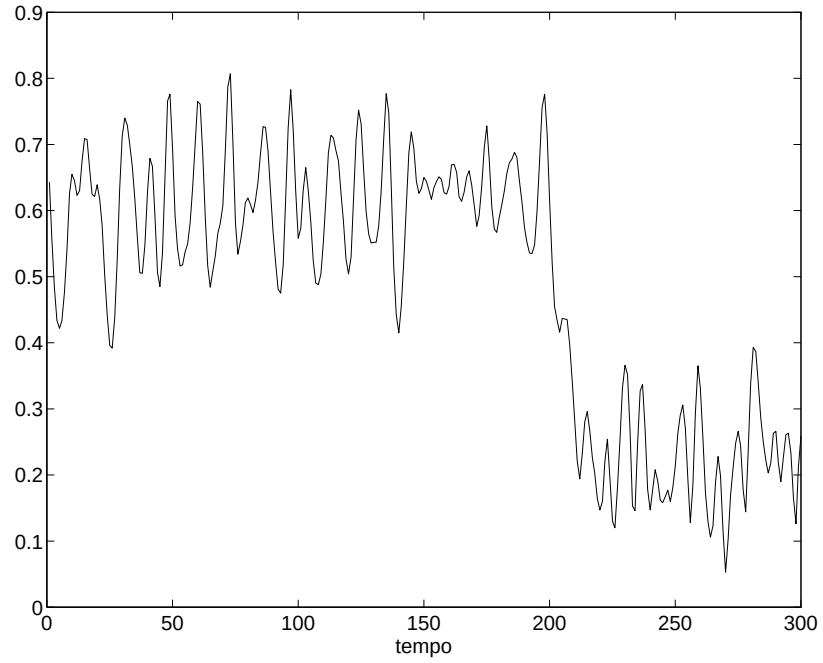

Figura 3.1: Fragmento da série D utilizada nos experimentos.

PUC-Rio - Certificação Digital Nº 9916205/CA