UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

CIÊNCIA DA COMPUTAÇÃO

FERNANDA SILVA EUSTÁQUIO

Um estudo sobre índices de validação de agrupamento fuzzy para dados de alta dimensionalidade

TRABALHO DE CONCLUSÃO DE CURSO

Salvador

2017

FERNANDA SILVA EUSTÁQUIO

Um estudo sobre índices de validação de agrupamento fuzzy para dados de alta dimensionalidade

Trabalho de Conclusão de Curso apresentado ao Departamento de Ciência da Computação, como parte dos requisitos necessários à obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Tatiane Nogueira Rios

Salvador
2017

Modelo de ficha catalográfica fornecido pelo Sistema Universitário de Bibliotecas da UFBA para ser confeccionada pelo autor

Eustáquio, Fernanda Silva
Um estudo sobre índices de validação de agrupamento fuzzy para dados de alta dimensionalidade / Fernanda Silva Eustáquio. -- Salvador, 2017.
265 f.

Orientadora: Tatiane Nogueira Rios.
TCC (Graduação - Ciência da Computação) -- Universidade Federal da Bahia, Departamento de Ciência da Computação, 2017.

1. Índices de validação fuzzy. 2. Agrupamento fuzzy. 3. Organização flexível de documentos. I. Rios, Tatiane Nogueira. II. Título.

RESUMO

Agrupar objetos em clusters constitui uma das tarefas não supervisionadas de aprendizagem de máquina. Nesta tarefa, os objetos são agrupados convencionalmente em um número *c* de clusters previamente informado, onde um objeto pertence a somente um dos clusters. No entanto, se o problema de aprendizagem está inserido em um contexto onde deseja-se que um objeto possa pertencer a mais de um cluster ao mesmo tempo, então a Teoria dos Conjuntos Fuzzy pode ser utilizada para que o agrupamento torne-se flexível. Esta flexibilidade em agrupar objetos será dada por um grau de pertinência que cada objeto terá em cada um dos *c* clusters.

Este número de clusters deve ser informado como parâmetro do algoritmo Fuzzy C-Means (FCM), utilizado neste trabalho para agrupar bases textuais de alta dimensionalidade. No entanto, se este valor não é conhecido e deseja-se saber qual o número ótimo de clusters que mais se ajusta a base agrupada, o algoritmo deverá ser executado para cada número *c* de clusters definido em um intervalo. Considerando que serão geradas diferentes partições para cada um dos valores de *c* utilizados, como escolher qual a melhor partição? Ou seja, como escolher a partição gerada que encontrou a melhor estrutura contida em uma base? A validação de agrupamento é então realizada para verificar se a partição gerada por um algoritmo é bem estruturada e qual o número de clusters desta partição. Para bases de dados bidimensionais, esta validação pode ser realizada através da visualização dos dados. Já no caso das bases de maiores dimensões esta validação será feita através de índices estatísticos, que quantificam a qualidade do agrupamento obtido, identificando a estrutura mais adequada de acordo com os dados e o problema. Estes índices estatísticos são chamados de índices de validação de agrupamento fuzzy.

Neste trabalho, os índices de validação serão avaliados ao estudar como estes índices validaram os agrupamentos gerados pelo FCM, ou seja, a partir da quantidade de clusters indicada por cada índice, estes terão seus desempenhos avaliados através do cálculo de média para quando o número ótimo de clusters escolhido for igual a quantidade correta, mínima ou máxima (intervalo de *c* informado ao FCM) de clusters. Estes valores de média permitiram: perceber que a alta dimensionalidade das bases textuais pode ser a principal causa do desempenho inesperado dos índices e sugerir os índices P, MPC, SF, K e T para validação de agrupamento de bases textuais, utilizando valores no intervalo de [7.0; 10.0] para o fator de fuzzificação *m.*

ABSTRACT

Clustering objects in clusters is one of the unsupervised tasks of machine learning. In this task, the objects are conventionally clustered into a number c of clusters previously informed, where an object belongs to only one of the clusters. However, if the learning problem is embedded in a context where it is desired that an object can belong to more than one cluster at a time, then Fuzzy Set Theory can be used to make flexible clustering. This flexibility in clustering objects will be given by a membership degree that each object will have in each clusters.

This number of clusters must be informed as a parameter of the Fuzzy C-Means (FCM) algorithm used in this work to clustering high dimensional textual sets. However, if this value is not known and we want to know what the optimum number of clusters that fit the clustered dataset, the algorithm should be executed for each number of clusters defined in a range. Considering that different partitions will be generated for each of the values of c used, how to choose the best partition? That is, how to choose the generated partition that found the best structure contained in a dataset? The validity clustering is performed to verify that the partition generated by an algorithm is well structured and how many clusters this partition has. For two-dimensional data sets, this validation can be performed by visualizing the data. In case of larger data sets, this validation will be done through statistical indexes, which quantify the quality of the clustering obtained, identifying the most appropriate structure according to the data and the problem. These statistical indexes are called fuzzy clustering validity indexes.

In this work, the validity indexes will be evaluated by studying how these indexes validated the clusters generated by FCM, that is, from the number of clusters indicated by each index, these will have their performances evaluated through the average calculation for when the optimal number of clusters chosen was equal to the correct, minimum or maximum amount (range of c given to FCM) of clusters. These average values allowed: to realize that the high-dimensionality of the textual set can be the main reason of the unexpected performance of the indexes and to suggest P, MPC, SF, K and T indexes for validate high dimensional data clustering using values in the range of [7.0; 10.0] for the fuzzification factor m.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Era da informação, da *big data*, das redes sociais, dos dispositivos móveis, da tecnologia das coisas. Estes são alguns adjetivos que podem ser atribuídos à atualidade, onde se é gerado um grande volume de dados, em tempo real, e formatos distintos, advindos de diversas fontes como interações em redes sociais, pesquisas em buscador web, aplicativos de monitoramento de saúde e serviços de geolocalização. Todos estes dados são armazenados e, quando inseridos em um contexto, tornam-se uma informação. A partir desta informação, se foi possível extrair alguma utilidade, então um conhecimento terá sido gerado.

Estes dados podem ser estruturados, organizados em tabelas de bancos de dados, e não estruturados, como documentos textuais, e-mail, vídeo e áudio. Quando os dados não são estruturados, a tarefa de extração de informação torna-se mais difícil, sendo então um desafio dos Sistemas de Recuperação de Informação (SRIs) conseguirem extrair a informação desejada pelo usuário ao utilizar qualquer motor de busca. Para isso, é desejável que os documentos estejam organizados automaticamente e de maneira flexível, já que o critério de relevância de um documento para um usuário é relativo.

A flexibilidade desejada nos SRIs pode ser obtida ao incorporar a Teoria de Conjuntos Fuzzy ao sistema de recuperação. Esta teoria irá permitir que um mesmo documento esteja presente em mais de um grupo (cluster) quando o conjunto de documentos for organizado. Esta organização flexível, automatizada e inteligente constitui a etapa de Extração de Padrões da Mineração de Dados que utiliza algoritmos de Aprendizagem de Máquina para agrupar os documentos. A Mineração de Dados é um processo exploratório de grande volume de dados com o intuito de descobrir padrões e tendências que podem auxiliar até mesmo a tomada de decisões. Quando os dados explorados neste processo consistem de documentos este será denominado Mineração de Texto.

Este processo é constituído de três etapas: Pré-processamento; Extração de Padrões e Pós-processamento. No pré-processamento os documentos são estruturados; na fase de extração de padrões, os algoritmos de Aprendizagem de Máquina são executados e, no pós-processamento, serão utilizadas diferentes métricas para avaliar os resultados gerados na etapa anterior. Esta última etapa, além de indicar o número ótimo de clusters de uma coleção (CARVALHO et al., 2016) (RIOS; REZENDE; CAMARGO, 2015) também é utilizada quando se deseja avaliar diferentes técnicas de pré-processamento (LIMA; EUSTÁQUIO; RIOS, 2017, No prelo), algoritmos (CARVALHO et al., 2016) e valores de seus parâmetros (PAL; BEZDEK, 1995) a fim de indicar qual técnica, algoritmo ou valor de parâmetro melhor se adequariam aos dados. Portanto, pode-se afirmar que as etapas de pré-processamento e extração de padrões podem influenciar as avaliações executadas no pós-processamento, assim como o mesmo também pode influenciar na escolha de uma técnica de pré-processamento, ou na escolha de um algoritmo e seus valores

de parâmetros, ou até mesmo na escolha entre métricas de validação em detrimento de outras (EUSTÁQUIO et al., 2017, No prelo).

Este estudo foi então elaborado para a escolha de valores de parâmetros e métricas de validação em detrimento de outras.

Neste trabalho, especificamente, foram executadas as fases de Extração de Padrões, onde as bases textuais são agrupadas de maneira flexível, utilizando o algoritmo Fuzzy C-Means, e a etapa de Pós-processamento, que avalia os agrupamentos gerados através da execução de índices de validação de agrupamento fuzzy. Como as bases utilizadas já foram obtidas pré-processadas, esta etapa não foi executada.

A validação de agrupamento é realizada para verificar se a partição gerada por um algoritmo é bem estruturada e qual o número de clusters desta partição. Para bases de dados bidimensionais, esta validação pode ser realizada através da visualização dos dados. Já no caso das bases de maiores dimensões esta validação será feita através de índices estatísticos, que quantificam a qualidade do agrupamento obtido, identificando a estrutura mais adequada de acordo com os dados e o problema (WANG; ZHANG, 2007).

Portanto, o processo de Mineração de Texto foi executado com o intuito de avaliar diferentes métricas utilizadas no pós-processamento, variando os valores dos parâmetros do algoritmo de agrupamento fuzzy ao agrupar bases textuais de alta dimensionalidade. Com isso, deseja-se saber: qual ou quais métricas seriam melhor indicadas para validar bases de alta dimensionalidade; os valores dos parâmetros do algoritmo utilizado; e se a alta dimensionalidade das bases é o principal fator prejudicial no desempenho dos índices.

Este trabalho está estruturado da seguinte maneira:

Capítulo 2: neste capítulo são apresentados o algoritmo de agrupamento fuzzy, Fuzzy C-Means, utilizado nos experimentos (Seção 2.1) com seus respectivos parâmetros e os índices de validação utilizados com suas definições e características (Seção 2.2).

Capítulo 3: neste capítulo, os trabalhos relacionados que executam diferentes índices de validação de agrupamento fuzzy para compará-los com a proposta de um novo índice ou para fazer uma revisão dos mesmos são apresentados.

Capítulo 4: neste capítulo são apresentados os materiais, bases textuais de alta dimensionalidade (Seção 4.1), e métodos utilizados, algoritmo Fuzzy C-Means com seus respectivos valores de parâmetros (Seção 4.3), além de ser apresentada a motivação (Seção 4.2) para este estudo e os questionamentos que o mesmo pretende responder.

Capítulo 5: são apresentados os resultados obtidos pelos índices ao validarem os agrupamentos gerados para os diferentes valores de parâmetros testados (Seção 5.1), a discussão acerca destes resultados (Seção 5.2) e como estes foram analisados (Seções 5.2.1, 5.2.2 e 5.2.3) para responder aos questionamentos levantados no capítulo anterior.

Capítulo 6: neste último capítulo são revistas as respostas, discutidas no Capítulo 5, além de enumerar as razões pelas quais um índice de validação pode não conseguir reconhecer a estrutura presente nos dados.

2 AGRUPAMENTO FUZZY

Para organizar documentos de maneira flexível é necessário utilizar algoritmos de agrupamento de Aprendizagem de Máquina. Dentre os algorimos de agrupamento fuzzy, o Fuzzy C-Means foi escolhido neste trabalho por, além de ser bem conhecido e utilizado na literatura, também ser muito utilizado como modelo para a criação de diversos índices de validação, que frequentemente utilizam conceitos e até mesmo funções do algoritmo em suas definições. O Fuzzy C-Means e detalhes de seu funcionamento são apresentados a seguir.

2.1 Fuzzy C-Means

O Fuzzy C-Means (FCM) (BEZDEK, 1981) é um algoritmo particional que deseja agrupar os objetos em clusters a partir de suas semelhanças ou similaridades. O FCM, em cada iteração, busca encontrar o ponto mais característico de cada cluster, denominado centroide, e em seguida calcula o grau de pertinência de cada objeto (documento) nos clusters. O grau de pertinência de um objeto em cada cluster representa o quão próximo este objeto está do centroide deste cluster. Esta busca pelo ponto mais característico de um cluster é executada ao minimizar a função objetivo do algoritmo definida como:

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} A_i^m(d_k)\,|d_k - v_i\,|^2. \tag{2.1}$$

Na Equação 2.1: d_k é um documento; n é o número de documentos da coleção; c é o número de clusters informado pelo usuário; v_i é o centroide do cluster i; $A_i(d_k)$ é o grau de pertinência do documento d_k no cluster i; e o fator de fuzzificação m é um número real, $1 < m < \infty$, que define o quão fuzzy as partições geradas pelo FCM serão. Valores de m próximos de 1 indicam partições fuzzy mais próximas de uma partição crisp e, a medida que este fator aumenta, as partições se tornam mais fuzzy, ou seja, se tornam mais distantes de partições crisp. Muitos usuários do FCM escolhem valores de m no intervalo (1; 10], mas valores de $m < 5$ são mais usuais. No entanto, $m = 2$ é a escolha mais comum na literatura (PAL; BEZDEK, 1995).

Um documento d_k é definido como:

Definição 2.1: Considere uma coleção de documentos D = $\{d_1, d_2, \ldots, d_n\}$ e seja T o número de termos nesta coleção. Uma matriz documento-termo W = $[d_{kj}]$ é composta por um documento d_k em cada linha, onde cada coluna corresponde a um termo t_j, j = 1, ..., T. Um documento d_k é representado por um vetor $[d_{k1}, d_{k2}, \ldots, d_{kT}]$, $1 \leq k \leq n$.

Cada execução do FCM produz uma partição fuzzy definida como:

Definição 2.2: Seja c o número de clusters e $A_i(d_k)$ o grau de pertinência do documento d_K no cluster i, $k = 1, \ldots, n$, $i = 1, \ldots, c$. Uma pseudo-partição fuzzy $U = [A_i(d_k)]$ é uma família de conjuntos fuzzy de documentos D (Definição 2.1) denodado por $P = \{A_1, A_2, \ldots, A_c\}$, que satisfaz as Equações 2.2 e 2.3.

$$\sum_{i=1}^{c} A_i(d_k) = 1 \tag{2.2}$$

$$0 < \sum_{k=1}^{n} A_i(d_k) < n \tag{2.3}$$

Como no FCM os objetos são agrupados a partir de suas semelhanças, aqueles que forem mais similares entre si serão agrupados em um mesmo cluster. Já os mais diferentes, ou seja, com maior dissimilaridade, deverão ser agrupados em clusters distintos. Esta medida de dissimilaridade é geralmente calculada através da Distância Euclidiana. Como as bases utilizadas neste estudo são bases textuais de alta dimensionalidade e esparsas, a Distância Euclidiana não é indicada porque pode considerar as comparações do tipo zero-zero (que não agregam informação) e alguns documentos serão indicados como similares pelo número de termos que ambos possuem. Por esse motivo, a dissimilaridade entre d_k e v_i ($\|d_k - v_i\|$) será medida ao transformar o coeficiente de similaridade do Coseno em uma medida de dissimilaridade (RIOS; REZENDE; CAMARGO, 2015) como definido nas Equações 2.4 e 2.5:

$$sim(d_k, v_i) = cos\theta = \frac{d_k . v_i}{|d_k||v_i|} \in [0;1] \tag{2.4}$$

$$|d_k - v_i| = 1 - sim(d_k, v_i) \in [0;1] \tag{2.5}$$

Durante o agrupamento, um processo iterativo de otimização é realizado atualizando os centroides dos clusters $V = \{v_1, v_2, \ldots, v_c\}$ e os graus de pertinência da matriz da pseudo-partição fuzzy U, definidos respectivamente pelas Equações 2.6 e 2.7. Esta atualização é realizada até o critério de parada ser satisfeito.

$$v_i = \frac{\sum_{k=1}^{n} [A_i(d_k)]^m d_k}{\sum_{k=1}^{n} [A_i(d_k)]^m}, 1 \leq i \leq c. \tag{2.6}$$

$$\mu_{ik} = A_i(d_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{|d_k - v_i|}{|d_k - v_k|}\right)^{\frac{1}{m-1}}} \tag{2.7}$$

O FCM possui dois parâmetros que podem ser utilizados como critério de parada. São eles: o número de iterações que o algoritmo irá executar, representado pela letra *T*, e o valor do erro ε que, quando satisfeito, significa que a escolha dos centroides foi estabilizada e os mesmos realmente são os pontos mais característicos dos clusters.

O algoritmo FCM é executado nos seguintes passos:

Passo 1: Dado o número de clusters *c*, o valor de *m*, valor do erro ε e/ou o número máximo de iterações *T*, a matriz de pertinência *U* é inicializada;

Passo 2: Calcula os centroides dos clusters para $i = 1, \ldots, c$, usando a Equação 2.6;

Passo 3: Atualiza os valores dos graus de pertinência $A_i(d_k)$, usando a Equação 2.7;

Passo 4: Se a melhora em relação a $J_m\ (U, V)$ é menor do que o erro ε estabelecido ou o número máximo de iterações *T* foi satisfeito, então o algoritmo para, caso contrário retorna ao passo 2.

O número de clusters em que os objetos de uma base de dados devem ser agrupados é informado pelo usuário como parâmetro de entrada do FCM. No entanto, por se tratar de uma tarefa não-supervisionada, o número *c* de clusters muitas vezes é desconhecido. Por este motivo, quando se deseja descobrir o número ótimo de clusters que uma determinada base pode ser melhor oganizada, deve ser informado ao FCM, um intervalo com a quantidade miníma e máxima de clusters que se deseja agrupar os dados. Este intervalo deve ter no mínimo dois e no máximo $(n - 1)$ clusters ($2 \leq c \leq n - 1$). Isto porque se $c = 1$, somente um cluster existirá na partição o que corresponde a base de dados original, e se $c = n$, cada objeto terá o seu próprio cluster, o que também não é desejável. A partir deste intervalo, o FCM será executado para cada valor de *c* e os índices de validação de agrupamento fuzzy irão indicar o número ótimo de clusters que melhor oganizou os dados.

Os índices de validação que serão utilizados, bem como suas definições e características, serão apresentados na próxima seção.

2.2 Índices de Validação de Agrupamento Fuzzy

Os índices de validação de agrupamento fuzzy são índices estatísticos que quantificam a qualidade de uma partição, informando o número ótimo de clusters que uma base de dados pôde ser organizada. Os índices de validação são classificados em três critérios que expressam a estratégia utilizada para validar um agrupamento: interno, externo e relativo. No critério externo, os índices avaliam um agrupamento de acordo com a estrutura estabelecida previamente por um especialista. Esta estrutura pode ser chamada de partição real ou estrutura conhecida. No

critério interno, os índices medem a qualidade da estrutura gerada a partir de informações do próprio conjunto de dados. Os critérios externo e interno são baseados em testes estatísticos e possuem um alto custo computacional. Estes critérios têm como objetivo medir o quanto o resultado obtido confirma uma hipótese pré-especificada, utilizando um teste de hipótese que depende da distribuição do índice sob uma hipótese nula H_o. A metodologia destes critérios são mostradas nas Figuras 1 e 2 (GAMA et al., 2011).

Figura 1 – Critério Externo (GAMA et al., 2011)

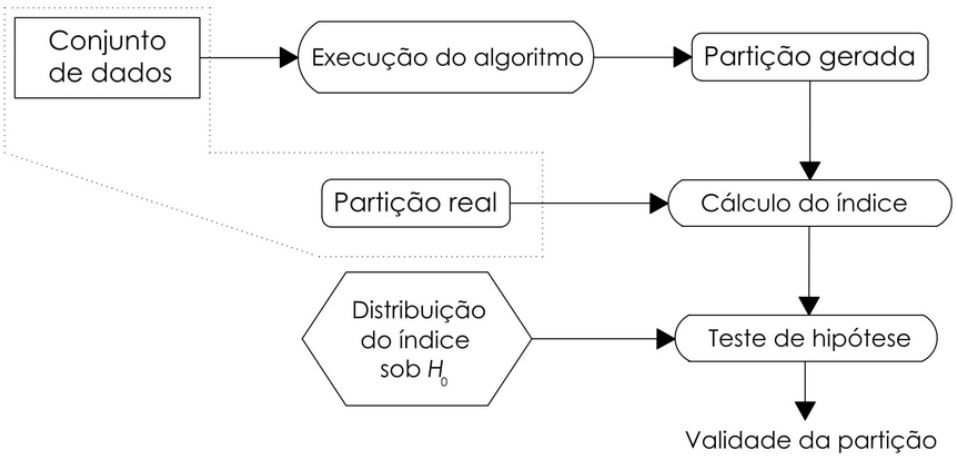
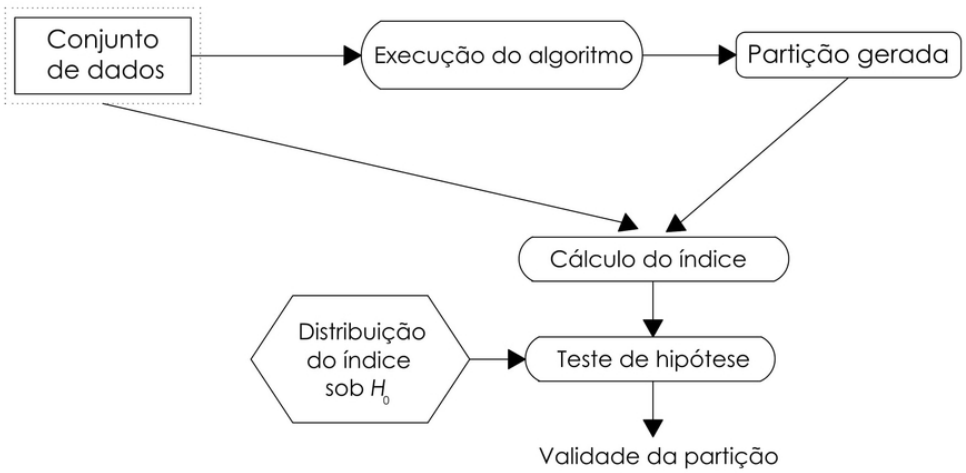
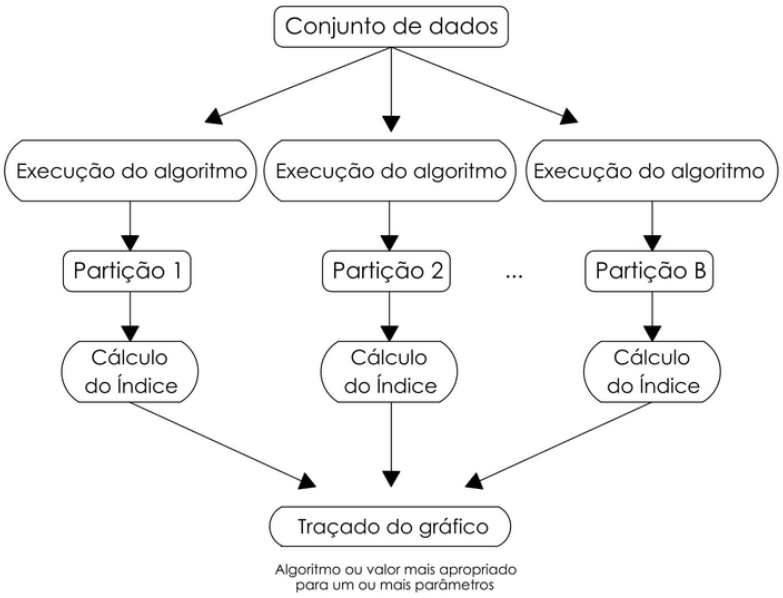


Figura 2 – Critério Interno (GAMA et al., 2011)



No critério relativo, os índices comparam diversos agrupamentos para decidir qual deles é o mais adequado aos dados. Este critério pode ser usado para comparar diversos algoritmos de agrupamento, índices de validação e valores de parâmetros dos algoritmos. Neste trabalho, um algoritmo será executado com diferentes valores de parâmetros, para diferentes conjuntos de dados e as partições geradas serão avaliadas por 16 índices de validação. Na Figura 3, pode ser verificada a metodologia do critério relativo. Neste estudo, para cada partição, com número distinto de clusters, um valor de índice é calculado e seus valores serão comparados, onde o melhor agrupamento será determinado pelo valor máximo (índices de maximização) ou mínimo (índices de minimização) calculado pelo índice.

Figura 3 – Critério Relativo (GAMA et al., 2011)



Neste trabalho, foram utilizados os índices internos com o critério relativo de avaliação porque deseja-se verificar se, o número ótimo de clusters, escolhido por um índice, corresponde ao número indicado por um especialista, e determinar o valor ou intervalo de valores para o fator de fuzzicação m ao agrupar bases de alta dimensionalidade. Além dos critérios de validação, os índices também podem ser classificados em duas classes de acordo com a informação que os mesmos utilizam para quantificar a qualidade das partições avaliadas. Estas duas classes correspondem aos índices que utilizam somente a matriz de pertinência ou índices que utilizam o conjunto de dados além da matriz dos graus de pertinência dos documentos. As definições das classes e seus respectivos representantes são definidos e apresentados nas seções seguintes.

2.2.1 Índices que utilizam somente a matriz de pertinência

Os índices que utilizam somente a matriz de pertinência da pseudo-partição U para o cálculo de seus valores são considerados distantes das informações inerentes aos dados, como

os valores de seus atributos ou a distância entre os objetos. Desta classe foram utilizados os seguintes índices, definidos nas próximas seções: Coeficiente da Partição; Entropia da Partição; Coeficiente da Partição Modificado; índice de Kim; índice de Chen e Linkens; índice robusto de Yang e o índice Graded Distance.

Nas definições dos índices apresentados nas seções seguintes: c é o número de clusters da pseudo-partição U; n é o número de objetos (documentos) nas bases e $A_i(d_k)$ corresponde ao grau de pertinência do documento d_k no cluster i.

2.2.1.1 Coeficiente da Partição (PC)

O PC (BEZDEK, 1974) calcula a média relativa de interseção fuzzy entre pares de subconjuntos fuzzy na pseudo-partição U. O PC é um índice de maximização e, portanto, uma boa partição será aquela de maior valor. Nesta partição, os graus de pertinência dos objetos (documentos) estão mais próximos de 1, indicando assim que os objetos estão próximos de seus centroides. O PC pode assumir valores entre [1/c; 1] e é definido como:

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(d_k). \tag{2.8}$$

2.2.1.2 Entropia da Partição (PE)

O PE (BEZDEK†, 1974) é um índice de minimização que mede o montante de fuzzificação em uma pseudo-partição U e pode assumir valores entre $[0; \log_a c]$. Como o valor da base do logarítmo não é definido na literatura, foi então utilizado o logarítmo decimal ($a = 10$). O PE é definido como:

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i(d_k)\log_a(A_i(d_k)). \tag{2.9}$$

Os índices PC e PE, definidos anteriormente, medem essencialmente a distância que uma pseudo-partição U está de ser crisp, isto é, eles medem a fuzzificação em U (PAL; BEZDEK, 1995). Ambos têm a desvantagem da monotonicidade, onde a medida que c cresce o valor do PC decresce e o oposto ocorre com o PE.

2.2.1.3 Coeficiente da Partição Modificado (MPC)

O Coeficiente da Partição Modificado (DAVE, 1996) é um índice de maximização e foi proposto para corrigir a tendência monotônica do PC. Este pode assumir valores entre [0; 1] e é definido como:

$$MPC = 1 - \frac{c}{c-1}(1 - PC). \tag{2.10}$$

2.2.1.4 Índice de Kim (KYI)

O KYI (KIM et al., 2004) é um índice de minimização que define o grau relativo de compartilhamento de dois clusters como a soma ponderada destes graus para todos os dados. Quanto menos sobreposição existir na pseudo-partição fuzzy U e menos vago os objetos estiverem nestas sobreposições, menor será o valor do índice. O KYI é definido como:

$$KYI = \frac{2}{c(c-1)} \sum_{p \neq q}^{c} \sum_{k=1}^{n} [c \times min(A_{F_p}(d_k), A_{F_q}(d_k)) \times h(d_k)], \tag{2.11}$$

$$h(d_k) = -\sum_{i=1}^{c} A_{F_i(d_k)} \log_a A_{F_i(d_k)},$$

onde F_p e F_q são dois clusters pertencentes à pseudo-partição U.

2.2.1.5 Índice de Chen e Linkens (P)

O índice de Chen e Linkens (P) (CHEN; LINKENS, 2004) é um índice de maximização composto por dois termos: o primeiro termo mede a compacidade dentro de um cluster e o segundo mede a separação entre os clusters da partição. O índice P é definido como:

$$P = \frac{1}{n} \sum_{k=1}^{n} max_i(A_i(d_k)) - \frac{1}{K} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \left[\frac{1}{n} \sum_{k=1}^{n} min(A_i(d_k), A_j(d_k)) \right], \tag{2.12}$$

onde $K = \sum_{i=1}^{c-1} i$. No primeiro termo, quanto mais próximo um documento d_k está de um centroide, mais próximo de 1 estará o grau máximo de pertinência $max_i(A_i(d_k))$. Portanto,

um valor alto do primeiro termo indica que os objetos similares estão próximos um dos outros e os clusters formados são compactos. No segundo termo, a interseção entre dois clusters é utilizada para avaliar a separação entre os mesmos (*i, j*). Se d_k é próximo do centroide do cluster i, $min(A_i(d_k), A_j(d_k))$ será mais próximo de 0 e, consequentemente, os clusters i e j estarão bem separados (WANG; ZHANG, 2007).

2.2.1.6 Índice robusto de Yang (MPO)

O MPO (HU et al., 2011) é um índice de maximização que visa diminuir a influência de ruídos e *outliers* que podem reduzir a performance dos índices ao avaliar as partições geradas pelo FCM. O MPO é definido pela diferença entre a compactação (Equação 2.14) e a separação (Equação 2.16) dos clusters. Quanto maior for a compactação dentro dos clusters e menor a separação entre os mesmos, melhor avaliada será a partição fuzzy. Na Equação 2.17 foi utilizado o valor de limiar T_o = 1/c.

$$MPO(U,c) = MPC(U,c) - Sep(U,c), \tag{2.13}$$

$$MPC(U,c) = \left(\frac{c+1}{c-1}\right)^{1/2} \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} {A_i}^2(d_k)}{A_M}, \tag{2.14}$$

onde

$$A_M = min_{1\leq i\leq c}\left\{\sum_{k=1}^{n} {A_i}^2(d_k)\right\}. \tag{2.15}$$

$$Sep(U,c) = \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} O_{ijk}(c,U)\right), \tag{2.16}$$

onde

$$O_{ijk}(c,U) = \begin{cases} 1 - |A_i(d_k) - A_j(d_k)| & |A_i(d_k) - A_j(d_k)| \geq T_o, i \neq j \\ 0 & |A_i(d_k) - A_j(d_k)| < T_o \end{cases} \tag{2.17}$$

A compactação foi definida pelos autores como uma modificação do índice MPC definido anteriormente na Seção 2.2.1.3. Na Equação 2.14, o termo A_M é utilizado para lidar com diferentes escalas e o termo $(c+1/c-1)^{1/2}$ é utilizado para ajustar o valor do MPC para que possa evitar a tendência monotônica para o número de clusters. Na Equação 2.16, o termo O_{ijk} pode diminuir a influência de ruídos e *outliers* porque esta equação utiliza o limiar T_o para remover os objetos dispersos nas fronteiras dos clusters. Na Equação 2.17, para o documento d_k e dois clusters i e j, O_{ijk} *(c, U)* indica o grau de separação de d_k destes dois clusters. O segundo termo do MPO (Equação 2.16) é a soma dos graus de separação de todos os clusters para todos os documentos.

2.2.1.7 Índice Graded Distance (GD)

O GD (JOOPUDI et al., 2013) é um índice de minimização que calcula a média da diferença entre o primeiro e segundo maiores graus de pertinência de todos os documentos. O GD é definido como:

$$GD_{index,c} = \frac{\sum_{k=1}^{n} A_1(d_k) - A_2(d_k)}{n} - \left(\frac{c}{n}\right). \tag{2.18}$$

O índice GD obtém o número ótimo de clusters: maximizando a força de associação de todos os objetos em relação aos seus respectivos maiores graus de pertinência em um cluster e minimizando a sobreposição entre os mesmos.

2.2.2 Índices que utilizam a matriz de pertinência e o conjunto de dados

Os índices definidos na Seção 2.2.1.1, por utilizarem somente informações da pseudo-partição *U,* podem apresentar as seguintes desvantagens (WANG; ZHANG, 2007):

1) Constante dependência do número de clusters c;
2) Sensibilidade ao fator de fuzzificação m;
3) Falta de conexão entre a geometria dos dados.

Devido as desvantagens listadas acima, também foram executados, neste estudo, os índices de validação que utilizam as informações das bases de dados além da matriz de pertinência. Estes índices estão definidos nas próximas seções e são eles: o índice de Fukuyama-Sugeno, Fuzzy Hypervolume Validity, índice de Xie-Beni, índice de Kwon, índice de Zahid, índice de Pakhira fuzzificado, índice de Wu e Yang, índice de Tang e Silhueta Fuzzy.

2.2.2.1 Índice de Fukuyama-Sugeno (FS)

O FS (FUKUYAMA; SUGENO, 1989) é um índice de minimização que mede a diferença entre a compactação, dada pela distância intra-cluster (primeiro termo), e a separação entre clusters, dada pela distância inter-cluster (segundo termo). Este índice é definido como:

$$FS = \sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(d_k)\,|d_k - v_i\,|^2 - \sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(d_k)\,|v_i - \bar{v}\,|^2, \tag{2.19}$$

onde v_i é o centroide do cluster *i,* definido na Equação 2.6 da Seção 2.1, e $\bar{v} = \sum_{i=1}^{c} v_i/c$ corresponde a média dos centroides dos c clusters da pseudo-partição U.

2.2.2.2 Fuzzy Hypervolume Validity (FHV)

O FHV (GATH; GEVA, 1989) é um índice de minimização baseado nos conceitos de hipervolume e densidade definido como:

$$FHV = \sum_{i=1}^{c}[det(F_i)]^{1/2}, \tag{2.20}$$

$$F_i = \frac{\sum\limits_{k=1}^{n} A_i^m(d_k)(d_k - v_i)(d_k - v_i)^T}{\sum\limits_{k=1}^{n} A_i^m(d_k)}, \tag{2.21}$$

onde F_i é a matriz de covariância fuzzy do cluster *i*.

2.2.2.3 Índice de Xie-Beni (XB)

O XB é um índice de minimização que mede a compactação (numerador) e separação (denominador) dos clusters. Originalmente, o XB foi definido com valor de m fixo em 2 (XB_2) (XIE; BENI, 1991) e em (PAL; BEZDEK, 1995) foi generalizado para qualquer valor de m. Um menor valor de XB indica que os clusters são mais compactos e separados. Este índice é definido como:

$$XB = \frac{\sum\limits_{i=1}^{c}\sum\limits_{k=1}^{n} A_i^m(d_k)\,|d_k - v_i\,|^2}{n \times min_{k\neq i}\,|v_i - v_k\,|^2}. \tag{2.22}$$

Na Equação 2.22, o numerador mede a compactação da partição utilizando a função objetivo J_m (Equação 2.1, Seção 2.1) do FCM e a separação é medida no denominador pela distância mínima entre os centroides dos clusters.

2.2.2.4 Índice de Kwon (K)

O índice K (KWON, 1998) foi proposto para eliminar a tendência monotônica do XB de decrescer quando o número de clusters aproxima-se da quantidade de documentos. Este é um índice de minimização definido como:

$$K = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} A_i^2(d_k)\,|d_k - v_i\,|^2 + \frac{1}{c}\sum_{i=1}^{c}|v_i - \bar{v}\,|^2}{min_{i\neq j}\,|v_i - v_j\,|^2}, \tag{2.23}$$

onde $\bar{v} = \sum_{k=1}^{n} d_k/n$ corresponde a média geral de todos os documentos d_k pertencentes a coleção agrupada. Na Equação 2.23, o primeiro termo do numerador mede a similaridade dentro de um cluster, ou seja, a distância intra-cluster. Quanto mais compactos forem os clusters da partição, menor será o valor deste termo. O segundo termo do numerador é a função adicionada ao XB que tenta eliminar a tendência de decrescimento em função de c. O denominador mede a separação entre os clusters, ou seja, a distância inter-cluster e quanto maior o valor desta distância, mais bem separados estarão os clusters.

2.2.2.5 Índice de Zahid (SC)

SC (ZAHID; LIMOURI; ESSAID, 1999) é um índice de maximização que combina além dos conceitos de compactação e separação (SC_1), também utiliza a união e interseção fuzzy para obter o grau fuzzy de compactação e separação (SC_2). O índice é definido como:

$$SC = SC_1(c) - SC_2(c), \tag{2.24}$$

onde

$$SC_1(c) = \frac{\sum_{i=1}^{c}|v_i - \bar{v}\,|^2/c}{\sum_{i=1}^{c}(\sum_{k=1}^{n} A_i^m(d_k)\,|d_k - v_i\,|^2/\sum_{k=1}^{n} A_i(d_k))} \tag{2.25}$$

e

$$SC_2(c) = \frac{\sum_{i=1}^{c}\sum_{l=i+1}^{n}(\sum_{k=1}^{n}(min(A_i(d_k),A_l(d_k)))^2 / \sum_{k=1}^{n} min(A_i(d_k),A_l(d_k)))}{\sum_{k=1}^{n}(max_{1\leq i\leq c}A_i(d_k))^2 / \sum_{k=1}^{n} max_{1\leq i\leq c}A_i(d_k)} \tag{2.26}$$

2.2.2.6 Índice de Pakhira fuzzificado (PBMF)

O PBMF (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004) é um índice de maximização composto por três fatores: 1/c, E_1/J_m e D_c definidos como:

$$PBMF = \left(\frac{1}{c} \times \frac{E_1}{J_m} \times D_c\right)^2, \tag{2.27}$$

onde

$$E_1 = \sum_{k=1}^{n} |d_k - v|, \tag{2.28}$$

$$D_c = max_{i,j=1}^{c} |v_i - v_j|, \tag{2.29}$$

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(d_k)\,|d_k - v_i|. \tag{2.30}$$

Na Equação 2.28, $v = \sum_{k=1}^{n} d_k/n$ é a média geral de todos os documentos d_k pertencentes a coleção agrupada. O primeiro fator (1/c) indica a divisibilidade de uma partição de c clusters. O segundo fator (E_1/J_m, Equação 2.27) mede a compactação da pseudo-partição U onde E_1 é a soma ponderada das distâncias de todos os documentos a um "documento central" v e corresponde a um valor fixo utilizado somente para eliminar a chance do segundo fator se tornar muito pequeno. O terceiro fator (D_c, Equação 2.29) mede a separação máxima entre dois clusters a partir da distância entre seus centroides.

2.2.2.7 Índice de Wu e Yang (PCAES)

O PCAES (WU; YANG, 2005) é um índice de maximização onde seu alto valor pode ser usado para detectar a estrutura dos dados com clusters compactos e bem separados. Este índice pode apresentar boa performance em bases de dados com muitos ruídos. O PCAES é definido como:

$$PCAES = \sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(d_k)/A_M - \sum_{i=1}^{c} exp\left(-min_{j\neq i}|v_i - v_j|^2/B_T\right), \tag{2.31}$$

onde A_M é definido como na Equação 2.15 (Seção 2.2.1.6); $\bar{v} = \sum_{k=1}^{n} d_k/n$ corresponde a média de todos os documentos d_k pertencentes a coleção agrupada e $B_T = \sum_{i=1}^{c} |v_i - \bar{v}|^2/c$ é a média da distância entre os centroides de cada cluster (v_i) e a média de todos os documentos ($\bar{v}$).

2.2.2.8 Índice de Tang (T)

O índice T (TANG; SUN; SUN, 2005) é um índice de minimização que segue a mesma referência do índice K (Seção 2.2.2.4) em acrescentar uma função para eliminar a tendência monotônica do XB quando o número de clusters aproxima-se da quantidade de documentos. O índice é definido como:

$$T = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(d_k)\,|d_k - v_i|^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{j=1}^{c}{}_{j\neq i}\,|v_i - v_j|^2}{min_{i\neq j}|v_i - v_j|^2 + 1/c}. \tag{2.32}$$

Na Equação 2.32, os primeiro fatores do numerador e denominador medem, respectivamente, a distância intra-cluster e a menor distância inter-cluster. Já os segundos fatores do numerador e denominador correspondem a funções de punição respectivamente quando $c \to n$ e quando $m \to \infty$, sendo a última para reforçar a estabilidade númerica quando m se aproxima do infinito.

2.2.2.9 Silhueta Fuzzy (SF)

O índice Silhueta Fuzzy (CAMPELLO; HRUSCHKA, 2006) é uma extensão fuzzy do critério de largura média da silhueta (ASWC) (KAUFMAN; ROUSSEEUW, 2005). Este é um índice de maximização que considera os dois clusters em que um documento d_k tem os dois maiores graus de pertinência. O SF é definido como:

$$SF = \frac{\sum_{k=1}^{n} (A_1(d_k) - A_2(d_k))S(d_k)}{\sum_{k=1}^{n} (A_1(d_k) - A_2(d_k))}, \tag{2.33}$$

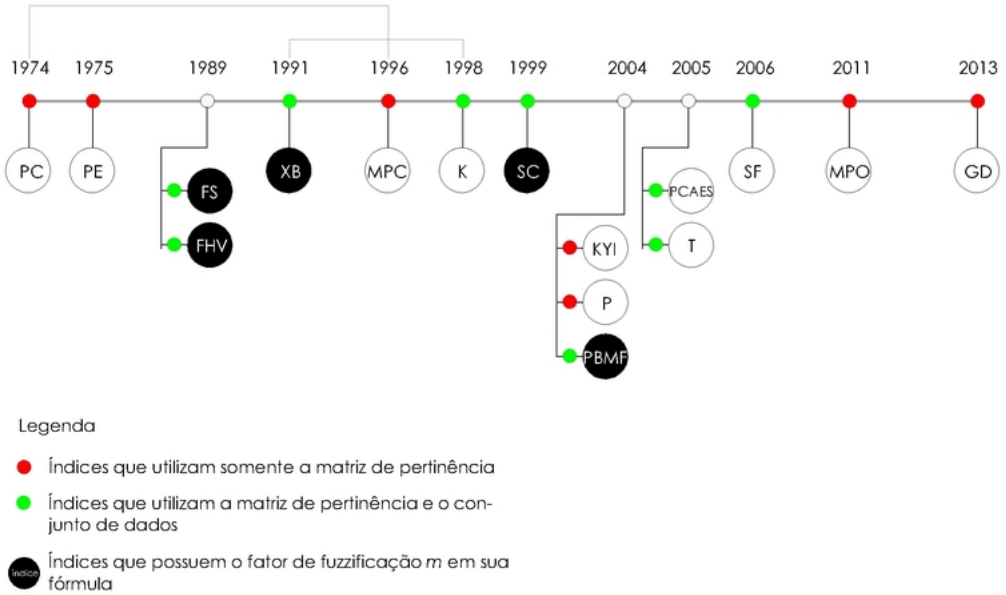
Na Equação 2.33, *S(d_k)* representa a silhueta do documento d_k, onde: d_k pertence ao cluster g_i, isto é, d_k possui maior grau de pertinência em g_i, com $g_i \in (g_1, g_2, ..., g_c)$; $\delta(d_k, g_i)$ é a média da distância entre d_k e todos os documentos pertencentes a g_i, isto é, a distância intra-cluster; $\beta(d_k, g_i)$ é a distância entre d_k e seu vizinho mais próximo a g_i, isto é, a distância inter-cluster. O denominador de *S(d_k)* é utilizado como um fator de normalização e quanto maior o valor de *S(d_k)*, mais o documento d_k é considerado pertencente ao cluster g_i (RIOS, 2013).

2.3 Considerações Finais

Nesta seção foram apresentados os três critérios de validação em que os índices podem ser classificados, sendo utilizados neste trabalho os índices de validação de agrupamento fuzzy do critério relativo. Estes índices são organizados pela informação que os mesmos utilizam para calcular os seus valores (índices que utilizam somente a matriz de pertinência e índices que utilizam a base original além da matriz de pertinência). Foram apresentados cada índice de validação com suas respectivas formulações que permitirão o cálculo de seus valores e, consequentemente, a escolha do número ótimo de clusters, seja pelo valor máximo (índices de maximização) ou mínimo (índices de minimização) de cada índice.

A Figura 4 mostra resumidamente a variedade dos índices utilizados neste trabalho e como estes estão organizados a partir de suas características. Foram utilizados 16 índices de validação de agrupamento fuzzy, dos mais tradicionais e conhecidos na literatura como o PC, PE, MPC, XB e FS, aos menos referenciados e mais recentes como o MPO e GD. A partir desta figura também é possível verificar que, apesar das desvantagens listadas na seção anterior, os índices que utilizam somente a matriz de pertinência continuam sendo propostos, como é o caso do MPO e GD.

Figura 4 – Linha cronológica dos índices de validação fuzzy



Na seção seguinte são apresentados os trabalhos relacionados, com seus respectivos agoritmos e índices de validação, bases de dados e valores de parâmetros utilizados, que motivaram e foram fontes de pesquisa para a investigação realizada neste trabalho.

3 TRABALHOS RELACIONADOS

Muitos artigos são constantemente publicados com revisões dos índices de validação de agrupamento fuzzy mais conhecidos na literatura ou com a proposta de um novo índice, sendo propriamente um novo índice, uma extensão ou até mesmo a versão fuzzy de um índice de validação de agrupamento crisp. Nas publicações em que um novo índice é proposto, este apresenta um desempenho superior aos demais índices ao acertar majoritariamente a quantidade de clusters das bases de dados utilizadas.

Dave propôs em (DAVE, 1996) o índice MPC a fim de eliminar a tendência monotônica apresentada pelo índice PC. O desempenho do índice proposto foi comparado com a de outros cinco índices, dentre eles o PC, PE, FHV. Quatro bases sintéticas bidimensionais foram utilizadas. O único valor de parâmetro do FCM informado foi o número de clusters que começou com dois e foi sendo incrementado em um até uma alteração no gradiente do índice ter sido medida.

Em (CAMPELLO; HRUSCHKA, 2006), o índice Silhueta Fuzzy foi proposto e sua validação é comparada a de outros cinco índices de validação, dentre eles o XB e FHV. Das seis bases utilizadas, cinco foram agrupadas pelo FCM. Destas cinco, quatro são bases sintéticas e bidimensionais e a última consiste da base Iris, sendo esta a de maior dimensão. O FCM foi executado com 10 iterações ($T = 10$) e os seguintes parâmetros foram utilizados: valor do fator de fuzzificação $m = 2.0$ e a quantidade c de clusters variando de dois até a quantidade de clusters das bases originais somado com três.

Em (RAWASHDEH; RALESCU, 2012) foi realizada uma revisão de alguns índices populares de validação de agrupamento fuzzy e uma generalização do SF foi proposta para adicionar duas matrizes com as distâncias intra e inter-cluster. Neste trabalho, dois exemplos são executados onde somente os parâmetros utilizados no FCM do segundo exemplo foram informados: o número de clusters c variou de 2 até 9 e o valor do fator de fuzzificação $m = 2.0$. A dimensão das bases utilizadas também não foi informada.

O índice MPO foi proposto em (HU et al., 2011) e sua validação foi comparada com a de outros nove índices, dentre eles o PC, PE, FS, XB, KYI e PCAES. Nesta publicação foram utilizadas seis bases de dados, sendo quatro sintéticas e duas reais, as bem conhecidas Iris e Wine. Os valores dos parâmetros do FCM utilizados não foram informados.

Em (JOOPUDI et al., 2013) o índice GD foi proposto e sua validação foi comparada com a de outros cinco índices, dentre eles o PC, PE, MPC e K. Neste trabalho, três bases, sendo duas sintéticas e a Iris, foram agrupadas pelo FCM variando a quantidade de clusters de 2 até 7 e com fator de fuzzificação $m = 2.0$.

Os índices de validação utilizados neste estudo foram desenvolvidos, em sua grande maioria, utilizando o FCM como modelo. No entanto, estes índices podem validar agrupamentos

gerados por outros algoritmos como Aprendizagem Participativa (AP) e Gustafson-Kessel (GK), sendo o último uma extensão do FCM que pode detectar clusters de diferentes orientações e formatos, assim como os índices inspirados em outros algoritmos também podem ser utilizados ao validar agrupamentos gerados pelo FCM. Este é o caso das publicações (ZHANG; QIAN, 2012) e (SILVA; GOMIDE, 2004). Na primeira, um novo índice de validação foi proposto e seu desempenho foi comparado com os índices PC, PE, XB, PBMF e KYI, onde: o algoritmo de agrupamento fuzzy utilizado foi o GK com $2 \leq c \leq 10$, $m = 2.0$ e $\varepsilon = 0.0001$; 10 bases foram agrupadas, sendo seis artificiais e quatro bem conhecidas da literatura como a Iris, Pima, Breas Cancer e Wine. Já em (SILVA; GOMIDE, 2004), os dois algoritmos AP e GK foram executados com $2 \leq c \leq 5$ ao agruparem quatro bases de dados, sendo a Iris a de maior dimensionalidade, sendo comparados os valores calculados pelos índices PC, PE, FS e XB.

Bezdek em (PAL; BEZDEK, 1995) faz uma revisão dos índices PC, PE, FS e XB a fim de testar a influência do fator de fuzzificação no comportamento dos mesmos. As bases Iris e Normal-4 foram utilizadas, tendo ambas quatro dimensões. Os valores dos parâmetros utilizados foram: número máximo de iterações $T = 100$; erro $\varepsilon = 0.00001$; $2 \leq c \leq 10$; $1.01 \leq m \leq 7.0$.

Uma das revisões mais completas sobre índices de validação fuzzy foi a publicada por Wang e Zhang em (WANG; ZHANG, 2007). Neste trabalho, foram revisados 25 índices de validação agrupados pelo FCM com os seguintes parâmetros: fator de fuzzicação $m = 2.0$ (com exceção do PBMF com $m = 1.2$), critério de parada $\varepsilon = 0.00001$ e número de clusters variando de 2 até $\sqrt{n}$. Nesta publicação foram utilizadas 16 bases de dados, sendo metade artificiais e a outra metade são bases conhecidas como, por exemplo, a Iris, Wisconsin Breast Cancer (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC) e Wine. Destas bases, a de maior dimensionalidade é a WDBC com 30 dimensões.

Na dissertação de mestrado (VENDRAMIN, 2012) foram estudados diferentes algoritmos de agrupamento fuzzy de dados, além do FCM, como os algoritmos que: buscam por grupos hiperelípticos; contornos; que lidam com valores ausentes e atributos categóricos, além de outros. Neste estudo foram revisados 10 dos 16 índices aqui utilizados, em que para cada algoritmo e índices revisados, foram realizadas as análises de complexidade dos mesmos em relaçao ao tempo e espaço.

Em (VALENTE, 2013) foi proposto um método de validação de agrupamento composto por uma etapa quantitativa e qualitativa, onde foi proposta uma maneira de visualização para a avaliação da qualidade das partições obtidas. Os índices de validação PC, PE, MPC, FS e XB foram revisados e utilizados na compação com o novo médoto. Foram agrupadas quatro bases sintéticas e quatro reais (Iris, Wine, Glass e WDBC) tendo o FCM sido executado com os seguintes parâmetros: número máximo de iterações $T = 200$; $2 \leq c \leq 9$; $m = 2.0$.

Em (LUCIEER; LUCIEER, 2009) o agrupamento fuzzy de dados foi utilizado a fim de desenvolver técnicas de classificação de sedimentos no fundo do mar, utilizando o FCM para identificar as zonas de transição indeterminadas e graduais entre estes sedimentos. Neste

trabalho, os dados utilizados são constituídos de uma combinação de dados físicos relacionados com a flora e fauna marinha. Dos índices de validação aqui utilizados, o PC, PE, MPC, FHV, XB, PBMF e PCAES foram revisados e executados nesta publicação. O FCM foi executado sobre os cinco principais componentes a qual a base foi reduzida, ao aplicar a técnica de Análise dos Componentes Principais (PCA), com os seguintes parâmetros: $2 \leq c \leq 20$; $m = 2.0$.

Todos os trabalhos citados anteriormente agrupam bases de dados de baixa dimensionalidade, sendo a maior com 30 dimensões, utilizando o fator de fuzzificação $m = 2.0$, a distância Euclidiana e número máximo de clusters e valor de critério de parada a escolha do autor. Além disto, todos os trabalhos citados e artigos publicados neste segmento ignoram bases de dados de alta dimensionalidade, como as utilizadas neste trabalho (Tabela 1, Seção 4.1), que são as que mais necessitam serem avaliadas por índices estatísticos, já que a visualização destes dados de forma original e a possível identificação de clusters é uma tarefa difícil.

Na Figura 5 foram resumidos os trabalhos relacionados, comentados anteriormente, com os respectivos índices de validação utilizados em comum com este trabalho, além de apresentar a maior dimensão encontrada entre as bases utilizadas.

Figura 5 – Trabalhos relacionados com os respectivos índices de validação e o maior número de dimensões das bases utilizadas

Trabalho	PC	PE	MPC	KYI	P	MPO	GD	FS	FHV	XB	K	SC	PBMF	PCAES	T	SF	Maior dimensão
(Dave, Rajesh N., 1996)	√	√	√						√								2
(R.J.G.B. Campello;E.R. Hruschka, 2006)									√	√						√	4
(Mohammad Rawashdeh;Anca L. Ralescu, 2012)	√								√	√			√			√	-
(Yating Hu et al., 2011)	√	√		√		√		√						√			13
(Sreeram Joopudi et al., 2013)	√	√	√				√				√						4
(Fangfang Zhang;Xuezhong Qian, 2012)	√	√		√						√			√				13
(N. R. Pal;J. C. Bezdek, 1995)	√	√						√		√							4
(Weina Wang;Yunjie Zhang, 2007)	√	√	√	√	√			√	√	√	√	√	√	√	√		30
(Rafael Xavier Valente, 2013)	√	√	√					√		√							30
(Lucas Vendramin, 2012)	√	√	√					√	√	√	√		√		√	√	8
(Leila R. S. da Silva;Fernando Gomide, 2004)	√	√						√		√							4
(V. Lucieer;A. Lucieer, 2009)	√	√	√						√	√			√	√			5

Além da grande diferença entre a maior dimensão apresentada pelas bases utilizadas nos trabalhos relacionados e no presente estudo com, respectivamente 30 e 66602 dimensões, foi possível perceber pela Figura 5 que os índices mais comumente utilizados ainda são o PC, PE, MPC, XB e FS propostos no intervalo de tempo de 1974 a 1996 (Figura 4).

3.1 Considerações Finais

Pela seção anterior foi possível perceber que muitos são os trabalhos que tanto utilizam agrupamento fuzzy para resolução de problemas finais (Sistemas de Recuperação de Informação, organização flexível de documentos e classificação de imagens) como também para tentar aperfeiçoar, sugerir ou até mesmo corrigir falhas, como é o caso do MPC, que possam interferir em um resultado indesejado dos algoritmos ou dos índices de validação. Este aperfeiçoamento ou correção de possíveis falhas só é possível de ser executado depois que uma investigação é realizada.

Este trabalho propõe-se a investigar se os índices de validação apresentarão o mesmo comportamento encontrado na literatura ao avaliarem partições de bases de alta dimensionalidade, utilizando os mesmos valores de parâmetros do FCM majoritariamente utilizados nos trabalhos relacionados. Na seção seguinte são apresentados os materiais e métodos utilizados no presente estudo com: as características das bases textuais utilizadas (Seção 4.1); as razões que motivaram esta investigação (Seção 4.2) e a metodologia aplicada para a mesma (Seção 4.3).

4 MATERIAIS E MÉTODOS

4.1 Bases Textuais

Como as bases textuais, também chamadas de coleções, possuem alta dimensionalidade, as mesmas foram escolhidas para serem agrupadas e terem suas partições avaliadas neste trabalho.

Estas coleções foram obtidas já pré-processadas e estão estruturadas em uma matriz Documento x Termo, onde cada documento é representado por um vetor e cada atributo é um termo encontrado na coleção (Definição 2.1). Esta matriz contém em suas células a razão entre a frequência de um termo particular em um documento e o inverso da frequência deste termo na coleção de documentos (*tf-idf* Term Frequency-Inverse Document Frequency) (RIOS; REZENDE; CAMARGO, 2015).

Algumas das coleções utilizadas estão disponibilizadas em [1] e um resumo é apresentado na Tabela 1.

Tabela 1 – Coleções de documentos com seus respectivos números de documentos (#docs), dimensões (#termos), tópicos (#classes) e domínio

Coleção	#docs	#termos	#classes	Domínio
NewYorkTimes	18	5951	9	Páginas Web
IAarticles	40	66602	4	News articles
Opinosis	51	10784	3	Análise Sentimental
CSTR	299	1725	4	Científico
SyskillWebert	334	4340	4	Páginas Web
Hitech	600	6593	6	News articles
WAP	1560	8068	20	Páginas Web
NSF	1600	2804	16	Científico
Irish-Sentiment	1660	8658	3	Análise Sentimental
20Newsgroups	2000	11026	4	E-mails
La1s	3204	13196	6	News articles
Reviews	4069	22926	5	News articles

[1] http://sites.labic.icmc.usp.br/text_collections/

4.2 Motivação

A fim de verificar se os valores dos parâmetros comumente utilizados no algoritmo Fuzzy C-Means também seriam indicados ao agrupar bases de alta dimensionalidade, o experimento realizado em Wang e Zhang (WANG; ZHANG, 2007) foi então reproduzido, utilizando os mesmos valores de parâmetro do FCM com erro $\varepsilon = 0.00001$ e fator de fuzzificação $m = 2.0$, tendo somente substituído a distância Euclidiana pela medida de Coseno (Equações 2.4 e 2.5, Seção 2.1), já que no presente trabalho, a base de menor dimensionalidade possui 1725 dimensões (Tabela 1).

Os resultados apresentados por Wang e Zhang e obtidos para as bases de alta dimensionalidade aqui utilizadas são apresentadas respectivamente nas Figuras 6 e 7. Como o único critério de parada adotado foi o erro ε, as bases que não conseguiram concluir o agrupamento com o valor de $\varepsilon = 0.00001$ não serão mostradas por considerar que os valores dos parâmetros do algoritmo deveriam ser os mesmos para que a comparação pudesse ser realizada.

Figura 6 – Número ótimo de clusters por Wang e Zhang (WANG; ZHANG, 2007)

Base de Dados	c	PC	PE	MPC	KYI	P	FS	FHV	XB	K	SC	PCAES	T
DataSet_3_3	3	2	2	**3**	2	**3**	**3**	**3**	2	2	**3**	**3**	2
DataSet_4_3	4	**4**	2	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
DataSet_4_2	4	2	2	2	2	14	11	14	13	**4**	**4**	**4**	**4**
DataSet_4noise	4	2	2	5	2	5	5	2	2	2	5	**4**	2
DataSet_5_2	5	4	2	**5**	**5**	**5**	**5**	**5**	4	4	**5**	4	4
DataSet_6_2	6	**6**	**6**	**6**	**6**	**6**	**6**	**6**	4	4	**6**	4	4
DataSet_10_2	10	2	2	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	4	**10**
DataSet_15_2	15	**15**	17	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**
Iris	2 ou 3	**2**	**2**	**2**	**2**	**2**	5	**3**	**2**	**2**	**3**	**2**	**2**
WBCD	2	**2**	**2**	**2**	**2**	**2**	12	**2**	**2**	**2**	**2**	**2**	**2**
WDBC	2	**2**	**2**	**2**	**2**	**2**	12	22	**2**	**2**	**2**	**2**	**2**
Wine	3	2	2	**3**	2	**3**	13	**3**	**3**	**3**	2	**3**	**3**
Liver Disorder	2	**2**	**2**	**2**	**2**	**2**	4	17	**2**	**2**	4	**2**	**2**
Butterfly	2	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Example_1	3	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
Example_2	4	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**

Figura 7 – Número ótimo de clusters obtido ao reproduzir os experimentos de Wang e Zhang

Base de Dados	c	PC	PE	MPC	KYI	P	FS	FHV	XB	K	SC	PCAES	T
NewYorkTimes	9	2	2	11	2	3	11	2	3	3	2	11	3
CSTR	4	2	2	5	2	5	6	2	3	3	2	5	5
Hitech	6	2	2	4	2	4	8	2	8	8	2	5	8
La1s	6	2	2	8	2	**6**	8	2	3	3	2	8	**6**
Reviews	5	2	2	4	2	4	7	2	4	4	2	4	4

Pôde-se perceber pela Figura 6 que todos os índices acertaram a quantidade de clusters em pelo menos metade das bases utilizadas, tendo sido o índice PE aquele que menos acertou o número ótimo das bases. Já na Figura 7 percebe-se que somente os índices P e T indicaram o número de clusters desejado da base La1s, tendo os índices PC, PE, KYI, FHV e SC indicado para todas as coleções $c = 2$.

Verificou-se então que, pela comparação dos resultados obtidos por Wang e Zhang e os reproduzidos para bases textuais, todos os índices não tiveram o desempenho esperado em relação a quantidade de acerto do número de clusters das bases como também da constante escolha de $c = 2$ para alguns índices como citado anteriormente.

A partir do resultado insesperado apresentado anteriormente e do estudo publicado por Bezdek em (PAL; BEZDEK, 1995), onde o mesmo examina a influência do fator de fuzzicação m em determinar a validade das partições geradas pelo FCM, as seguintes hipóteses foram levantadas:

1) "O desempenho dos índices de validação foi influenciado somente pela alta dimensão dos dados agrupados ou para dado valor de m, diferente de 2.0, os índices terão resultados semelhantes aos dos dados de baixa dimensionalidade mostrados na Figura 6?".

2) "Qual ou quais são os índices de validação que obtém o melhor desempenho ao validar uma partição fuzzy de dados de alta dimensionalidade?"

3) "Qual o valor de fator de fuzzicação m ou intervalo em que este esteja inserido que resultará em um melhor desempenho do FCM e dos índices de validação fuzzy ao agrupar bases de alta dimensionalidade? "

A fim de responder estas questões, o FCM foi executado com diferentes valores para o parâmetro m, onde 12 coleções foram agrupadas e 16 índices de validação avaliaram as partições geradas. Os valores dos parâmetros utilizados neste trabalho e por Bezdek em (PAL; BEZDEK, 1995) são detalhados na seção seguinte.

4.3 FCM

Como este trabalho visa testar diversos valores de parâmetro m do FCM e verificar a sua infuência no agrupamento e nos resultados dos índices, o artigo do Bezdek (PAL; BEZDEK, 1995) foi utilizado como principal referência à metodologia utilizada. Nesta publicação, são utilizadas as bases de dados Iris e Normal-4, ambas com quatro dimensões, e o FCM é executado com os seguintes parâmetros: número de iterações $T = 100$, erro $\varepsilon = 0.00001$, distância Euclidiana e o número de clusters c variando de 2 até 10. Para cada base, o FCM foi executado com13 valores de m no intervalo de [1.01; 7.0] (1.01; 1.1; 1.2; 1.5; 1.8; 2.0; 2.2; 2.5; 2.8; 3.0; 3.5; 4.0; 7.0).

No presente trabalho, o FCM foi executado com os seguintes parâmetros: não foi definido um número de iterações sendo somente considerado como critério de parada o erro $\varepsilon = 0.00001$, medida de dissimilaridade do Coseno, para os cálculos de distância, e valor de c variando de 2 ($c(min)$) até a quantidade de classes das coleções ($c(classes)$) acrescentado de 2, que corresponde a quantidade mínima de clusters utilizada ($c(max) = c(classes) + c(min)$).

A regra de ouro citada por Bezdek (PAL; BEZDEK, 1995) para a definição do valor máximo de *c*, $c(max) = c \leq \sqrt{n}$, não foi utilizada neste trabalho por considerar que a raiz quadrada do número de documentos das coleções utilizadas não seria um valor apropriado, visto que a base com menor número de documentos, a NewYorkTimes (*n = 18*), teria um valor de $c(max) \leq 4(\sqrt{18} \simeq 4.2)$ menor do que o número de clusters da base (*c(classes)* = 9) e a base com maior número de documentos, a Reviews (*n = 4069*), teria um valor de *c(max),* $c(max) \leq 64(\sqrt{4069} \simeq 63.8)$, muito superior ao número ótimo da coleção (*c(classes)* = 5).

Por algumas coleções não conseguirem concluir o agrupamento com valor de erro utilizado por Bezdek, estas tiveram que ter este valor de erro mínimo aumentado. As coleções IAarticles, Opinosis, SyskillWebert, Irish-Sentiment e 20Newsgroups utilizaram ε = 0.0001 e as coleções WAP e NSF usaram ε = 0.01.

Para cada coleção, o FCM foi executado com os valores de *m* no intervalo de [1.01; 10.0]. No total foram 32 valores de *m* utilizados: 1.01; 1.015; 1.02; 1.025; 1.03; 1.035; 1.04; 1.045; 1.05; 1.1; 1.3; 1.5; 1.8; 2.0; 2.2; 2.5; 2.8; 3.0; 3;5, 4.0; 4.5; 5.0; 5.5; 6.0; 6.5; 7.0; 7.5; 8.0; 8.5; 9.0; 9.5 e 10.0.

Para os valores iniciais de *m*, foram utilizados intervalos pequenos de 0.005, a partir de *m = 1.1* o intervalo variou entre 0.2 e 0.3 e a partir de *m = 3.0* os intervalos ficaram maiores com diferença de 0.5 entre os valores.

A Tabela 2 apresenta um resumo dos valores de parâmetros, bases de dados e índices de validação utilizados por Bezdek em (PAL; BEZDEK, 1995) e os utilizados neste trabalho.

Tabela 2 – Comparação dos valores e quantidade de parâmetros utilizados nas duas investigações

Parâmetros	Bezdek (PAL; BEZDEK, 1995)	Presente trabalho
Quantidade de bases	2	12
Base de maior dimensionalidade	4	66602
Medida de dissimilaridade	Distância Euclidiana	Coseno (Equações 2.4 e 2.5)
Variação do número de clusters c	$2 \leq c \leq 10$	$2 \leq c \leq c(max)$
Número de iterações	100	-
Erro	0.00001	0.00001; 0.0001; 0.01
Parâmetro m	[1.01; 7.0]	[1.01; 10.0]
Total de valores de m	13	32
Índices de validação	PC, PE, FS, XB_2, XB	PC, PE, MPC, KYI, P, MPO, GD, FS, FHV, XB, K, SC, PBMF, PCAES, T, SF
Total de índices avaliados	5	16

4.4 Considerações Finais

Os valores calculados pelos índices para cada partição gerada pelo FCM, ao utilizar os valores de parâmetros citados anteriormente, e a discussão acerca destes resultados são apresentados na próxima seção.

5 RESULTADOS E DISCUSSÃO

5.1 Resultados

Os resultados dos índices de validação de todas as coleções para cada valor de *m* utilizado do intervalo de [1.01; 10.0] são apresentadas em Anexos através das figuras que mostram o valor de cada índice ao variar o parâmetro *c* de *c(min)* até *c(max)*, tendo sido destacado em vermelho o número ótimo de clusters obtido. A seguir são apresentadas as análises realizadas para cada índice a partir dos resultados obtidos pelos mesmos.

5.1.1 PC

O PC é um índice de maximização com valores no intervalo de [1/c; 1] definidos por seus limites quando $m \to 1$ e $m \to \infty$ (PAL; BEZDEK, 1995):

$$\lim_{m \to 1} PC = 1; \tag{5.1}$$

$$\lim_{m \to \infty} PC = 1/c. \tag{5.2}$$

A partir dos resultados apresentados do PC, nas figuras das Tabelas 5 - 16 Anexo A, para todas as coleções foi possivel perceber que o PC não acertou a quantidade de clusters de nenhuma coleção, tendo somente a base NewYorkTimes obtido, para alguns valores de *m*, uma quantidade de clusters diferente de 2. Também foi perceptível que a tendência monotônica do índice foi muito presente em todas as coleções e para a grande maioria dos valores do fator de fuzzificação, além do valor de *m = 1.01* ter apresentado o maior valor de PC em todas as coleções, com NewYorkTimes e NSF tendo o maior (0.996) e menor valor (0.587) respectivamente. O fato de para todas as coleções, o valor de *m = 1.01* ter obtido o maior valor de PC, mostra a grande influência que este índice sofre da quantidade de clusters e do valor do parâmetro *m*.

Para cada base, foram considerados como valores de PC mais próximos de 1 (Equação 5.1) aqueles que estiveram na segunda metade do intervalo [1/c; 1]. Neste caso, os valores de *m* a partir de 1.04 foram considerados mais próximos do infinito ($m \to \infty$) do que de 1 ($m \to 1$).

5.1.2 PE

O PE é um índice de minimização com valores no intervalo de [0; $\log_a c$] definidos por seus limites quando $m \to 1$ e $m \to \infty$ (PAL; BEZDEK, 1995):

$$\lim_{m \to 1} PE = 0; \tag{5.3}$$

$$\lim_{m \to \infty} PE = \log_a c. \tag{5.4}$$

Nas figuras das Tabelas 17 - 28 do Anexo B pôde-se perceber que o PE constantemente selecionou *c = 2* como número ótimo de clusters para todas as coleções, com exceção da NewYorkTimes que selecionou *c = 11* para *m* igual a 1.01 e 1.015. Isto deve-se ao limite de PE quando *m* tende ao infinito (Equação 5.4), resultando em *c = 2* ser frequentemente selecionado por ser o valor que irá minimizar $log_{10}c$.

A partir dos resultados apresentados do PE para todas as coleções foi possível perceber que somente a base NewYorkTimes obteve, para *m = 1.01* e *m = 1.015*, uma quantidade de clusters diferente de 2. Também foi perceptível a tendência monotônica do índice que, ao contrário do PC, tem seu valor aumentado a medida que o número *c* de clusters também aumenta. Esta tendência também pôde ser verificada pelo fato do valor de *m = 1.01* ter apresentado o menor valor de PE em todas as coleções, com NewYorkTimes e NSF tendo o menor (0.007) e maior valor (0.250) respectivamente. Assim como o PC, o PE apresentou uma grande infuência da quantidade de clusters e do parâmetro *m*.

Para cada base, foram considerados como valores de PE mais próximos de 0 (Equação 5.3) aqueles que estiveram na primeira metade do intervalo [0; $log_{10}c$] . Neste caso, os valores de *m* a partir de 1.04 foram considerados mais próximos do infinito ($m \to \infty$) do que de 1 ($m \to 1$).

5.1.3 MPC

O MPC é um índice de maximização com valores no intervalo de [0; 1] definidos por seus limites quando $m \to 1$ e $m \to \infty$ (PAL; BEZDEK, 1995):

$$\lim_{m \to 1} MPC = 1; \tag{5.5}$$

$$\lim_{m \to \infty} MPC = 0. \tag{5.6}$$

A partir dos resultados apresentados nas figuras das Tabelas 29 - 40 do Anexo C foi perceptível que o índice conseguiu corrigir a tendência monotônica presente no PC, como

proposto em (DAVE, 1996). Devido aos limites do MPC (Equações 5.5 e 5.6), o valor de m = *1.01* novamente apresentou o maior valor de MPC em todas as coleções, com NewYorkTimes e NSF tendo o maior (0.996) e menor valor (0.180) respectivamente. Assim como o PC e o PE, o MPC também apresentou uma grande infuência do parâmetro m em seus resultados.

Foram considerados como valores de MPC mais próximos de 1 aqueles que estivessem na segunda metade do intervalo [0; 1], tendo estes valores de m mais próximos de 1 (Equação 5.5), sendo o contrário considerado como valores de m mais próximos do infinito (Equação 5.6). No caso do MPC, o valor de m a partir de 1.045 já pôde ser considerado como mais próximo do infinito ($m \rightarrow \infty$).

5.1.4 KYI

Nas figuras das Tabelas 41 - 52 do Anexo D, pôde-se perceber que KYI não acertou o número de clusters de nenhuma coleção, apresentando a mesma tendência monotônica verificada no PE, além do valor de KYI ter crescido com o aumento do valor de m. O fator de fuzzificação m = *1.01* apresentou os menores valores de KYI em todas as coleções, com NewYorkTimes e NSF tendo o menor (0.0007) e maior valor (394.794) respectivamente.

5.1.5 P

Nas figuras das Tabelas 53 - 64 Anexo E, pôde-se perceber que o valor do índice P diminuia com o aumento do valor de m. O fator de fuzzificação m = *1.01* apresentou os maiores valores de P em todas as coleções, com NewYorkTimes e NSF tendo o maior (0.998) e menor valor (0.181) respectivamente.

5.1.6 MPO

O MPO (figuras das Tabelas 65 - 76, Anexo F) apresentou a mesma tendência monotônica verificada no índice PE, onde, a medida que o número de clusters aumentava, o valor do MPO também aumentou. Este índice, apesar de também ter sofrido a influência do fator de fuzzificação m, onde a medida que o valor de m aumentava, o valor do índice diminuia, esta influência não foi constante como ocorrido nos índices anteriores. Isto foi comprovado pelo fato do fator de fuzzificação m = *1.01* ter apresentado os maiores valores de MPO na maioria das coleções com exceção da 20Newsgroups (m = *1.025*), NewYorkTimes (m = *4.5*) e Irish-Sentiment (m = *4.5*). Entre as coleções, a SyskillWebert e 20Newsgroups obtiveram o maior (152.104) e menor valor de MPO (7.444) respectivamente.

5.1.7 GD

O índice GD (figuras das Tabelas 77 - 88, Anexo G) apresentou a mesma tendência monotônica verificada no índice PC, onde, a medida que o número de clusters aumenta, o valor do índice diminuia. Pôde-se perceber pelas figuras apresentadas que o índice foi fortemente

influenciado pelo fator de fuzzificação, onde a medida que o mesmo aumentava, o valor do índice decrescia. O fator de fuzzificação *m = 1.01* apresentou os maiores valores de GD em todas as coleções, com SyskillWebert e NSF tendo o maior (0.956) e menor valor (0.177) respectivamente.

5.1.8 FS

O FS (figuras das Tabelas 89 - 100, Anexo H) apresentou a mesma tendência monotônica verificada no PC. Esta tendência pôde ser verificada pela frequente escolha de *c = c(max)* em todas as coleções. Entre estas, a Opinosis obteve o menor valor de FS, igual a -0.547, e a CSTR, SyskillWebert, Hitech, Irish-Sentiment, 20Newsgroups, La1s e Reviews tiveram o maior valor de FS, igual a 0, entre as coleções.

O FS apresentou valores indeterminados para o número de clusters em 10 das 12 coleções. Esta indeterminação deve-se ao menor valor de FS calculado para diferentes clusters ter sido igual a 0. Os valores de *m* em que esta convergência para 0 ocorreu deve-se ao limite do índice quando $m \rightarrow \infty$ (PAL; BEZDEK, 1995):

$$\lim_{m \to \infty} FS = 0 \tag{5.7}$$

A seguir são apresentadas as coleções que tiveram valor indeterminado de número de clusters e para quais valores de *m* esta indeterminação ocorreu, ou seja, quais valores de *m* se aproximaram do ∞.

- IAarticles: *m = 8.5* e *m = 10.0* (Tabela 90);
- Opinosis: *m = 8.0* e *m = 8.5* (Tabela 91);
- CSTR: $m \geq 8.0$ (Tabela 92);
- SyskillWebert: *m = 10.0* (Tabela 93);
- Hitech: $m \geq 6.5$ (Tabela 94);
- WAP: $m \geq 5.0$ (Tabela 95);
- NSF: $m \geq 5.0$ (Tabela 96);
- 20Newsgroups: $m \geq 6.5$ (Tabela 98);
- La1s: $m \geq 8.0$ (Tabela 99);
- Reviews: $m \geq 9.0$ (Tabela 100).

5.1.9 FHV

O índice FHV apresentou a mesma tendência monotônica verificada no índice PE. Esta tendência resultou na escolha de $c = c(min)$ em todas as coleções e valores de m, com exceção da IAarticles e Opinosis.

Pôde-se também perceber pelas figuras das Tabelas 101 - 112 do Anexo I que a variação do menor valor de FHV entre os valores de m foi muito pequena, considerando-se assim que o índice não foi influenciado pelo fator de fuzzificação. Entre as coleções, a NSF obteve o menor valor de FHV, igual a 1.874, e a La1s obteve o maior valor igual a 93.972.

5.1.10 XB

Nas figuras das Tabelas 113 - 124 do Anexo J, pôde-se perceber que o XB comportou-se da seguinte maneira enquanto o fator de fuzzificação m aumentava: o valor de XB aumentava até atingir um pico, entre $m = 6.0$ ou $m = 6.5$, e partir deste valor começou a decrescer. Este comportamento também foi verificado pelo fato do valor de $m = 10.0$ ter obtido o menor valor de XB em todas as coleções, com a WAP e Opinosis tendo respectivamente o menor (9.797E-12) e maior valor (5.12E-7). Para alguns valores de c e m, o XB obteve valor infinito, sendo então representado pela descontinuidade apresentada nas figuras.

5.1.11 K

Pôde-se perceber, pelas figuras das Tabelas 125 - 136 do Anexo K, que o índice K comportou-se de maneira semelhante ao XB em relação ao fator de fuzzificação m. O índice obteve bom desempenho ao acertar a quantidade de clusters das coleções, não acertando a quantidade de clusters da Hitech e WAP. Entre as coleções, a CSTR obteve o menor valor de K, igual a 0.895 para $m = 9.0$, e a WAP obteve o maior valor igual a 47.071, para $m = 3.5$.

5.1.12 SC

O SC (Tabelas 137 - 148) apresentou a mesma tendência monotônica verificada no PC. Esta tendência resultou na escolha de $c = c(min)$ em todas as coleções e valores de m, com exceção da IAarticles, única base em que acertou a quantidade de clusters somente para $m = 1.01$. A NewYorkTimes obteve o menor valor de SC, igual a 0.456, e as demais coleções obtiveram maior valor igual a 0.5.

5.1.13 PBMF

Nas figuras das Tabelas 149 - 160 do Anexo M pôde-se perceber que o PBMF comportou-se da seguinte maneira enquanto o fator de fuzzificação m aumentava: o valor de PBMF inicialmente decrescia seguido de um crescimento exponencial. Este comportamento também foi verificado pelo fato do valor de $m = 10.0$ ter obtido o maior valor de PBMF em cada coleção,

tendo a WAP obtido o maior valor (2.828E21) e a Opinosis e Irish-Sentiment menor valor (1.526E11). Este índice não acertou a quantidade de clusters somente da coleção NewYorkTimes.

5.1.14 PCAES

Nas figuras das Tabelas 161 - 172 do Anexo N pôde-se perceber que o PCAES, a medida que o fator de fuzzificação *m* aumentava comportou-se da seguinte maneira: o valor de PCAES inicialmente decresceu até atingir um valor mínimo, começou a crescer e em seguida ficou constante. Este comportamento também foi verificado pelo fato do valor de *m = 1.01* ter obtido o maior valor de PCAES em todas as coleções com exceção da NewYokTimes. A base SyskillWebert e 20Newsgroups obtiveram respectivamente o maior (150.839) e menor (7.539) valores de PCAES. Este índice não acertou a quantidade de clusters somente da coleção SyskillWebert.

5.1.15 T

Nas figuras das Tabelas 173 - 184 do Anexo O pôde-se perceber que o valor do índice T oscilou (inicialmente crescente e depois decrescente) enquanto o fator de fuzzificação *m* aumentava. O índice T não acertou a quantidade de clusters das coleções NewYorkTimes e WAP. A SyskillWebert e 20Newsgroups obtiveram respectivamente o menor (0.009) e maior (0.406) valor de T.

5.1.16 SF

O SF (Tabelas 185 - 196, Anexo P) foi o único índice que acertou a quantidade de clusters de todas as coleções. Na 20Newsgroups, para *m = 9.5* e *m = 10.0*, o número de clusters *c* ficou indeterminado porque o valor de SF para cada valor de *c* foi igual a 0. A IAarticles e Irish-Sentiment obtiveram respectivamente o maior (0.999) e menor (0.174) valor de SF.

5.1.17 Considerações Finais

A partir dos resultados apresentados na Seção 5.1, foi possível analisar os resultados sob três dimensões: a primeira dimensão verifica o desempenho dos índices de validação para cada coleção de documentos (Seção 5.2.1); a segunda avalia o desempenho dos índices de validação ao variar o fator de fuzzificação *m* (Seção 5.2.2); a última verifica o comportamento dos diferentes valores de *m* (Seção 5.2.3) para cada coleção. Esta análise tridimensional permite detalhar para cada coleção quais os índices e valores de *m* mais apropriados e, para cada índice, quais valores de *m* proporcionariam resultados superiores aos apresentados na Figura 7 da Seção 4.2.

5.2 Discussão

As avaliações apresentadas nas seções seguintes são mostradas em função de três parâmetros: *c(min)*, *c(classes)* e *c(max)*. O primeiro e último parâmetros representam quando o número ótimo de clusters *c* foi respectivamente igual a quantidade mínima (*c = c(min) = 2*) ou máxima (*c = c(max)*) de clusters estabelecidas como intervalo para a execução do FCM como explicado na Seção 4.3. Já o paramêtro *c(classes)* representa quando o número ótimo de clusters *c* foi igual a quantidade de classes (*c = c(classes)*) previamente estabelecida pelos especialistas nas coleções utilizadas, considerando assim que o índice de validação acertou a escolha do número de clusters.

Nas figuras apresentadas a partir desta seção o eixo Y de cada gráfico corresponde ao valor de média de *c(min)*, *c(classes)* e *c(max)* representados respectivamente pelas cores amarelo, vermelho e laranja. O cálculo e significado destas médias são explicados nas seções seguintes.

5.2.1 Coleções e índices de validação

Para verificar o desempenho dos índices de validação para cada coleção de documentos, foi feita uma média da quantidade de vezes que um índice obteve como número ótimo de clusters *c(min)*, *c(classes)* ou *c(max)* pelo total de valores de *m* utilizados. Ou seja, ao avaliar quando *c = c(classes)* por exemplo, para dada coleção, se um índice não acertou a quantidade de clusters em nenhum valor de *m*, este terá valor 0, enquanto se, para a mesma coleção, se o mesmo índice tiver acertado a quantidade de clusters em todos valores de *m* então este terá média de *c(classes)* igual a 1.

Um índice será considerado como tido bom desempenho se o mesmo tiver média *c(classes)* > 0 e média de *c(min)* e *c(max)* abaixo de 0.5. Foi considerado o valor de limiar de 0.5 para *c(min)* e *c(max)* por considerar que um valor de média acima de 0.5 para estes dois parâmetros representaria uma tendência significativa dos índices em escolher como número ótimo de clusters de uma coleção o valor igual a *c(min)* ou *c(max),* independente da escolha do valor de *m* utilizado como parâmetro no algoritmo FCM.

Nas Figuras 8 - 9 é possível verificar que em todas as coleções, pelo menos um índice acertou a quantidade de clusters em ao menos um valor de *m*. A base NewYorkTimes (Figura 8) foi a que menos índices acertaram (K, FS, MPO, PCAES, SF). O contrário aconteceu com a base IAarticles, que apesar de ter a maior dimensão entre as bases utilizadas, foi a que o maior número de índices, no total 12, acertaram a quantidade de clusters, ou seja, tiveram média *c(classes)* > 0.

Figura 8 – Média de *c(min)*, *c(classes)* e *c(max)* para as coleções NewYorkTimes, IAarticles, Opinosis, CSTR, SyskillWebert e Hitech

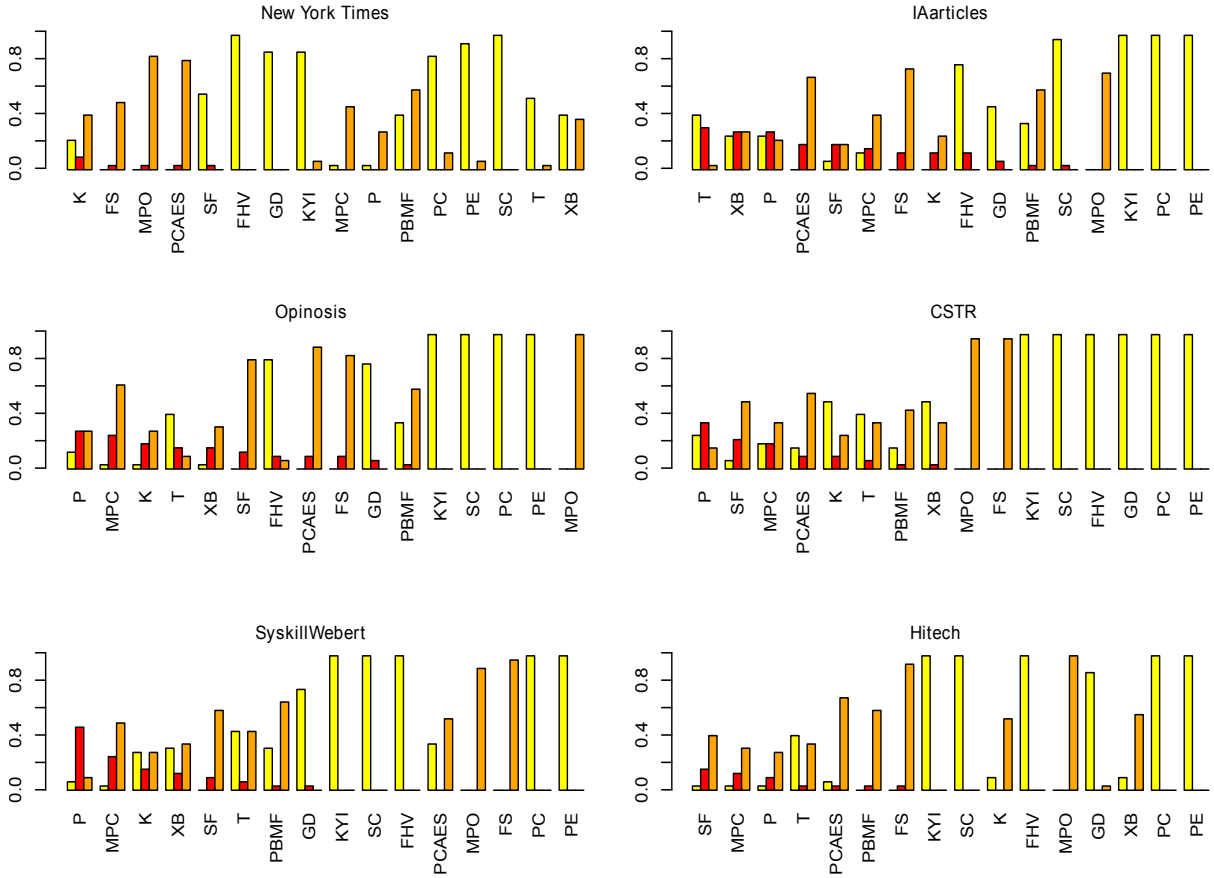
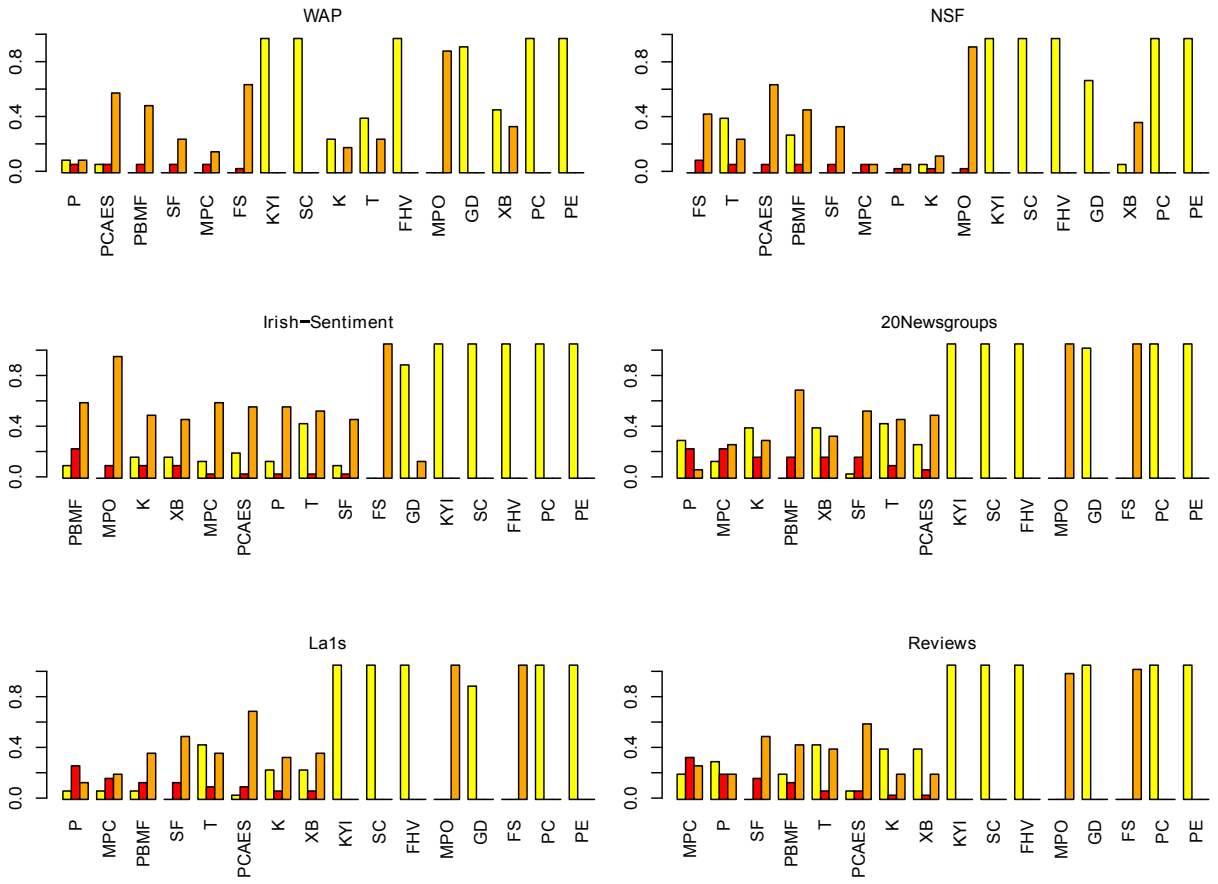


Figura 9 – Média de *c(min)*, *c(classes)* e *c(max)* para as coleções WAP, NSF, Irish-Sentiment, 20Newsgroups, La1s e Reviews



Em todas as coleções, os índices PC, PE, KYI não acertaram o número ótimo de clusters para nenhum valor de m, tendo além destes, os índices GD e SC também apresentado média de *c(min)* muito próxima de 1, com exceção de GD para a base IAarticles (Figura 8). O oposto ocorreu com o MPO, que teve como menor valor de média 0.72 para *c(max)* novamente na base IAarticles, e o FS com valores de média abaixo de 0.51 para *c(max)* somente nas coleções NewYorkTimes e NSF.

A partir das Figuras 8 e 9 serão indicados na Tabela 3, por ordem decrescente de média de acerto, os índices que obtiveram melhor desempenho para cada base, isto é, dos índices que tiveram valor de média *c(classes)* > 0, os mais indicados para cada coleção são aqueles com valores de média de *c(min)* e *c(max)* abaixo ou igual a 0.5.

Tabela 3 – Índices com maiores médias de c(classes) de cada coleção

Coleção	Índices
NewYorkTimes	K, FS
IAarticles	T, XB, P, SF, MPC, K, GD
Opinosis	P, K, T, XB
CSTR	P, SF, MPC, K, T, PBMF, XB
SyskillWebert	P, MPC, K, XB, T
Hitech	SF, MPC, P, T
WAP	P, SF, MPC
NSF	FS, T, PBMF, SF, MPC, P, K
Irish-Sentiment	K, XB, SF
20Newsgroups	P, MPC, K, XB, T, PCAES
La1s	P, MPC, PBMF, SF, T, K, XB
Reviews	MPC, P, SF, PBMF, T, K, XB

Nenhum índice possuiu unanimidade de bom desempenho nas coleções. O índice P obteve melhor resultado em metade das coleções e pode ser indicado para todas as bases com exceção da NewYorkTimes e Irish-Sentiment, seguido do índice K que obteve melhor resultado nestas mesmas duas bases, não sendo indicado somente para as bases Hitech e WAP. Além dos índices P e K, os índices T, MPC, XB e SF também podem ser indicados em mais da metade das coleções tendo T e MPC boa indicação em nove e XB e SF em oito das 12 coleções.

O destaque desta análise são os índices: FS que foi bem indicado para as bases NewYork-Times e NSF, sendo o de melhor desempenho na última, apesar de frequentemente apresentar valores de média elevadas de c(max); SF obteve melhor desempenho somente na Hitech; o bom desempenho geral dos índices P, K, T, MPC, XB e SF, com destaque para o índice P que obteve a maior média de acerto (*c(classes)*) de um índice em todas as coleções (maior média de acerto aproximadamente igual a 0.5 para a base SyskillWebert), o que significa que o índice foi capaz de obter bons resultados em quase metade dos valores de *m* utilizados.

5.2.2 Índices de validação e fator de fuzzificação *m*

Para avaliar o desempenho dos índices de validação ao variar o fator *m*, foi feita uma média da quantidade de vezes que um índice obteve como número ótimo de clusters *c(min)*, *c(classes)* ou *c(max)* pelo total de coleções. Ou seja, ao avaliar quando $c = c(classes)$ por exemplo, para dado valor de *m*, se um índice não acertou a quantidade de clusters de nenhuma coleção, este terá valor 0, enquanto se o mesmo, para um diferente valor de *m*, tiver acertado a quantidade de clusters em todas as coleções, então este terá valor 1.

Nas Figuras 10 - 13 percebe-se que os índices PC, PE e KYI (Figura 10) não acertaram o número de clusters *c* em nenhuma coleção e os índices MPO, GD, FS (Figura 11), FHV e SC (Figura 12) acertaram o valor de *c* em poucas coleções e para poucos valores de *m*. Destes índices, o PC, PE, KYI, FHV e SC apresentaram média muitas vezes igual ou muito próxima de 1, independente do valor de *m*. O índice GD (Figura 11), apesar de também ter apresentado a mesma tendência em escolher $c = c(min)$, somente os valores de *m* no intervalo [1.03; 1.05] e $m = 9.0$ obtiveram média abaixo de 0.7. Já os índices MPO e FS (Figura 11) apresentaram uma tendência para escolha da quantidade máxima de clusters ($c = c(max)$) como número ótimo na maioria das coleções e valores de *m*.

Figura 10 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das partições avaliadas pelos índices PC, PE, MPC e KYI.

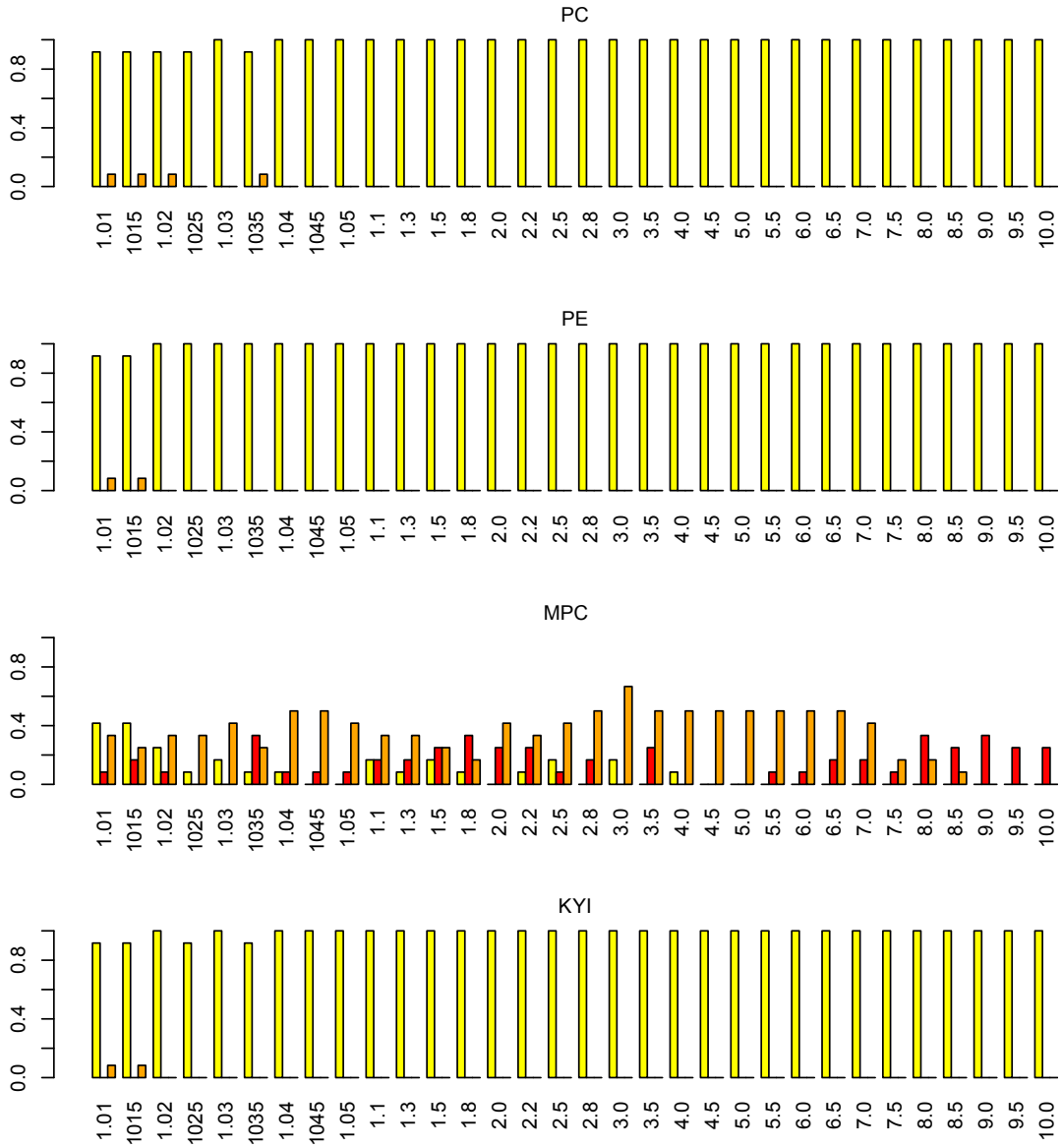


Figura 11 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das partições avaliadas pelos índices P, MPO, GD e FS.

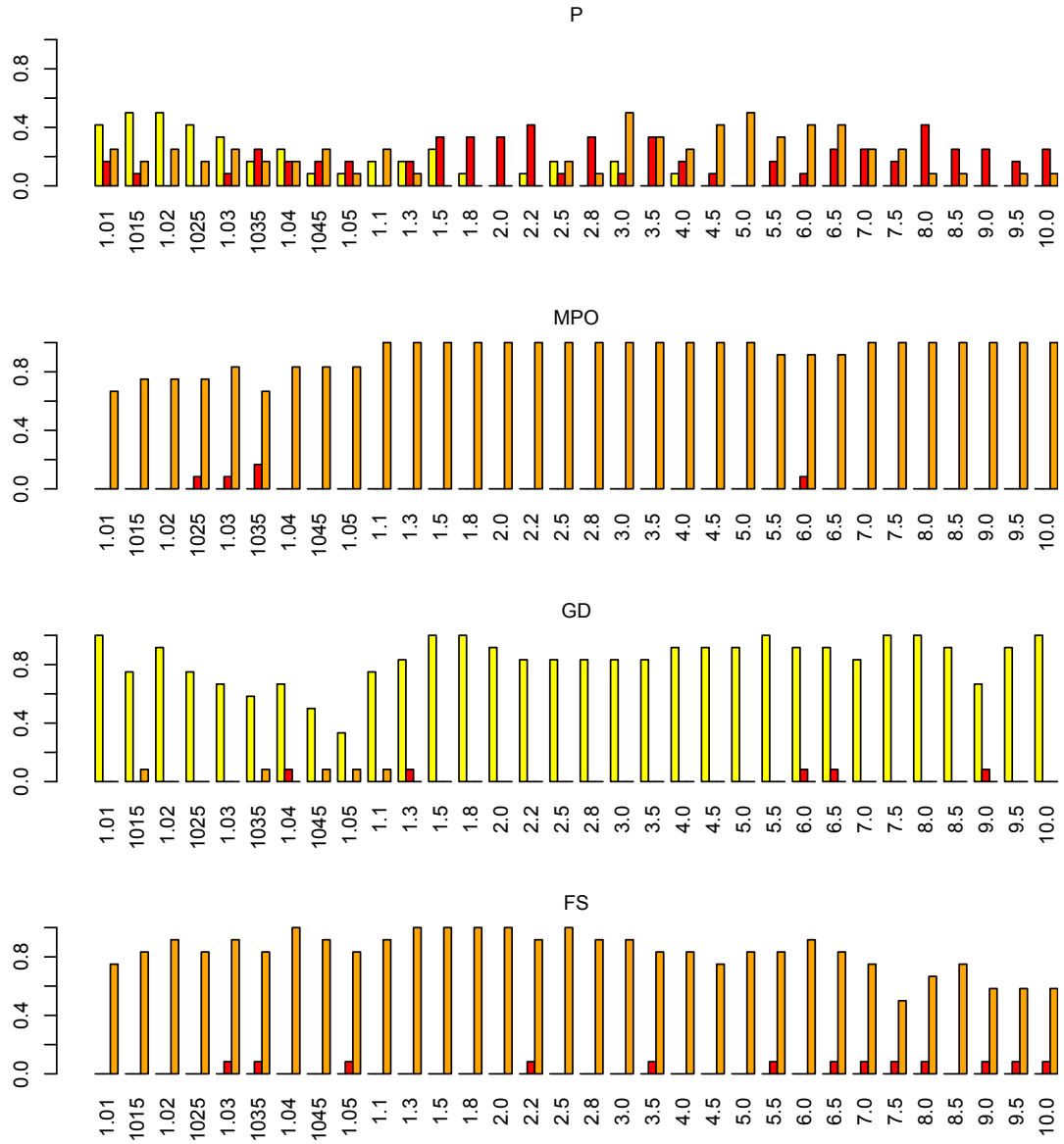


Figura 12 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das partições avaliadas pelos índices FHV, XB, K e SC.

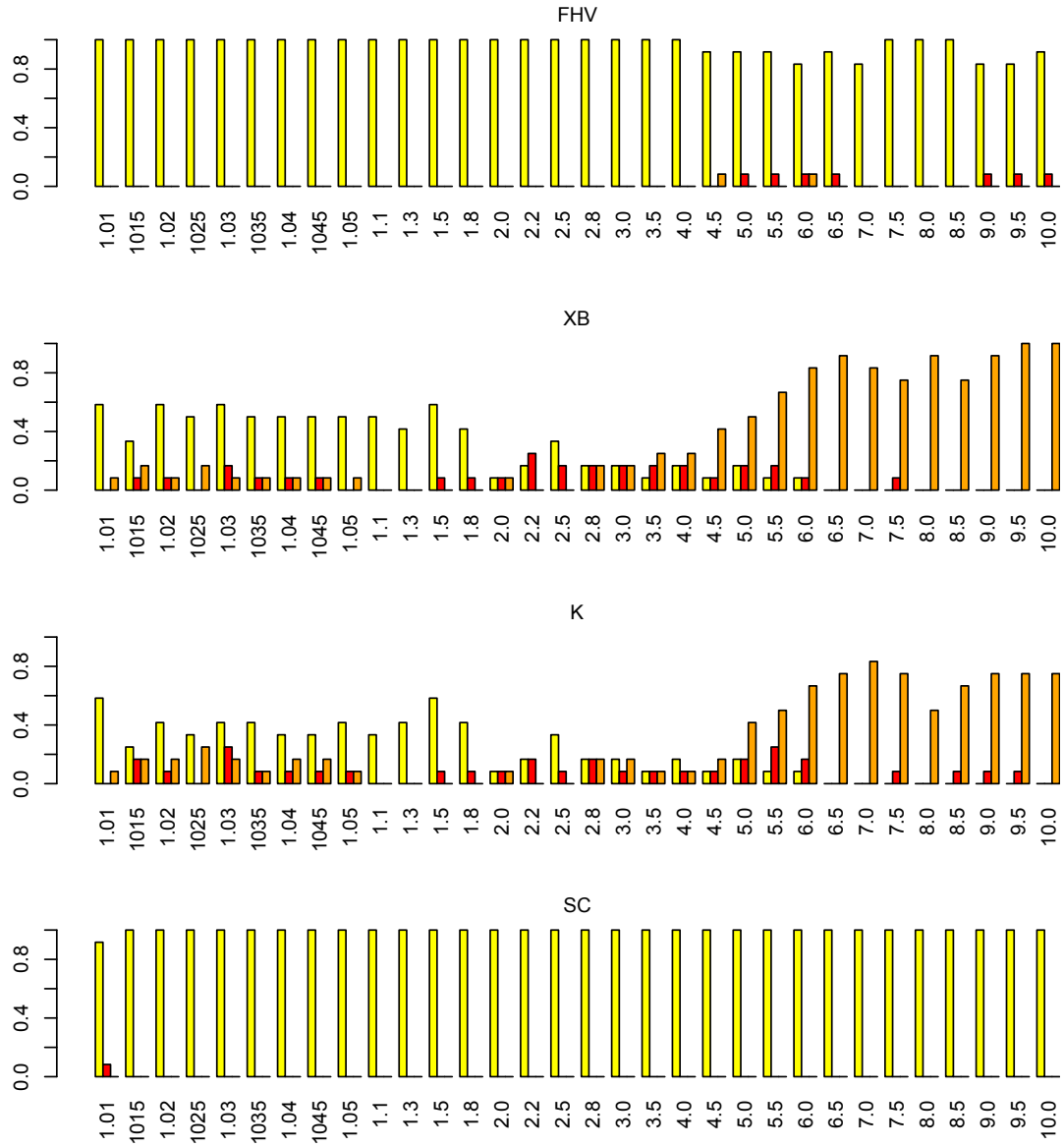
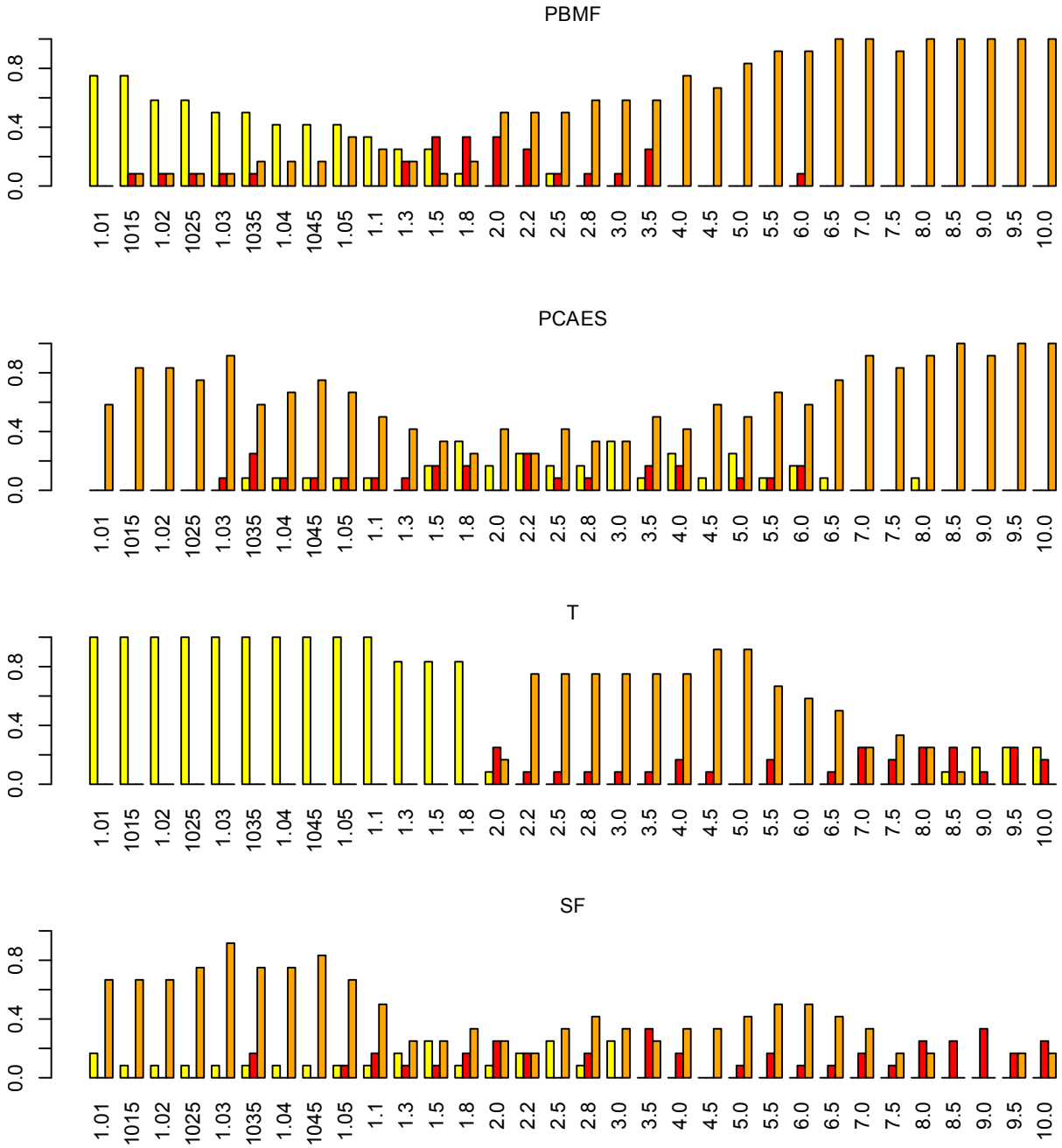


Figura 13 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das partições avaliadas pelos índices PBMF, PCAES, T e SF.



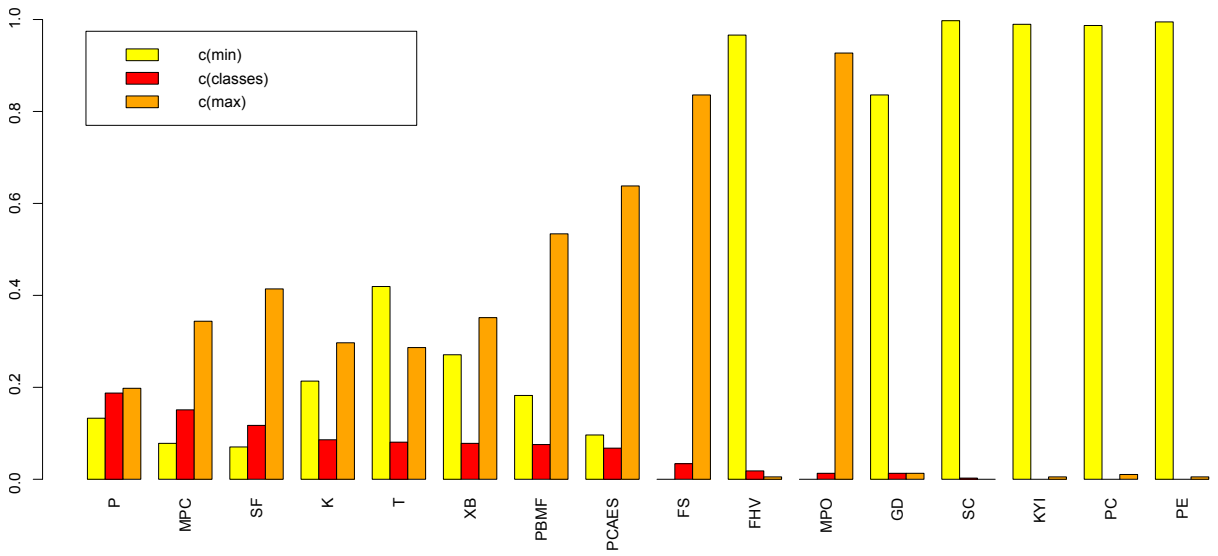
Os índices MPC (Figura 10), PCAES e SF (Figura 13) tiveram comportamento oscilatório em relação a média de *c(max)*. Já nas Figuras 11 - 13, é possível perceber que os índices P, XB, K, PBMF e T tiveram médias de *c(min)* mais altas concentradas nos menores valores de *m*, no intervalo [1.01; 1.8], e médias de *c(max)* mais altas nos maiores valores de *m*, com PBMF e T no intervalo [2.2; 10.0] e XB e K no intervalo [5.0; 10.0]. Segundo os resultados apresentados por estes índices, a medida que o fator de fuzzificação *m* aumenta e as partições se tornam mais fuzzy, então mais clusters serão formados.

Os índices P (Figura 11), MPC (Figura 10) e K (Figura 12) apresentarem valores de

média de *c(classes)* mais distribuídas entre os valores de *m*, com estes não acertando a quantidade ótima de clusters em respectivamente quatro, seis e oito valores de *m.* A partir das Figuras 10 - 13, pôde-se perceber que nenhum índice utilizando os valores de *m* no intervalo estabelecido de [1.01; 10.0] obteve média de acerto acima de 0.5, ou seja, nenhum índice acertou a quantidade de clusters de pelo menos metade das coleções, tendo o índice P o maior valor de média para *m = 2.2* e *m = 8.0,* ambos com média igual a 0.42.

A partir dos valores de média apresentados nas Figuras 10 - 13, foram calculados três valores de média geral de cada índice, para cada um dos três parâmetros *c(min)*, *c(classes)* e *c(max)*, pelo total de valores de *m* e coleções. Na Figura 14 são apresentados os índices de validação com cada um dos seus três valores de média geral em ordem decrescente de média de *c(classes)*.

Figura 14 – Média geral de *c(min)*, *c(classes)* e *c(max)* de cada índice



A partir da Figura 14 foi possível novamente verificar que os índices PC, PE, KYI e SC obtiveram média para *c(classes)* igual a 0 e valores muito próximos de 1 para *c(min)*. Já os índices PBMF, PCAES, FS, FHV, MPO e GD, apesar de possuírem um valor de média de *c(classes)* abaixo de 0.1, o PBMF, PCAES, FS e MPO apresentaram valores de média para *c(max)* acima de 0.5 e o FHV e GD apresentaram média para *c(min)* acima de 0.8.

Os índices P, MPC, SF, K, T e XB foram considerados como os de melhores desempenhos por, além de terem acertado mais vezes a quantidade de clusters, terem apresentado valores de média para *c(min)* e *c(max)* abaixo de 0.5, como explicado na Seção 5.2.1. Apesar destes índices terem apresentado melhor desempenho ao avaliar as partições geradas pelo FCM, a média geral de acerto, quando *c = c(classes)*, foi considerada muito baixa, com o índice P tendo valor aproximado de média de 0.2. Esta baixa média de acerto da quantidade de clusters de uma coleção significa que os índices não conseguiram reconhecer a estrutura presente nas bases.

Portanto, com a análise realizada anteriormente, pôde-se concluir que todos os índi-

ces analisados tiveram seu desempenho influenciado pela alta dimensionalidade dos dados e não pelo valor do fator de fuzzificação *m* utilizado, o que responde a questão levantada na Seção 4.2: "O desempenho dos índices de validação são influenciados somente pela alta dimensão dos dados agrupados ou para dado valor de *m*, diferente de 2.0, os índices terão resultados semelhantes aos dos dados de baixa dimensionalidade mostrados na Figura 6 ¿'.

No entanto, dos índices avaliados, o P, MPC, SF, K, T e XB foram os mais indicados para avaliar agrupamentos fuzzy de conjuntos de dados de alta dimensionalidade, respondendo a seguinte questão levantada na Seção 4.2: "Qual ou quais são os índices de validação que obtém o melhor desempenho ao validar uma partição fuzzy de dados de alta dimensionalidade?" Pela Figura 4 da Seção 2.3 é possível perceber que os dois índices melhor avaliados, P e MPC, utilizam somente a informação da matriz de pertinência para os seus cálculos, o que comprova que os índices desta classe também podem apresentar bons resultados, e no caso deste trabalho, apresentaram os melhores resultados.

O valor do fator de fuzzificação *m* ou o intervalo de valores mais adequado para conjuntos de alta dimensão será discutido na seção seguinte.

5.2.3 Coleções e fator de fuzzificação *m*

Para avaliar quais valores de *m* seriam mais adequados para cada coleção de documentos, visto que cada coleção possui características específicas, foi feita uma média da quantidade de vezes que uma coleção obteve como número ótimo de clusters *c(min)*, *c(classes)* ou *c(max)* pelo número de índices bem avaliados na seção anterior. Ou seja, ao avaliar quando $c = c(classes)$ por exemplo, para dado valor de *m* e uma dada coleção, se nenhum dos seis índices tiver acertado a quantidade de clusters, a média de *c(classes)* terá valor 0, enquanto se para mesma coleção e diferente valor de *m*, se todos os índices tiverem acertado a quantidade de clusters da coleção então a média de *c(classes)* terá valor igual a 1.

Nas Figuras 15 - 18 é possível perceber que todas as coleções, com exceção da SyskillWebert, tiveram médias mais altas de *c(min)* concentradas nos menores valores de *m*, no geral $m < 2.0$, e médias mais altas de *c(max)* concentradas nos maiores valores de *m*, no geral $m > 3.0$, em todas as coleções com exeção da NewYorkTimes e Hitech que tiveram média altas de *c(max)* também concentradas no início do intervalo. Estas coleções ilustram o comportamento relatado na seção anterior em relação aos índices P, XB, K, PBMF e T.

Com as Figuras 15 e 17 é possível verificar que as coleções NewYorkTimes, WAP, Irish-Sentiment e NSF tiveram menor número de valores de *m* acertando (média $c(classes) > 0$) a quantidade de clusters das coleções, enquanto a SyskillWebert e a IAarticles, base de maior dimensionalidade (Tabela 1, Seção 4.1), tiveram acerto em 20 dos 32 valores de *m* utilizados. Esta última também foi destacada na Seção 5.2.1 como sendo a coleção com maior número de índices que acertaram a quantidade ótima de clusters e agora com maior número de valores de *m* com média $c(classes) > 0$. Em contrapartida, além da NewYorkTimes ter sido citada na mesma

seção como tendo menor número de índices obtido média de acerto *c(classes)* > 0, a mesma obteve acerto em somente quatro valores de *m* (Figura 15) apresentando o menor valor de soma de média de *c(classes)* de aproximadamente 0.7.

Figura 15 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das coleções NewYorkTimes, IAarticles e Opinosis.

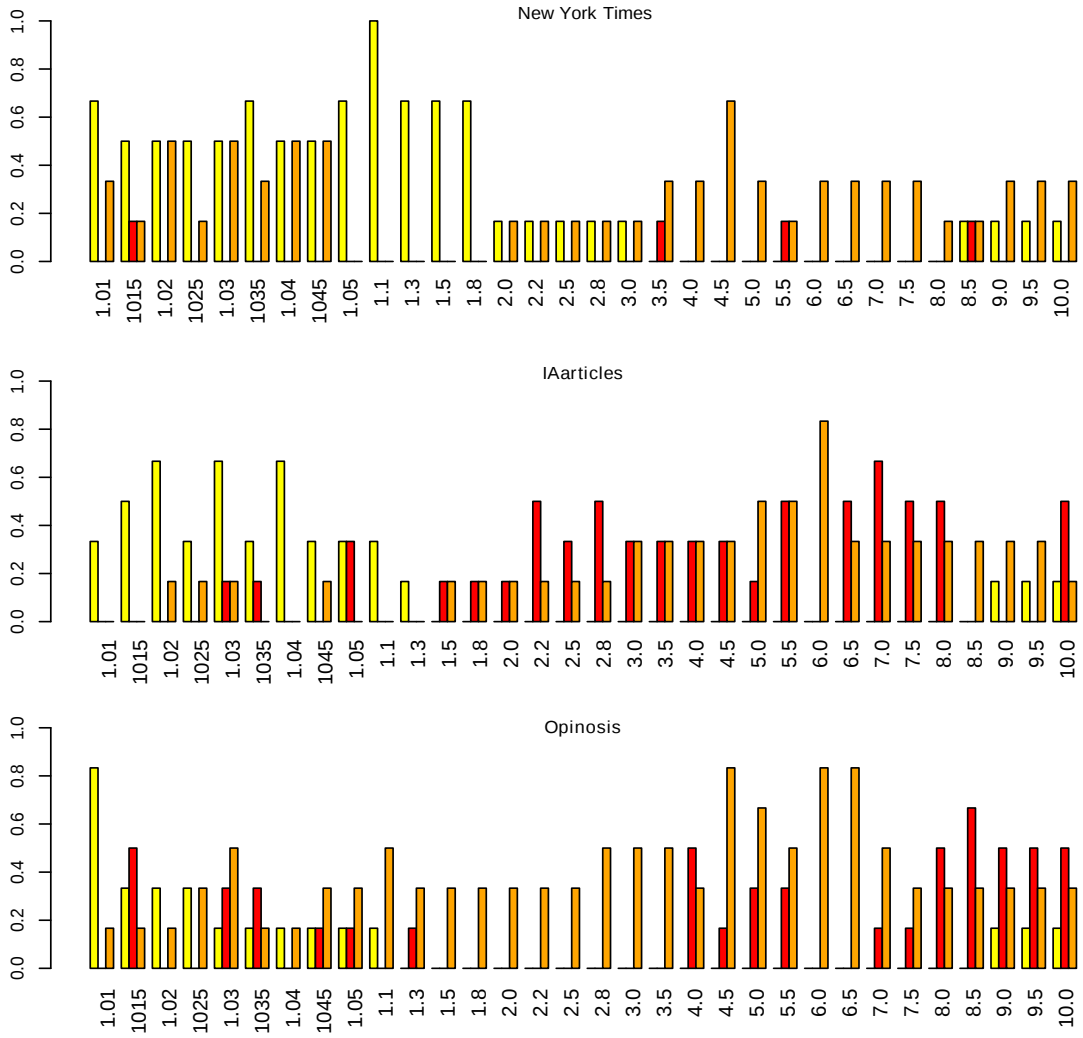


Figura 16 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das coleções CSTR, SyskillWebert e Hitech.

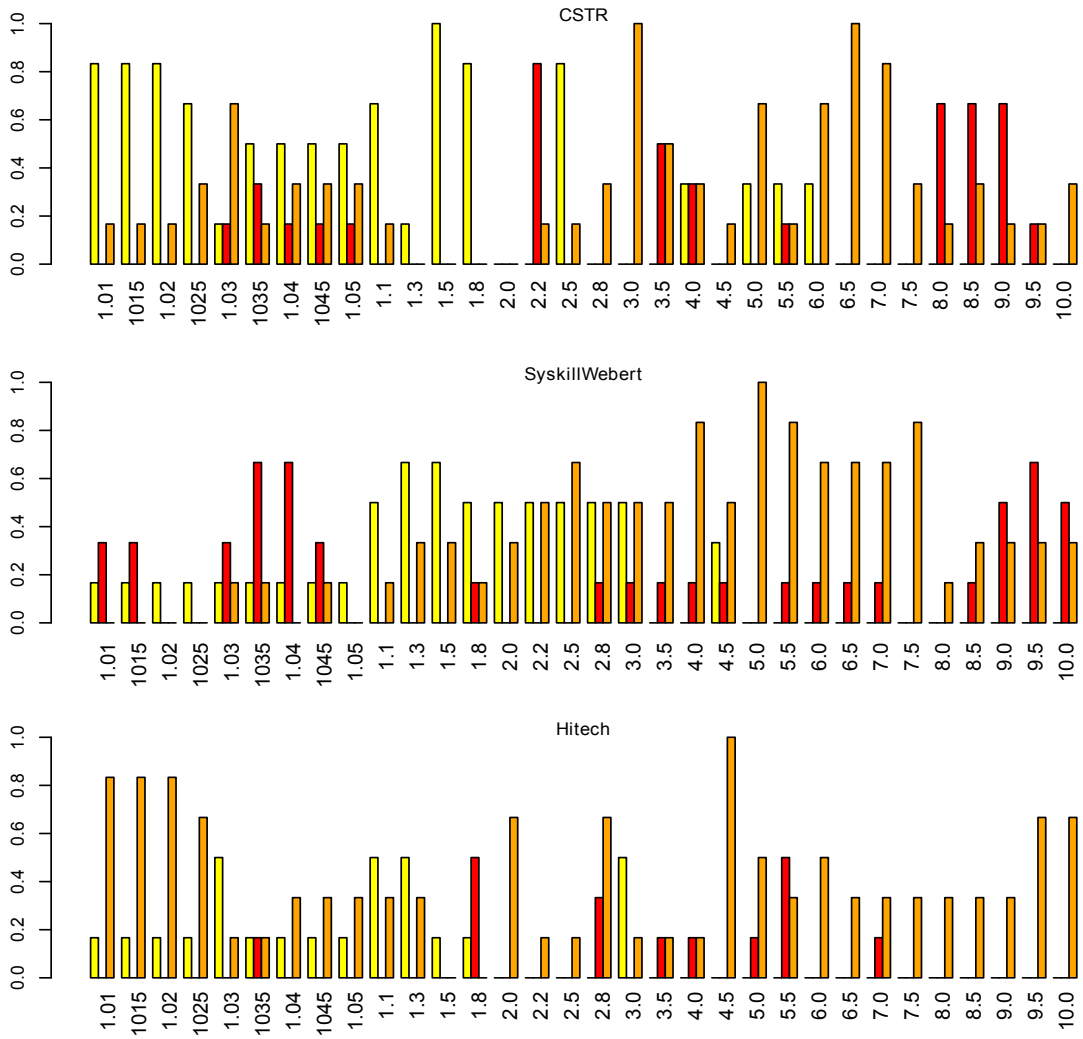


Figura 17 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das coleções WAP, NSF e Irish-Sentiment.

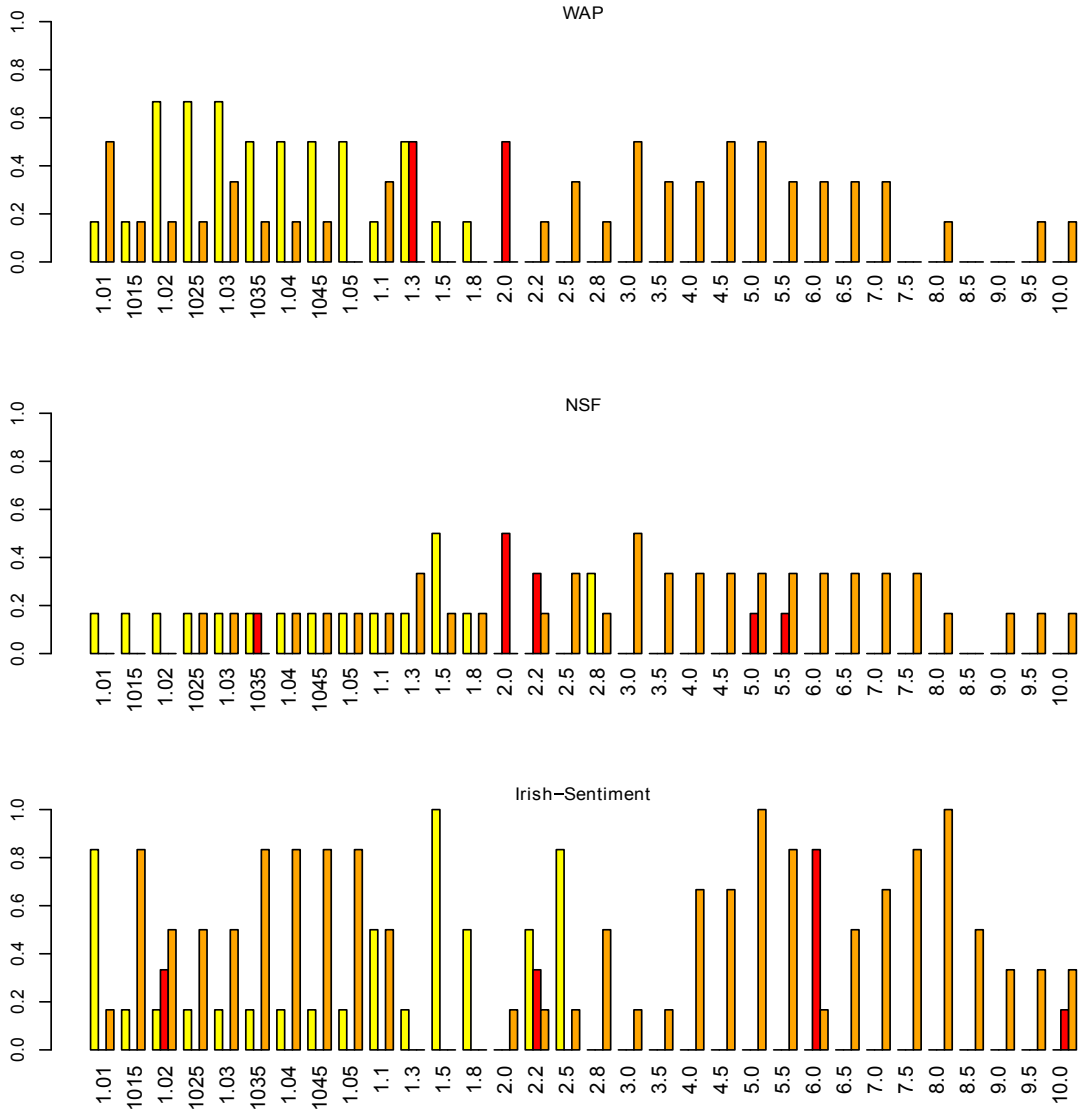
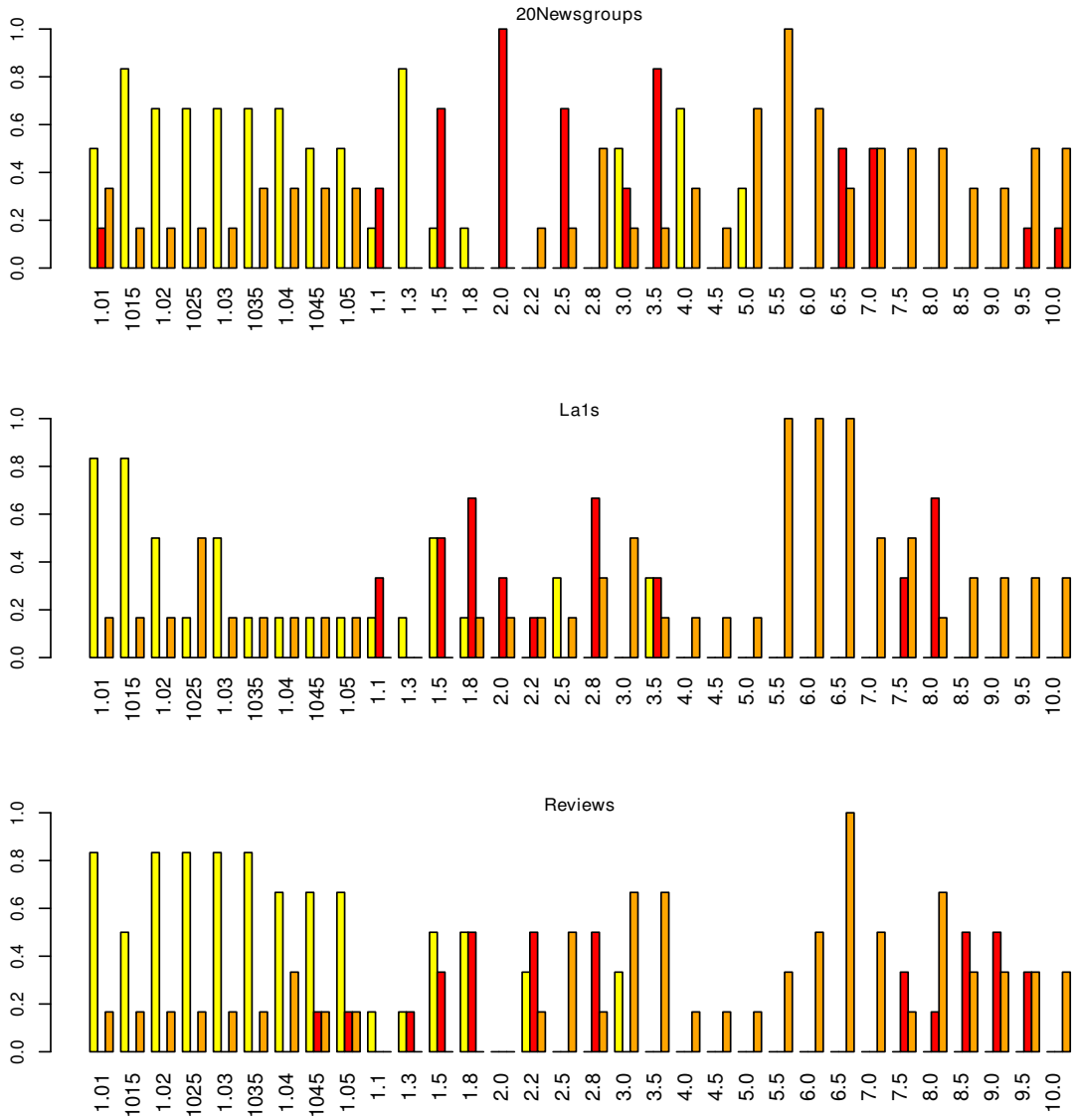
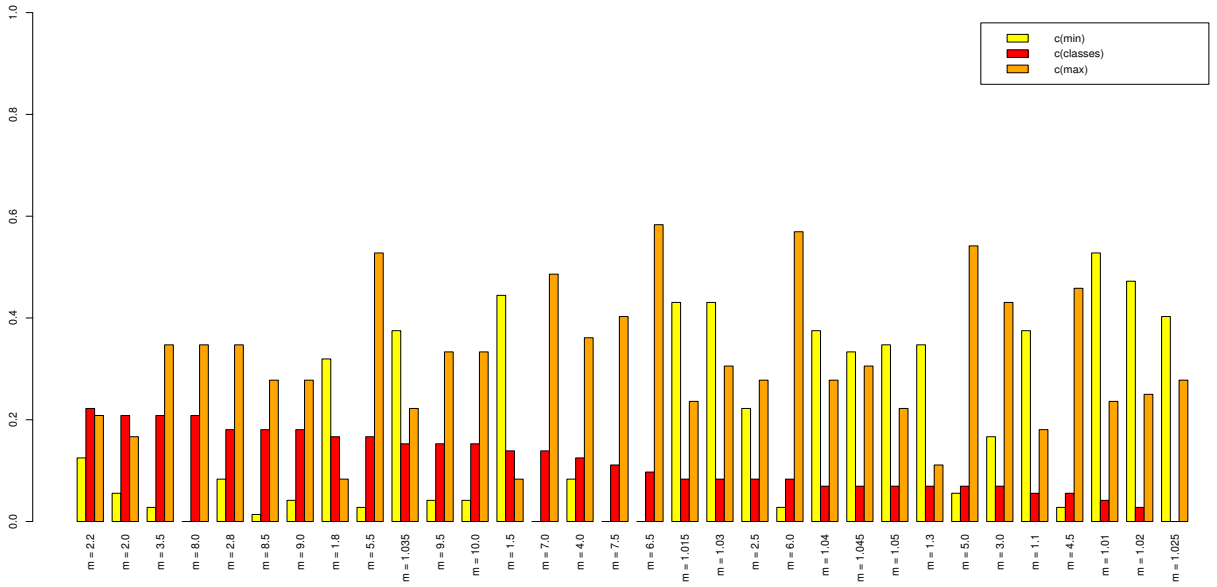


Figura 18 – Média de *c(min)*, *c(classes)* e *c(max)* para os valores de *m* das coleções 20Newsgroups, La1s e Reviews.



A partir dos valores de média apresentados nas Figuras 15 - 18, foram calculados três valores de média geral de cada valor de *m*, para cada um dos três parâmetros *c(min)*, *c(classes)* e *c(max)*, pelo total de coleções e índices bem avaliados. A Figura 19 apresenta os valores de *m* com cada um dos seus três valores de média geral em ordem decrescente de média de *c(classes)*, onde é possível verificar que a maior média de acerto ocorreu com $m = 2.2$, com média igual a 0.22 e o oposto para $m = 1.025$, com média 0. Os melhores valores de *m* na sequência foram 2.0, 3.5 e 8.0, todos com média aproximada de *c(classes)* igual a 0.21, seguido dos valores de *m* igual a 2.8, 8.5 e 9.0, todos com 0.18 de média.

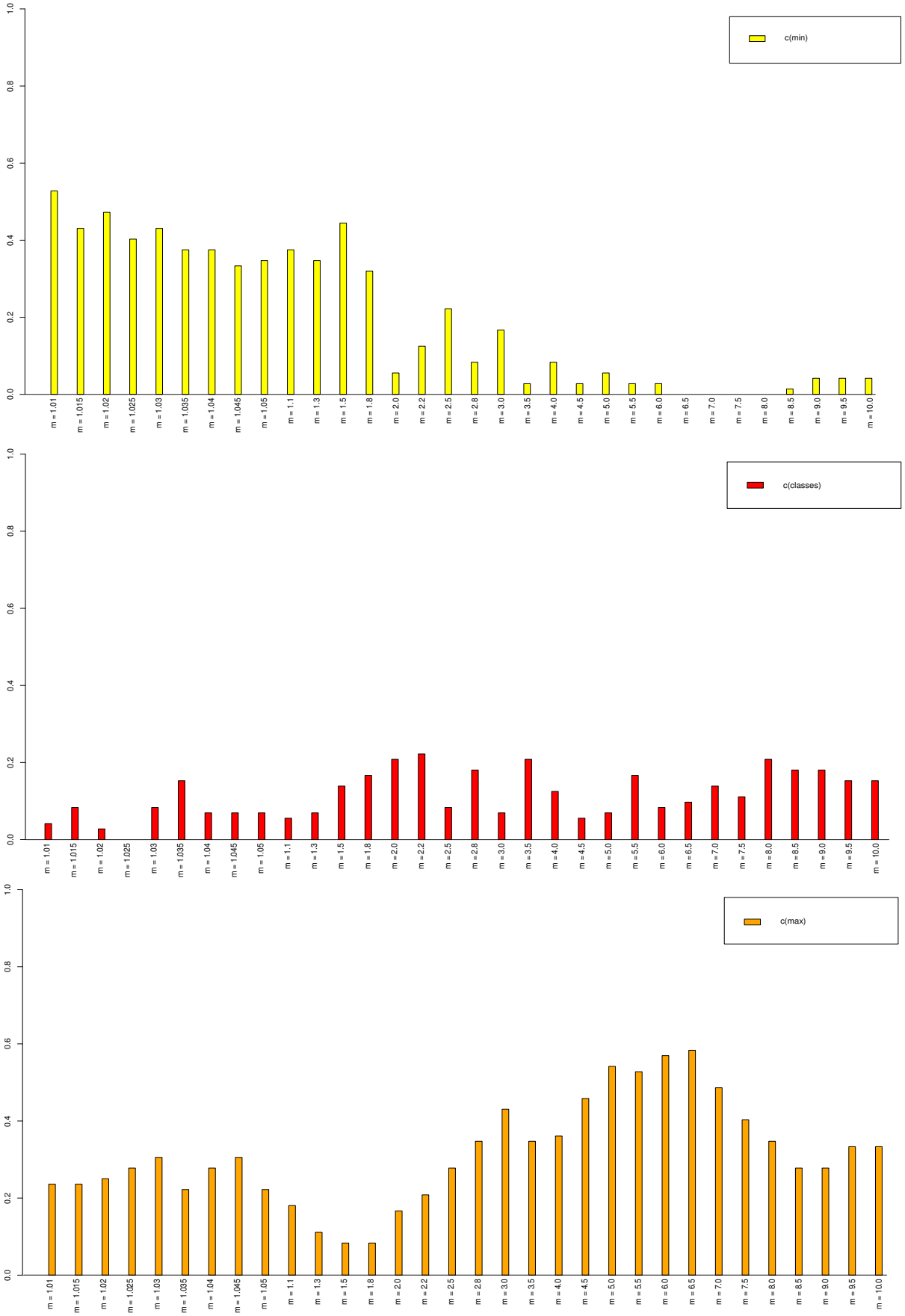
Figura 19 – Média geral de *c(min)*, *c(classes)* e *c(max)* para cada valor de *m*



Na Figura 20 pôde-se perceber que a média de *c(min)* comportou-se de maneira decrescente enquanto o valor de *m* aumentava, com os maiores valores de média para $m < 2.0$. Já as médias de *c(classes)* e *c(max)* comportaram-se de maneira oscilatória com médias de *c(max)* acima de 0.5 no intervalo [5.0; 6.5] e maiores valores de médias de *c(classes),* acima de 0.1, localizados nos intervalos [1.5; 2.2] e [7.0; 10.0], com picos fora dos intervalos com *m* igual a 1.035, 2.8, 3.5 e 5.5. Apesar dos valores de *m* no intervalo [3.5; 4.0] possuírem média de acerto acima de 0.1, este foi desconsiderado como intervalo, mantendo somente $m = 3.5$ como pico, por somente dois valores deste intervalo terem sido experimentados (3.5 e 4.0), além da média de *c(classes)* decrescer entre $m = 3.5$ e $m = 4.0$.

Como a análise realizada neste trabalho considera, não somente o acerto ($c = c(classes)$), como também quando os índices indicam como número ótimo de clusters os valores de *c(min)* ou *c(max)*, de cada coleção, dos valores de *m* com média acima de 0.1, citados anteriormente, os que tiveram média de *c(min)* ou *c(max)* acima de 0.5 foram descartados, sendo então desconsiderado $m = 5.5$.

Figura 20 – Média geral ordenada por valor de *m* para *c(min)*, *c(classes)* e *c(max)*



A fim de priorizar a indicação de bons intervalos para o fator de fuzzificação *m* e não valores isolados, os valores de *m* iguais a 1.035, 2.8 e 3.5 foram descartados restando somente os intervalos [1.5; 2.2] e [7.0; 10.0], tendo o primeiro coincidido quase que totalmente com o intervalo de [1.5; 2.5] sugerido por Bezdek em (PAL; BEZDEK, 1995).

A partir dos resultados apresentados na Seção 5.1 e dos intervalos de valores de *m* ([1.5; 2.2] e [7.0; 10.0]) aqui sugeridos a serem utilizados como valor de parâmetro para o FCM ao agrupar bases de alta dimensionalidade, foi realizada uma comparação entre as partições através dos valores obtidos por cada um dos seis índices, melhor avaliados anteriormente, para cada um dos valores de *m*, dentro e fora do intervalo estabelecido, que acertaram a quantidade de clusters em cada coleção.

O propósito desta comparação é verificar qual valor de *m* que produziu a melhor partição entre as partições com a quantidade correta de clusters avaliadas por um mesmo índice. Esta comparação foi realizada da seguinte forma: dada uma coleção, dos valores que cada índice obteve para os diferentes valores de *m*, aquele de maior ou menor valor, a depender se o índice é de maximização ou minimização, corresponderá a melhor partição. O melhor valor de *m* de cada índice para cada coleção será aquele que produziu a melhor partição (Tabela 4).

Para exemplificar, dada a coleção IAarticles e o índice P, Tabela 53 da Seção E em Anexos, pôde-se verificar que, dos valores de *m* que produziram as partições com número correto de clusters (*m = 10.0* e nos intervalos [1.5; 2.2], [6.5; 8.0]), aquela gerada com *m = 1.5* foi a melhor partição avaliada pelo índice P.

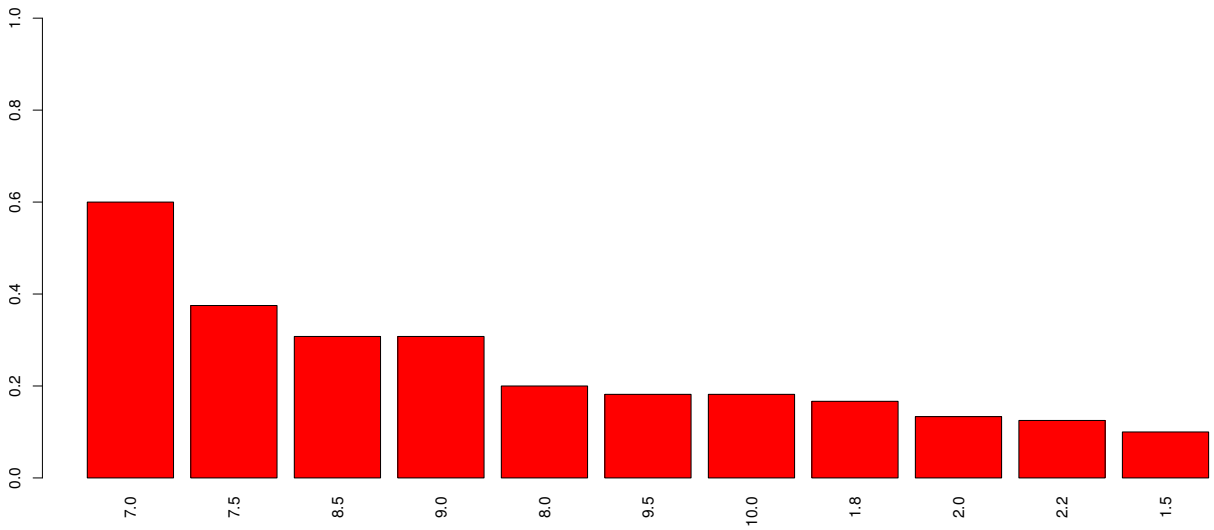
Tabela 4 – Valor de *m* mais indicado para cada índice e coleção

Coleção	P	MPC	SF	K	T	XB
NewYorkTimes	-	-	3.5	5.5	-	-
IAarticles	1.5	7.0	7.0	2.8	7.0	5.5
Opinosis	1.035	1.015	9.0	4.0	8.5	5.5
CSTR	1.03	1.035	9.0	9.0	8.5	2.2
SyskillWebert	1.01	1.01	9.0	1.03	9.5	1.03
Hitech	5.5	1.035	5.0	-	7.0	-
WAP	1.3	1.3	1.3	-	-	-
NSF	2.2	2.0	2.0	5.0	5.5	-
Irish-Sentiment	6.0	6.0	6.0	1.02	10.0	1.02
20Newsgroups	1.01	7.0	7.0	3.0	10.0	3.0
La1s	7.5	8.0	8.0	1.8	8.0	1.8
Reviews	8.5	1.045	8.5	7.5	9.5	7.5

A partir desta comparação, foi calculada uma média do número de vezes que um valor de *m*, nos intervalos [1.5; 2.2] e [7.0; 10.0], gerou a melhor partição pela quantidade total em que o mesmo valor de *m* acertou a quantidade de clusters em todas as coleções e para os seis índices. Isto é, para o valor de $m = 7.0$ a média de 0.6 (Figura 21) foi dada pelo número de vezes que apareceu na Tabela 4, neste caso seis, pelo número total de acertos, 10 (4 acertos na IAarticles, 3 na 20Newsgroups e 1 nas coleções Opinosis, SyskillWebert e Hitech).

Na Figura 21 são apresentadas as médias calculadas para cada valor de *m* ordenados de maneira decrescente. Com esta análise, o fator de fuzzificação igual a 7.0 apresentou as partições de melhor qualidade com quase o dobro do valor obtido para $m = 7.5$. Já o valor de $m = 2.2$, que foi o melhor avaliado anteriormente por ter obtido maior média de acerto, nesta análise, obteve média igual a 0.125, sendo considerada a melhor partição somente quando foi o único valor de *m* a acertar a quantidade de clusters (NSF com o índice P e CSTR para XB).

Figura 21 – Média de melhores partições por valor de *m*



A partir das médias da Figura 21 pôde-se perceber que os valores de *m* no intervalo de [7.0; 10.0] obtiveram maior número de partições com a quantidade correta de clusters do que o intervalo [1.5; 2.2] comumente indicado como mais adequado para o algoritmo FCM.

5.3 Considerações Finais

Portanto, baseado na análise realizada a cerca da média de acerto obtida ao utilizar os valores de *m* no intervalo [1.01; 10.0], é possível indicar um intervalo geral de [7.0; 10.0], como mais adequado para agrupamento fuzzy de bases textuais de alta dimensionalidade, respondendo a segunda questão levantada na Seção 4.2: “Qual o valor de fator de fuzzicação *m* ou intervalo em que este esteja inserido que resultará em um melhor desempenho do FCM e dos índices de validação de fuzzy ao agrupar bases de alta dimensionalidade ¿‘

6 CONCLUSÃO

As execuções do algoritmo Fuzzy C-Means e dos índices de validação fuzzy para agrupamento de bases textuais apresentam peculiaridades devido a alta dimensionalidade das coleções. Neste trabalho, foram investigadas as possíveis causas destas peculiaridades além de indicar os índices e o intervalo de valor do fator de fuzzificação m que melhor se adéquam às bases textuais agrupadas com o FCM.

Para esta investigação, 12 coleções de documentos foram agrupadas com o FCM utilizando 32 valores para o parâmetro m e 16 índices de validação avaliaram as partições geradas.

A partir dos resultados obtidos, os índices e valores de m identificados com melhor desempenho foram aqueles que apresentaram maior média de acerto em relação a quantidade de clusters de uma coleção, além de não apresentarem uma tendência expressiva em escolher como número ótimo c de clusters o valor mínimo ($c = c(min)$) ou máximo ($c = c(max)$).

Com as análises discutidas na Seção 5.2, foi possível encontrar as seguintes respostas para as três questões levantadas na Seção 4.2:

1) A alta dimensionalidade das bases textuais utilizadas pode ser considerada como o principal agente do desempenho inferior dos índices;

2) A partir dos resultados obtidos, dos índices testados, o P, MPC, SF, K e T obtiveram melhor desempenho sendo assim sugeridos para serem utilizados ao avaliar agrupamento fuzzy de alta dimensionalidade;

3) O intervalo [7.0; 10.0] de valores de m foi o mais indicado ao se agrupar bases de alta dimensionalidade utilizando o FCM.

Além destas questões, também foi possível identificar os valores de m que se aproximaram do infinito para os índices PC, PE ($m \geq 1.04$) e MPC ($m \geq 1.045$) que têm os valores dos seus limites previamente definidos (Seções 5.1.1, 5.1.2 e 5.1.3).

Neste trabalho, os índices de validação não apresentaram o mesmo desempenho visto na literatura (CAMPELLO; HRUSCHKA, 2006; WANG; ZHANG, 2007; DAVE, 1996; PAL; BEZDEK, 1995) ao avaliar as partições geradas pelo FCM. No contexto em que o trabalho está inserido (agrupamento fuzzy, utilizando o algoritmo Fuzzy C-Means, de bases textuais estruturadas por meio de uma matriz Documento x Termo como definido na Seção 2.1), é possível afirmar que a principal razão deste desempenho deve-se a alta dimensionalidade das bases.

Em (PAL; BEZDEK, 1995) Bezdek cita as razões pela qual, se uma base é bem estruturada, ou seja, se seus objetos (documentos) são bem separados em c reconhecíveis clusters,

como é o caso de todas as coleções utilizadas, já que as mesmas tiveram o número de classes definido por especialistas, o porquê desta estrutura não ser descoberta. São elas:

1) A representação numérica dos objetos pode não possuir informação adequada para diferenciar os clusters;

2) O algoritmo utilizado não consegue extrair a estrutura da base por ser, por exemplo, um algoritmo que busca somente por clusters hiperesféricos;

3) Os parâmetros apropriados do algoritmo que produzem uma interpretação bem-sucedida da base nunca são usados;

4) Se a representação numérica for adequada, o algoritmo for capaz de encontrar a forma compatível dos clusters e os parâmetros utilizados no mesmo estiverem apropriados, os índices de validação podem falhar ao indicar que os bons clusters definidos pelos especialistas são realmente bons.

Neste trabalho, o item 3 foi investigado com os parâmetros do FCM sendo testados a fim de encontrar os valores que produziriam os melhores resultados ao agrupar bases textuais de alta dimensionalidade.

A alta dimensionalidade só poderá ser considerada como única causa do mau desempenho dos índices de validação se as bases textuais forem representadas de uma outra maneira numérica, se diferentes algoritmos de agrupamento fuzzy forem executados, sendo necessário testar os valores dos parâmetros de cada algoritmo a fim de encontrar os seus valores mais adequados, e se forem avaliados por mais índices de validação, além dos 16 aqui estudados.

Se a tarefa não supervisionada de agrupar bases de dados de maneira crisp já é uma tarefa difícil, agrupar bases textuais de alta dimensionalidade de maneira fuzzy pode se tornar ainda mais difícil. Bezdek declarou em (PAL; BEZDEK, 1995): “no matter how good your index is, there is a data set out there waiting to trick it (and you).” [1] Inspirado nesta declaração pode-se adicionar: “ainda se seu índice for bom em avaliar qualquer base de dados, existe uma base de alta dimensionalidade esperando para enganá-lo (e você)“.

6.1 Produções

Duas produções serão publicadas ao utilizar os índices de validação de agrupamento fuzzy para avaliar a organização flexível de bases textuais. O primeiro artigo intitulado: *On Fuzzy Cluster Validity Indexes for High Dimensional Feature Space* foi aceito e será publicado na *Conferência da Sociedade Europeia de Lógica Fuzzy e Tecnologia* (Conference of the European Society for Fuzzy Logic and Technology - EUSFLAT 2017) (EUSTÁQUIO et al., 2017, No

[1] Tradução: “não importa quão bom seu índice seja, existe uma base de dados esperando para enganá-lo (e você)“

prelo). Neste artigo foram executados os índices PC, PE, MPC, FS, XB e SF a fim de escolher o índice que apresentaria a melhor partição resultando em uma melhor extração de descritores dos clusters. O segundo artigo intitulado: *Influência de Técnicas Não-supervisionadas de Redução de Dimensionalidade para Organizacão Flexível de Documentos* foi aceito e será publicado no *Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana* (Brazilian Symposium in Information and Human Language Technology - STIL) (LIMA; EUSTÁQUIO; RIOS, 2017, No prelo). Neste artigo os índices de validação foram utilizados para avaliar e comparar o desempenho obtido pelo FCM ao agrupar bases pré-processadas por duas técnicas distintas.

REFERÊNCIAS

BEZDEK, J. C. Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, v. 1, n. 1, p. 57 – 71, 1974. ISSN 1432-1416.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

BEZDEK†, J. C. Cluster Validity with Fuzzy Sets. *Journal of Cybernetics*, v. 3, n. 3, p. 58 – 73, 1974.

CAMPELLO, R.; HRUSCHKA, E. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, v. 157, n. 21, p. 2858 – 2875, 2006. ISSN 0165-0114.

CARVALHO, N. V. et al. Flexible Document Organization by Mixing Fuzzy and Possibilistic Clustering algorithms. In: *IEEE International Conference on Fuzzy Systems*. [S.l.: s.n.], 2016. p. 790 – 797.

CHEN, M.; LINKENS, D. Rule-base self-generation and simplification for data-driven fuzzy models. *Fuzzy Sets and Systems*, v. 142, n. 2, p. 243 – 265, 2004. ISSN 0165-0114. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016501140300160X>.

DAVE, R. N. Validating Fuzzy Partitions Obtained Through C-shells Clustering. *Pattern Recognition Letter*, v. 17, n. 6, p. 613 – 623, may 1996. ISSN 0167-8655.

EUSTÁQUIO, F. et al. On Fuzzy Cluster Validity Indexes for High Dimensional Feature Space. In: CONFERENCE OF THE EUROPEAN SOCIETY FOR FUZZY LOGIC AND TECHNOLOGY - EUSFLAT, 2017, No prelo. [S.l.], 2017, No prelo.

FUKUYAMA, Y.; SUGENO, M. A new method of choosing the number of clusters for fuzzy c-means method. In: *Fuzzy systems Symposium*. [S.l.: s.n.], 1989. p. 247 – 250.

GAMA, J. et al. *Inteligência artificial*: uma abordagem de aprendizado de máquina. Grupo Gen - LTC, 2011. ISBN 9788521618805. Disponível em: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

GATH, I.; GEVA, A. B. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 11, n. 7, p. 773 – 780, 1989. ISSN 0162-8828.

HU, Y. et al. A robust cluster validity index for fuzzy c-means clustering. In: *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*. [S.l.: s.n.], 2011. p. 448 – 451.

JOOPUDI, S. et al. A New Cluster Validity Index for Fuzzy Clustering. *IFAC Proceedings Volumes*, v. 46, n. 32, p. 325 – 330, 2013. ISSN 1474-6670. 10th IFAC International Symposium on Dynamics and Control of Process Systems. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1474667015382781>.

KAUFMAN, L.; ROUSSEEUW, P. *Finding Groups in Data*: An Introduction to Cluster Analysis. Wiley, 2005. (Wiley Series in Probability and Statistics). ISBN 9780471735786. Disponível em: <https://books.google.com.br/books?id=yS0nAQAAIAAJ>.

KIM, Y. et al. A cluster validation index for GK cluster analysis based on relative degree of sharing. *Information Sciences*, v. 168, n. 1, p. 225 – 242, 2004. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025504000209>.

KWON, S. H. Cluster validity index for fuzzy clustering. *Electronics Letters*, v. 34, n. 22, p. 2176 – 2177, 1998. ISSN 0013-5194.

LIMA, B.; EUSTÁQUIO, F.; RIOS, T. Influência de Técnicas Não-supervisionadas de Redução de Dimensionalidade para Organização Flexível de Documentos. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY - STIL, 2017, No prelo. [S.l.], 2017, No prelo.

LUCIEER, V.; LUCIEER, A. Fuzzy clustering for seafloor classification. *Marine Geology*, v. 264, n. 3, p. 230 – 241, 2009. ISSN 0025-3227. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0025322709001431>.

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, v. 37, n. 3, p. 487 – 501, 2004. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320303002838>.

PAL, N. R.; BEZDEK, J. C. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, v. 3, n. 3, p. 370 – 379, 1995. ISSN 1063-6706.

RAWASHDEH, M.; RALESCU, A. L. Fuzzy Cluster Validity with Generalized Silhouettes. In: *MAICS*. [S.l.: s.n.], 2012.

RIOS, T. N. *Organização flexível de documentos*. 2013. 128 p. Tese (Ciências de Computação e Matemática Computacional) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

RIOS, T. N.; REZENDE, S. O.; CAMARGO, H. A. Flexible Document Organization: Comparing Fuzzy and Possibilistic Approaches. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015.

SILVA, L. R. S. da; GOMIDE, F. Um estudo comparativo entre as funções de validação para agrupamento nebuloso de dados. In: IV CONGRESSO BRASILEIRO DE COMPUTAÇÃO, 2004. [S.l.], 2004. p. 266 – 270.

TANG, Y.; SUN, F.; SUN, Z. Improved validation index for fuzzy clustering. In: *Proceedings of the 2005, American Control Conference, 2005.* [S.l.: s.n.], 2005. p. 1120–1125 – 2. ISSN 0743-1619.

VALENTE, R. X. *Uma Avaliação da Utilização de Matrizes de Afinidades na Validação de Agrupamentos de Dados*. 2013. 71 p. Dissertação (Engenharia Elétrica) — Universidade Federal de Minas Gerais, Belo Horizonte.

VENDRAMIN, L. *Estudo e desenvolvimento de algoritmos para agrupamento fuzzy de dados em cenários centralizados e distribuídos*. 2012. 138 p. Dissertação (Ciência da Computação e Matemática Computacional) — Instituto de Ciências Matemáticas e de Computação - ICMC - USP, São Carlos.

WANG, W.; ZHANG, Y. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, v. 158, n. 19, p. 2095 – 2117, 2007.

WU, K.; YANG, M. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, v. 26, n. 9, p. 1275 – 1291, 2005. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865504003629>.

XIE, X. L.; BENI, G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 13, n. 8, p. 841 – 847, Aug 1991. ISSN 0162-8828.

ZAHID, N.; LIMOURI, M.; ESSAID, A. A new cluster-validity for fuzzy clustering. *Pattern Recognition*, v. 32, n. 7, p. 1089 – 1097, 1999. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320398001575>.

ZHANG, F.; QIAN, X. A New Validity Index for Fuzzy Clustering. *Journal of Computational Information Systems*, 2012.

Anexos

Neste anexo são apresentados os resultados obtidos por cada índice de validação ao avaliar as partições geradas com fator de fuzzificação no intervalo [1.01; 10.0] para todas as coleções de documentos.

ANEXO A – PC

Tabela 5 – NewYorkTimes

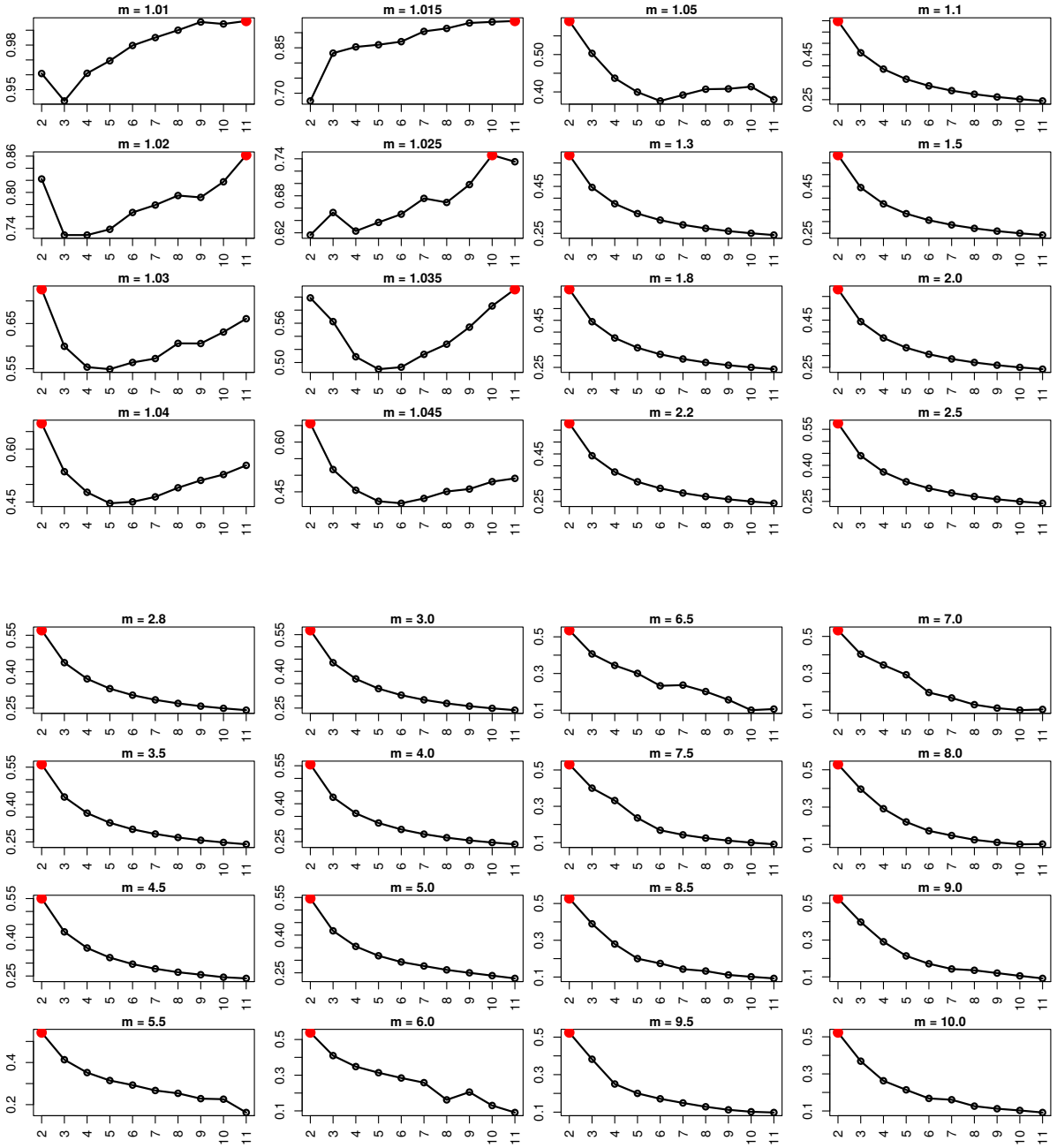


Tabela 6 – IAarticles

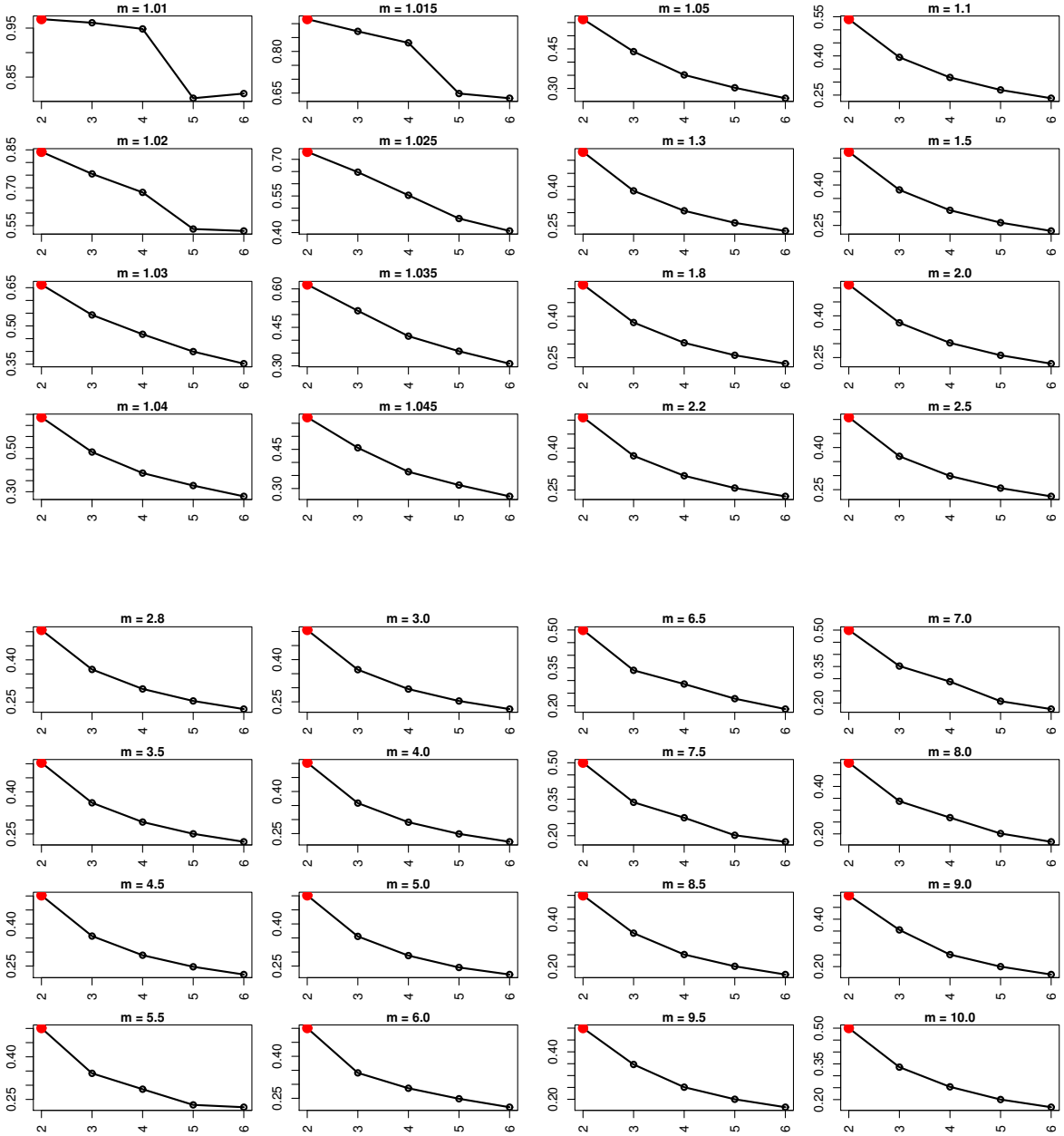


Tabela 7 – Opinosis

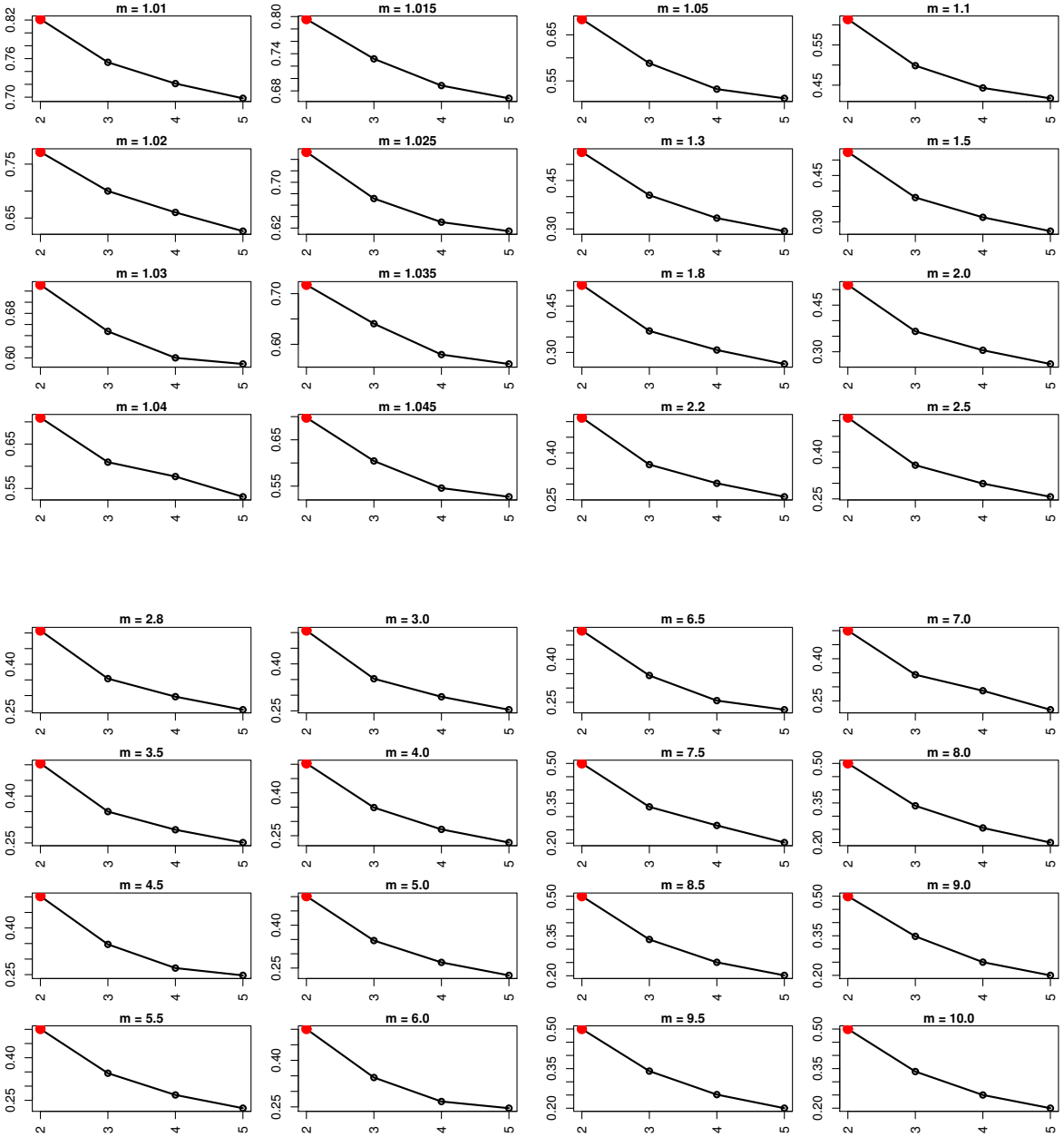


Tabela 8 – CSTR

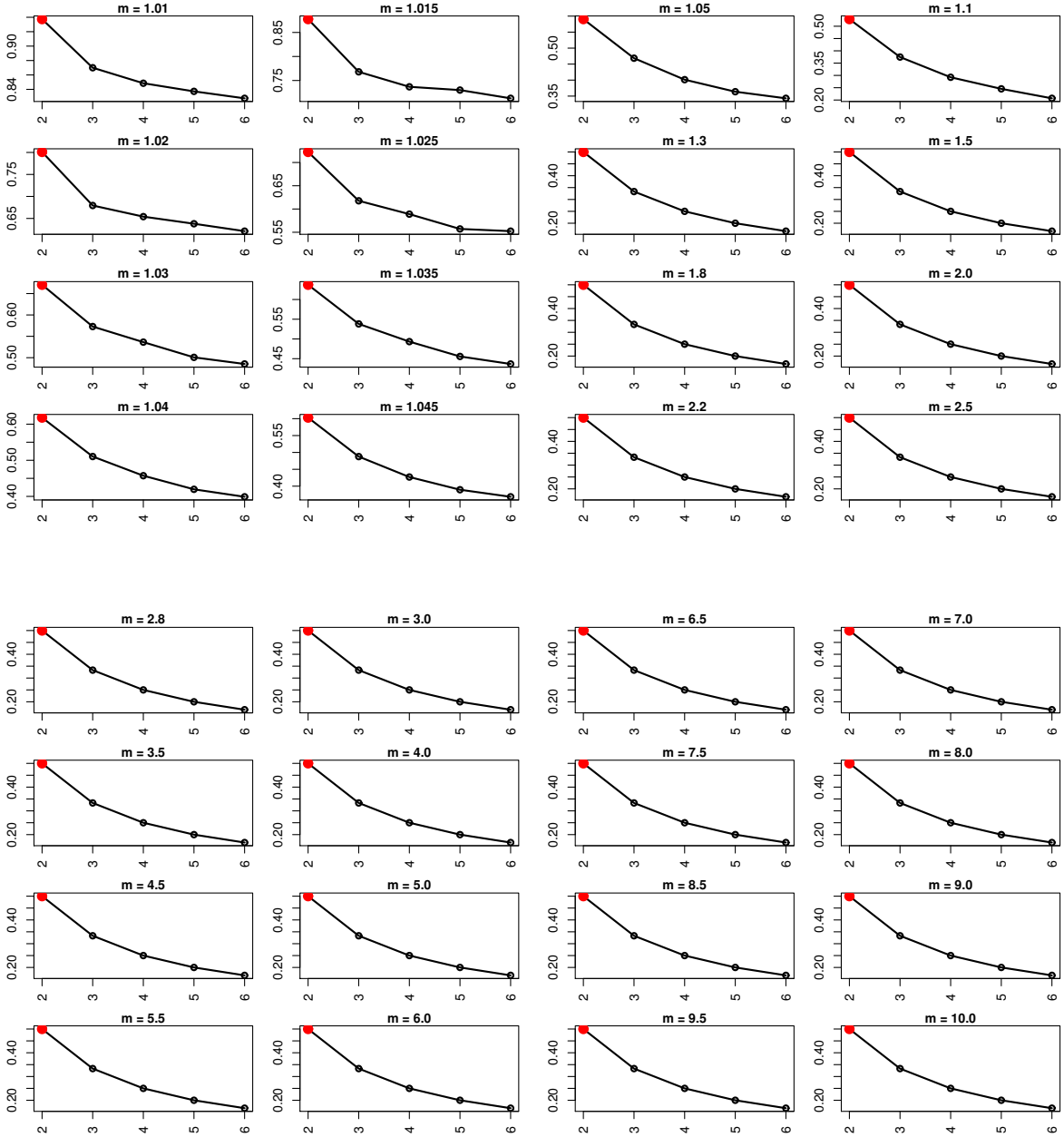


Tabela 9 – SyskillWebert

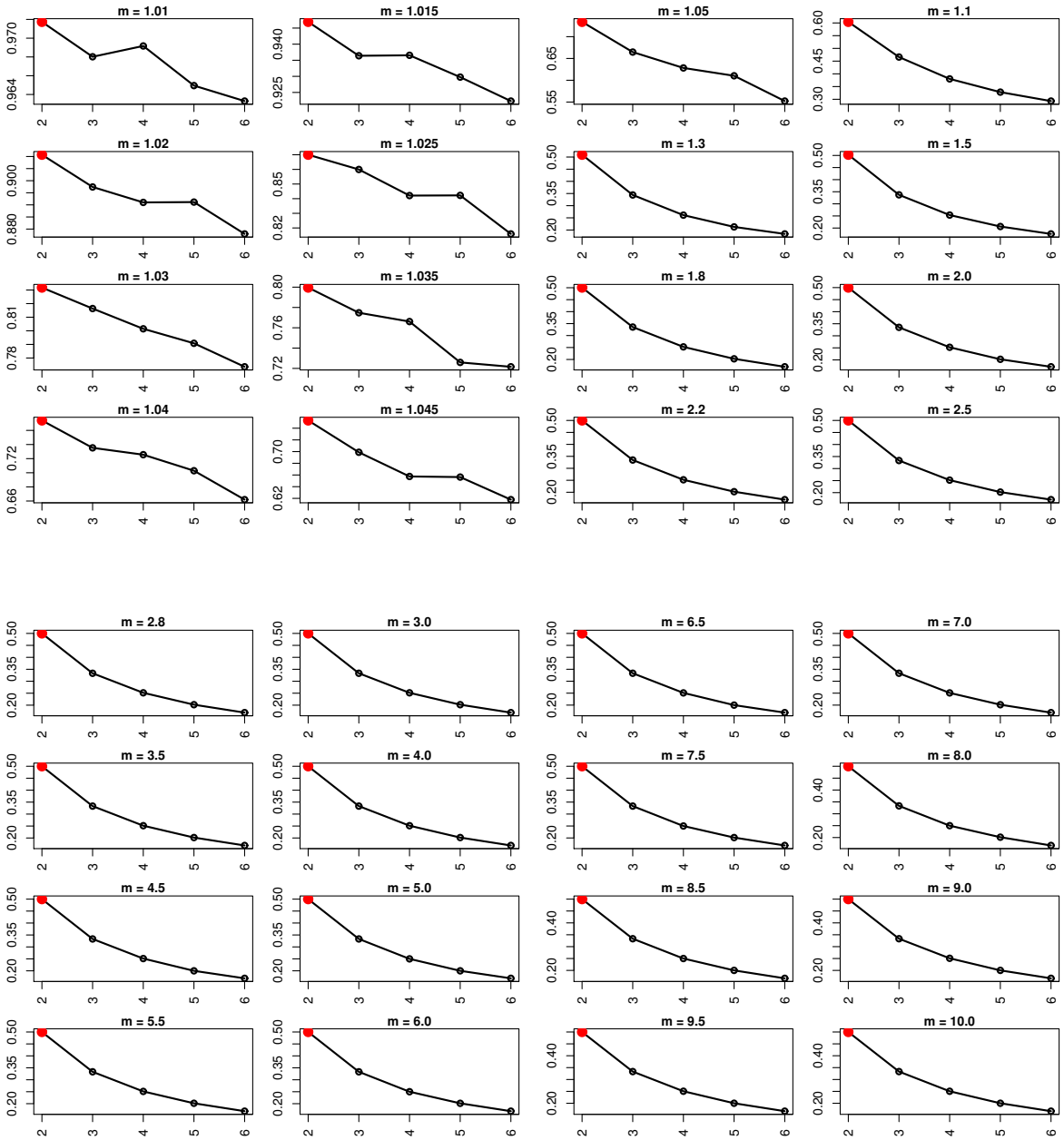


Tabela 10 – Hitech

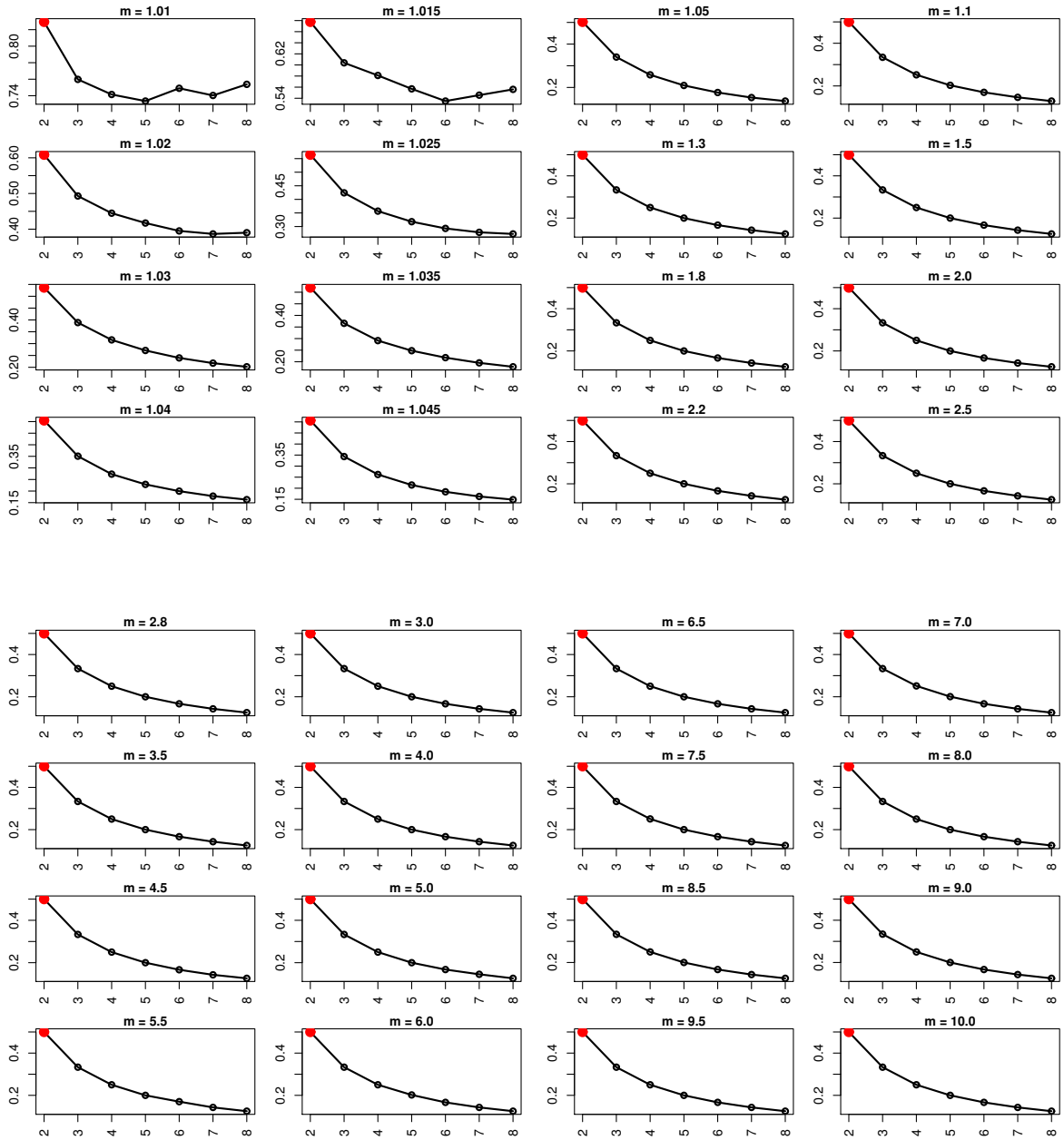


Tabela 11 – WAP

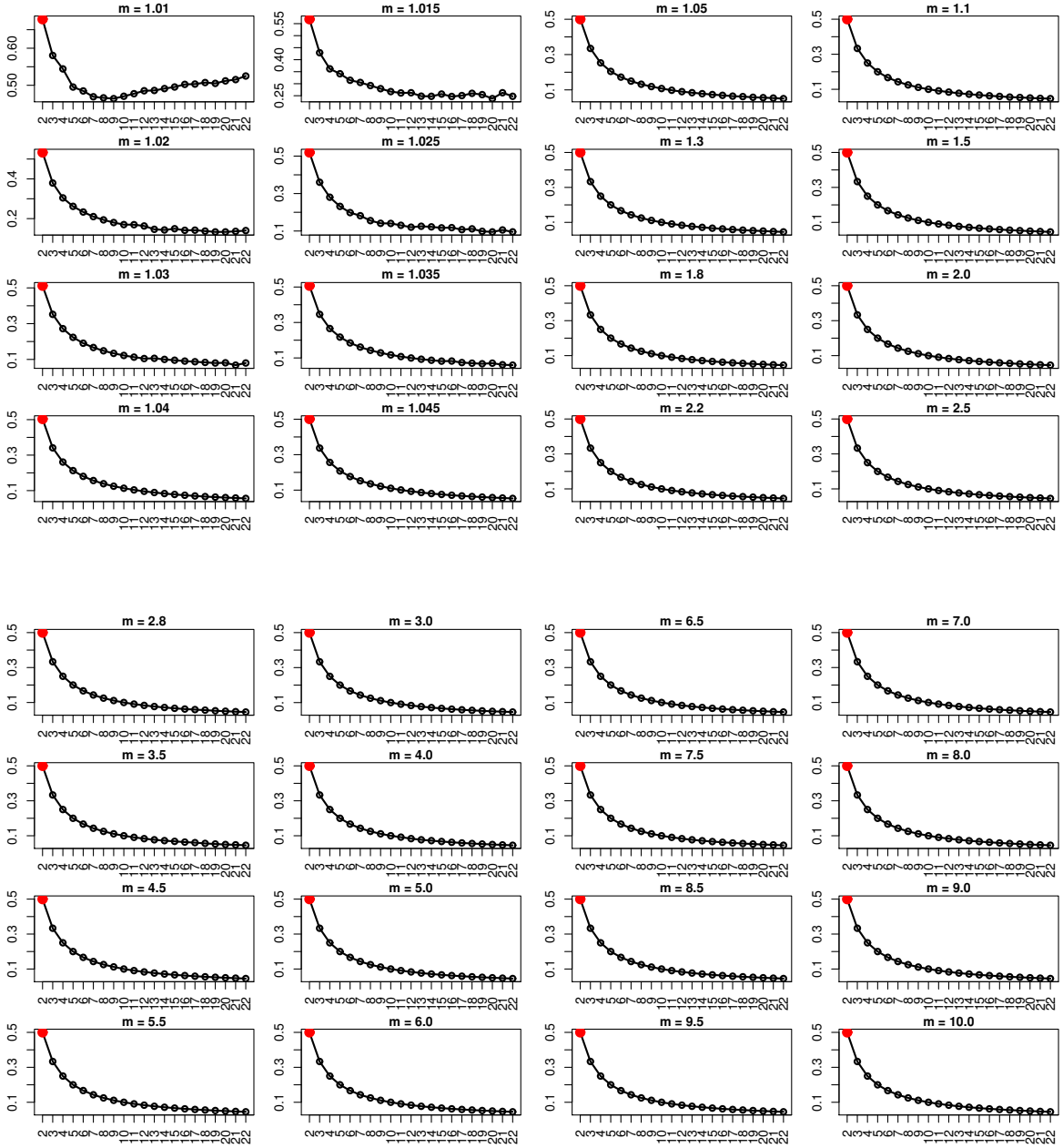


Tabela 12 – NSF

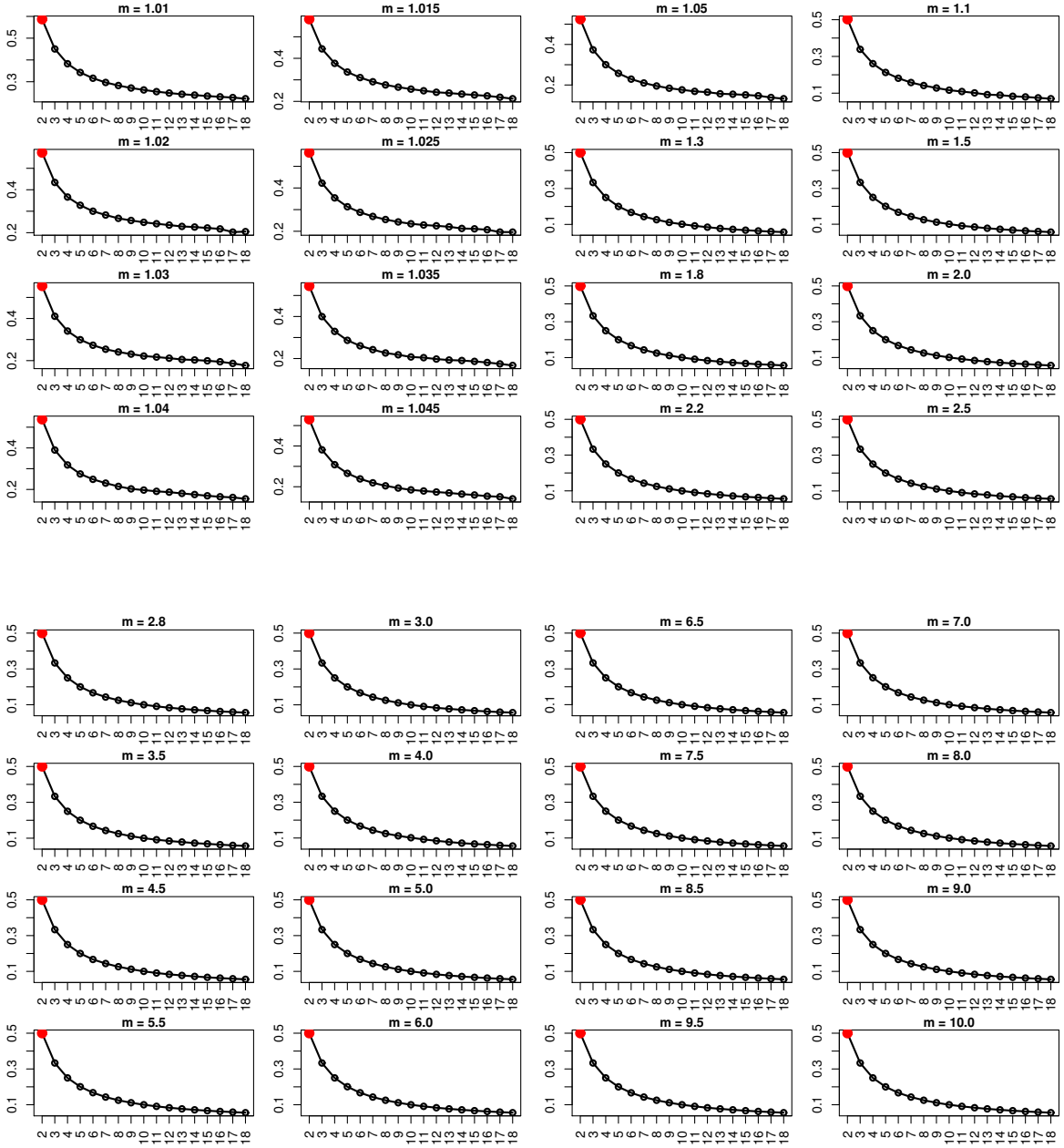


Tabela 13 – Irish-Sentiment

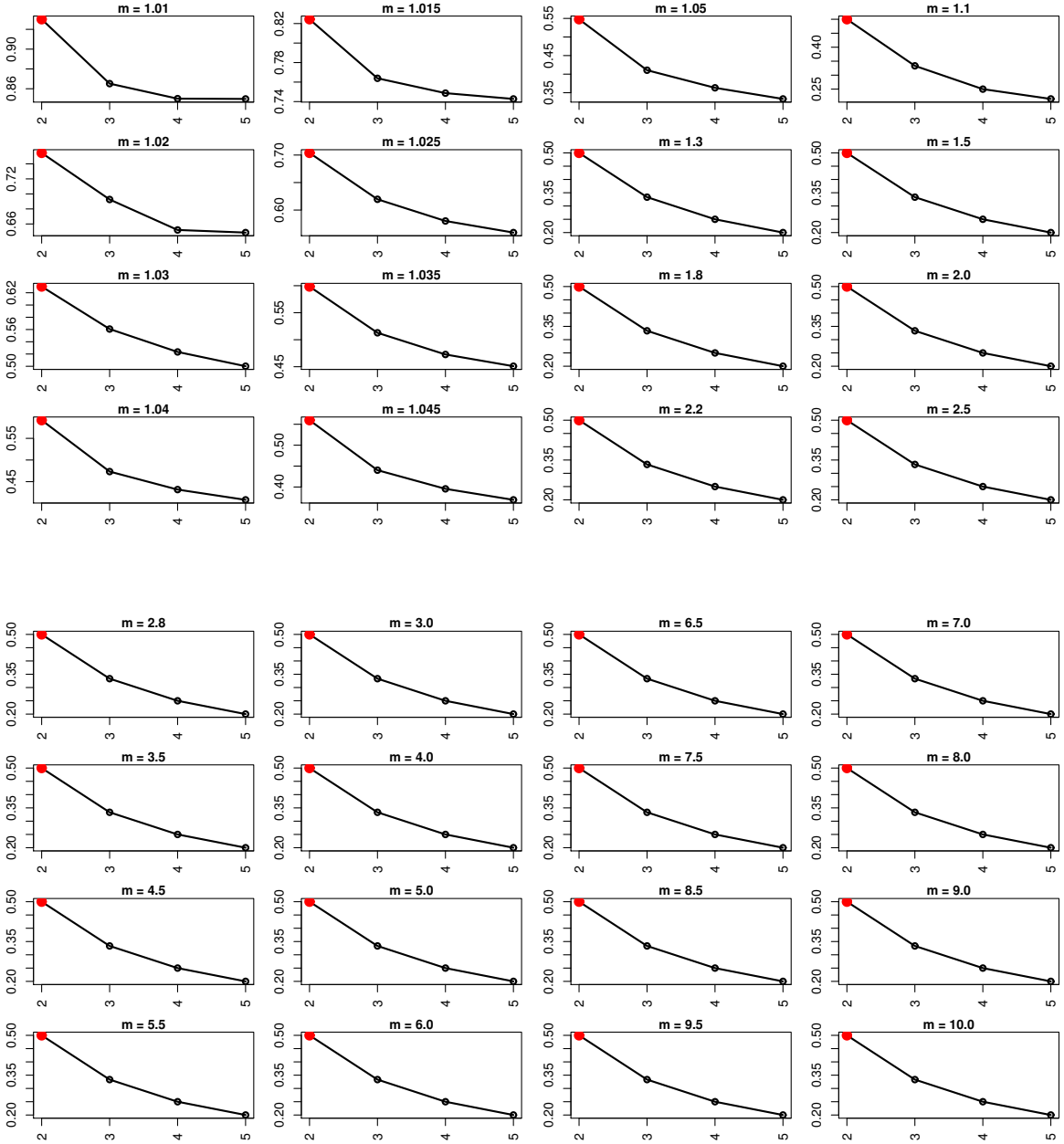


Tabela 14 – 20Newsgroups

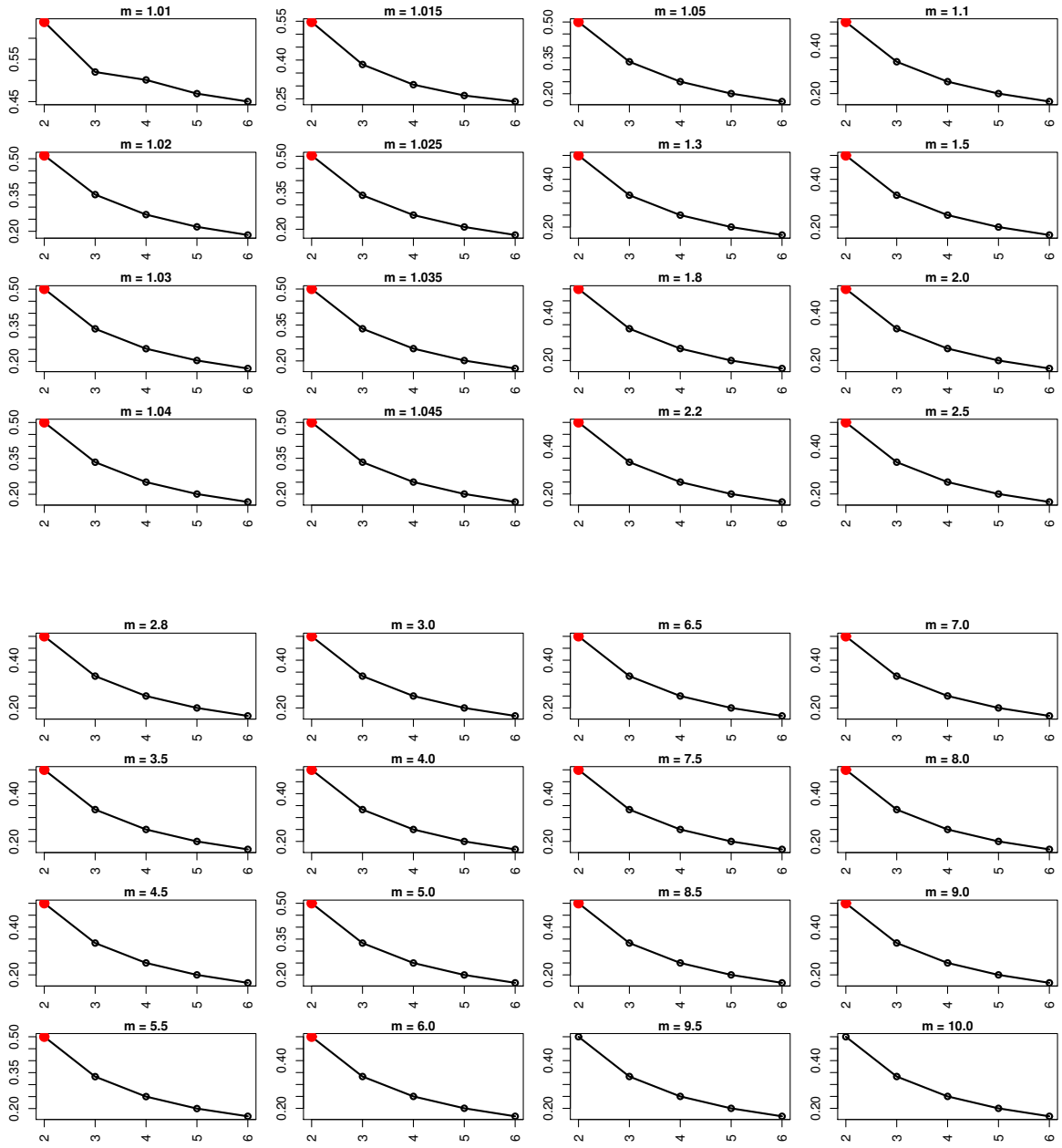


Tabela 15 – La1s

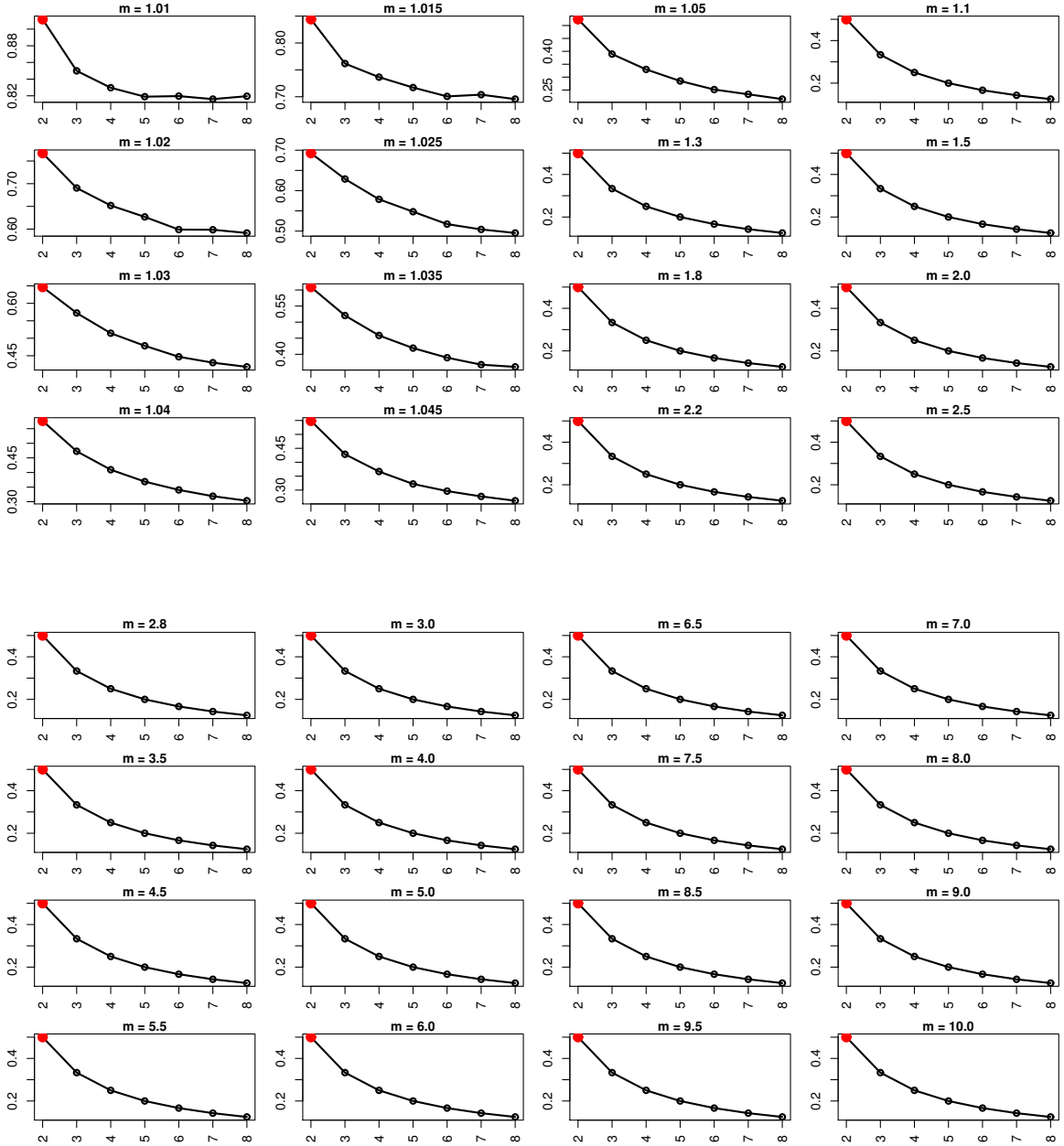
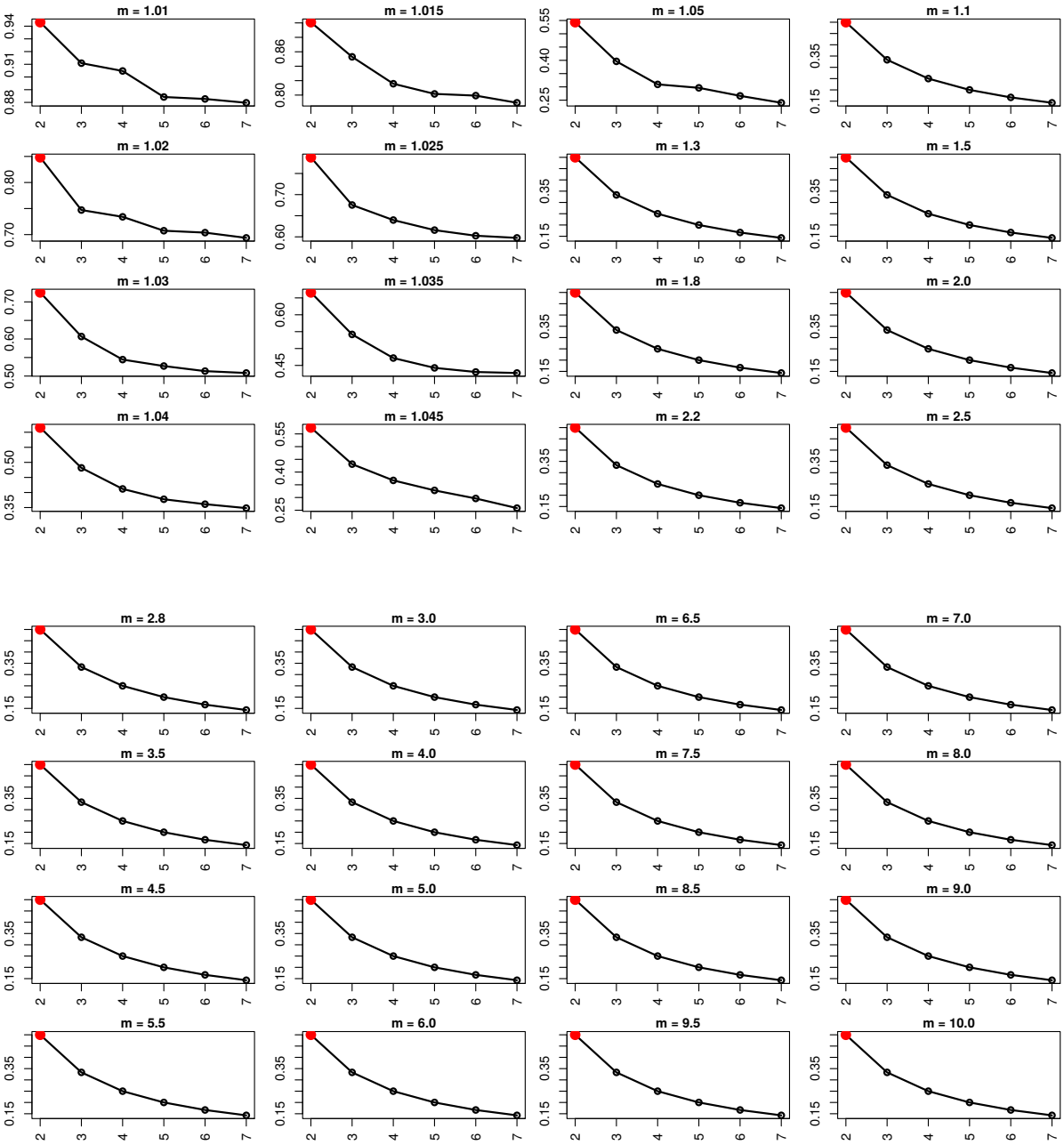


Tabela 16 – Reviews



ANEXO B – PE

Tabela 17 – NewYorkTimes

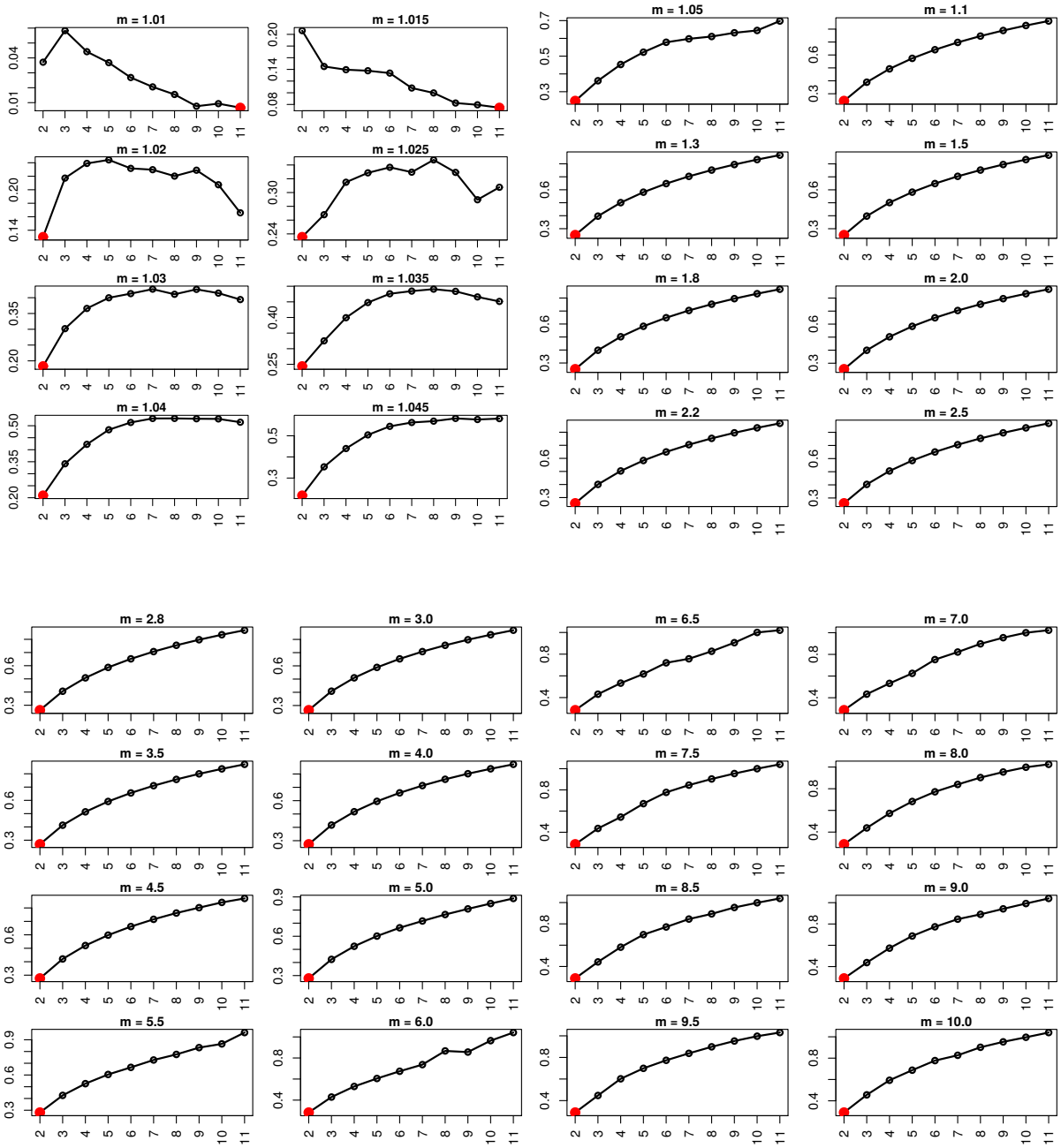


Tabela 18 – IAarticles

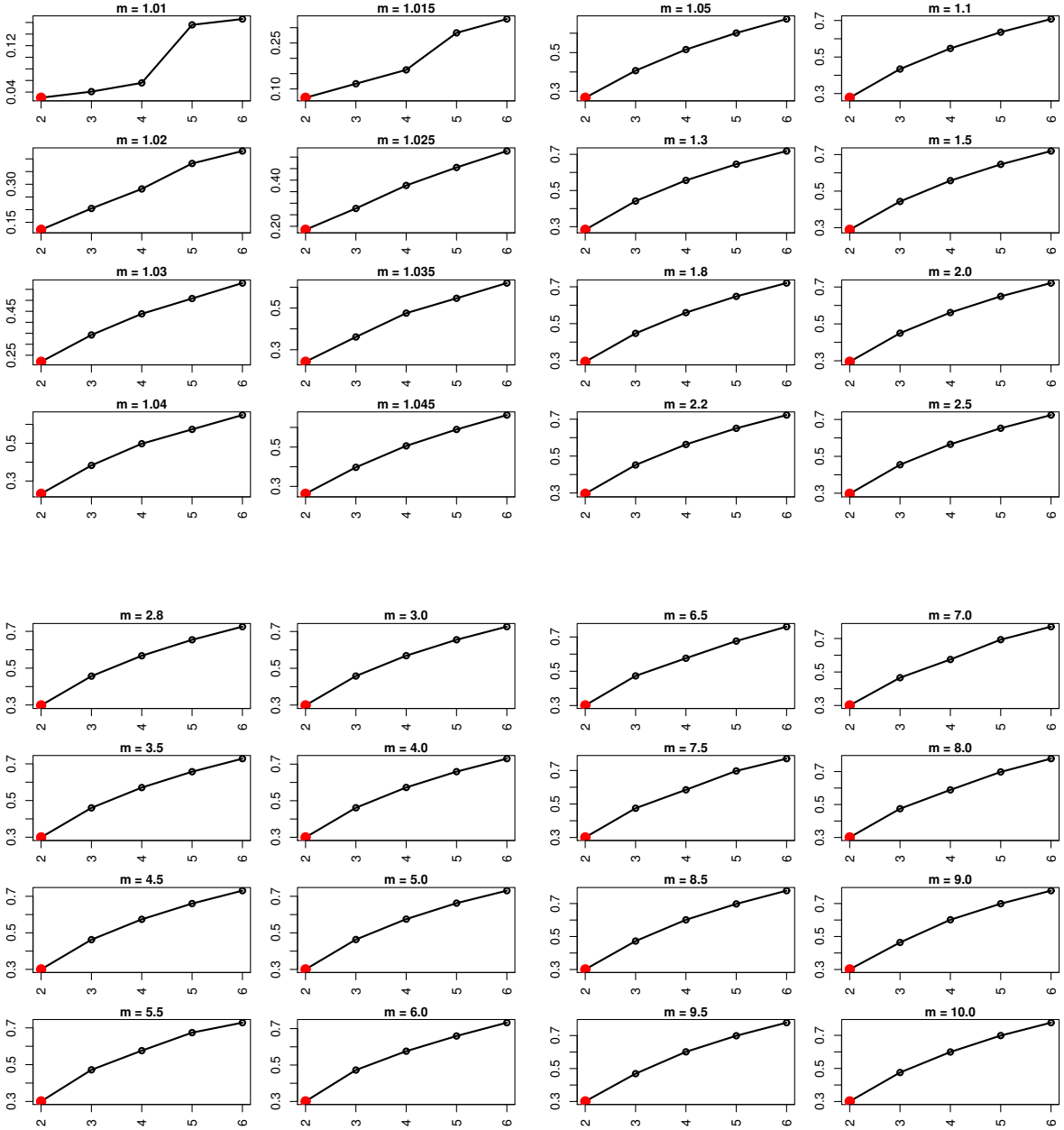


Tabela 19 – Opinosis

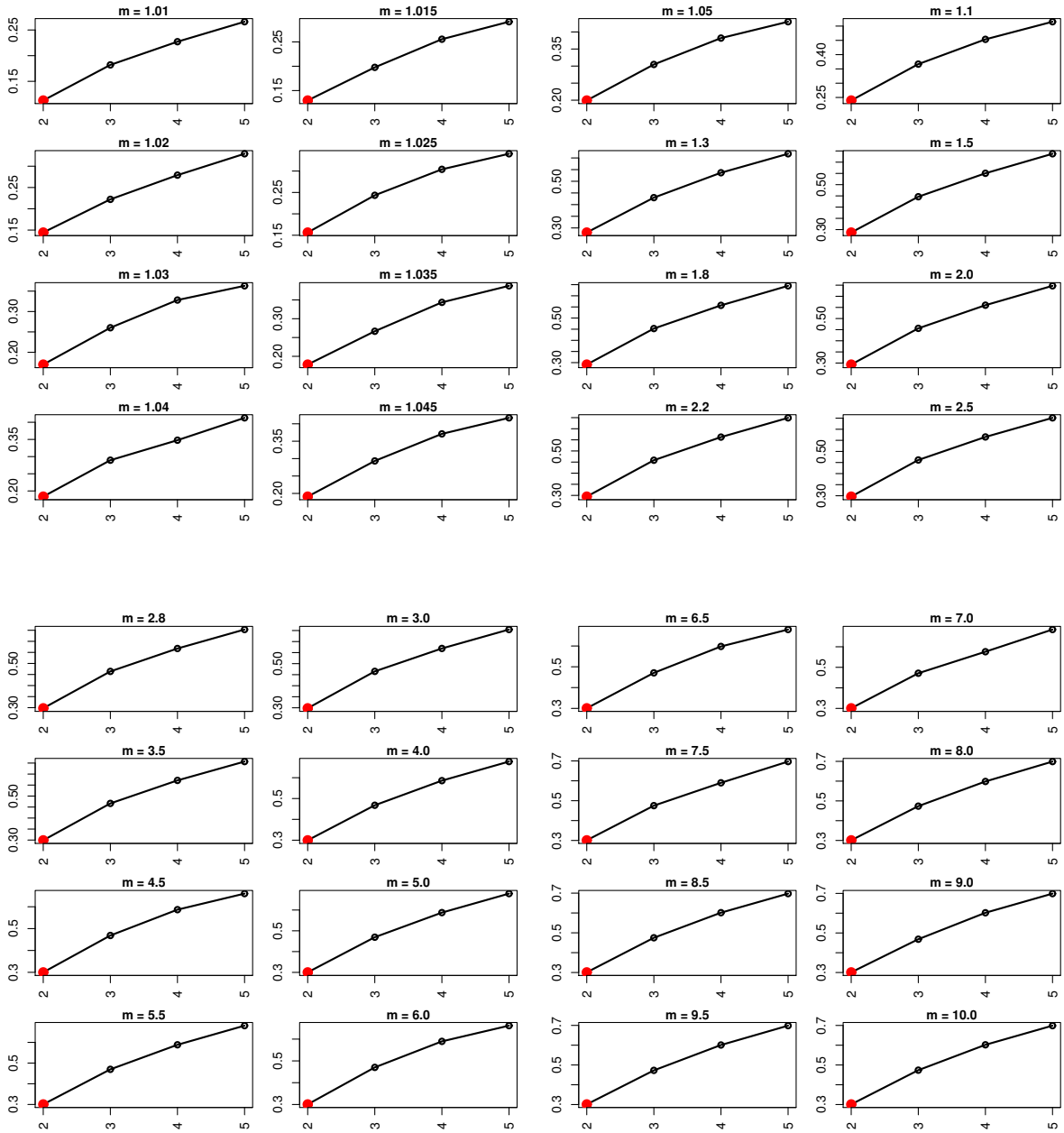


Tabela 20 – CSTR

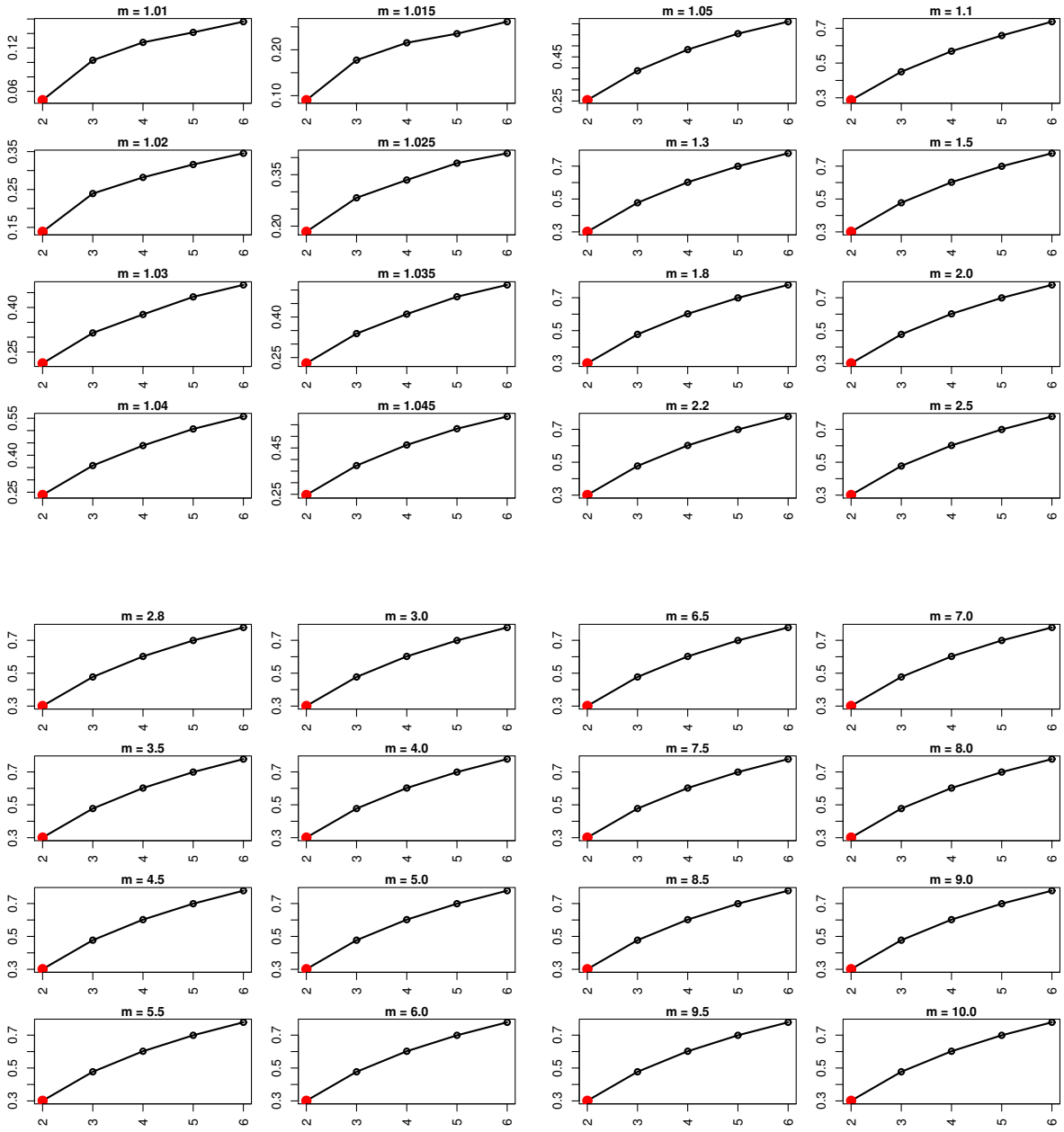


Tabela 21 – SyskillWebert

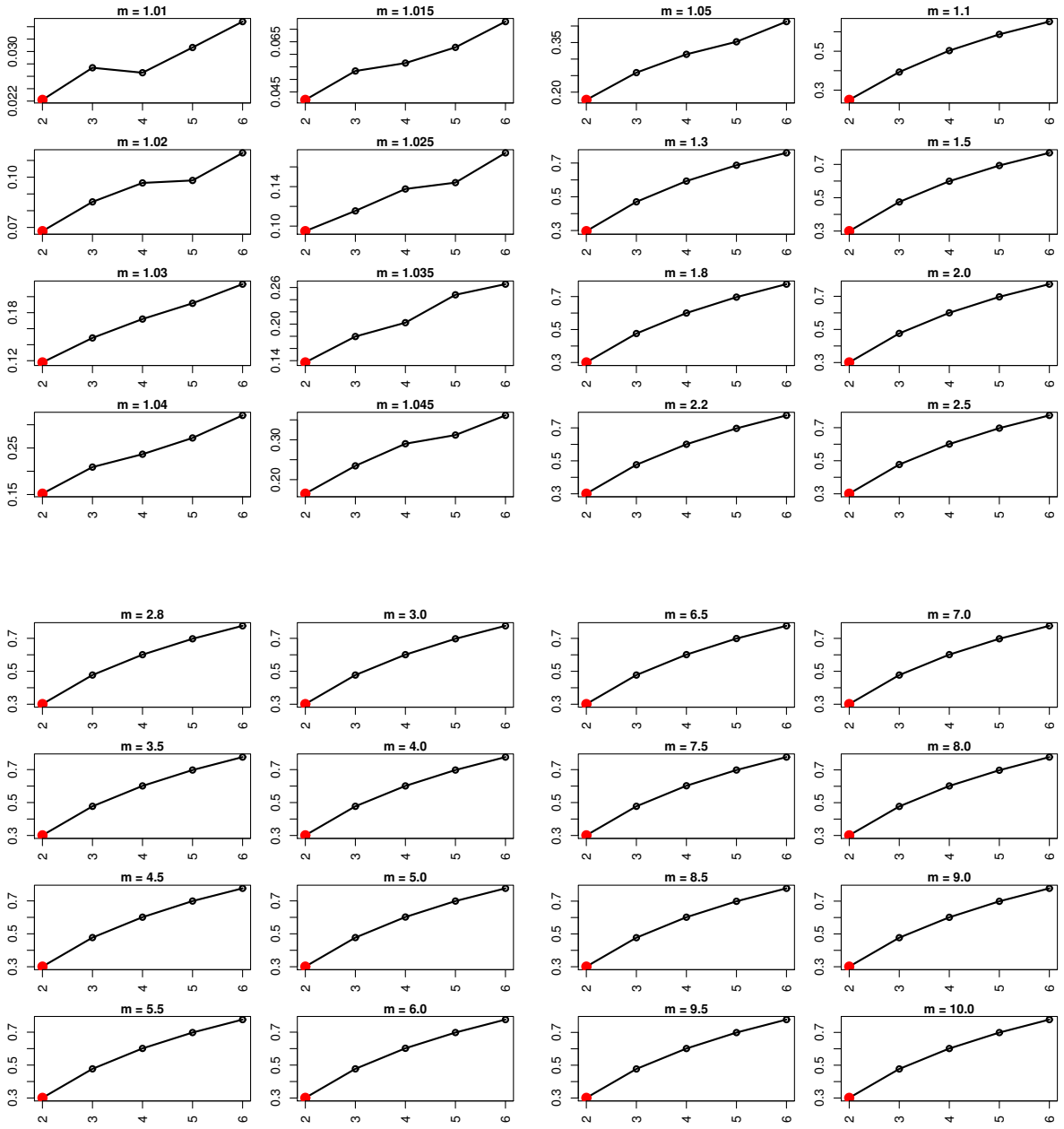


Tabela 22 – Hitech

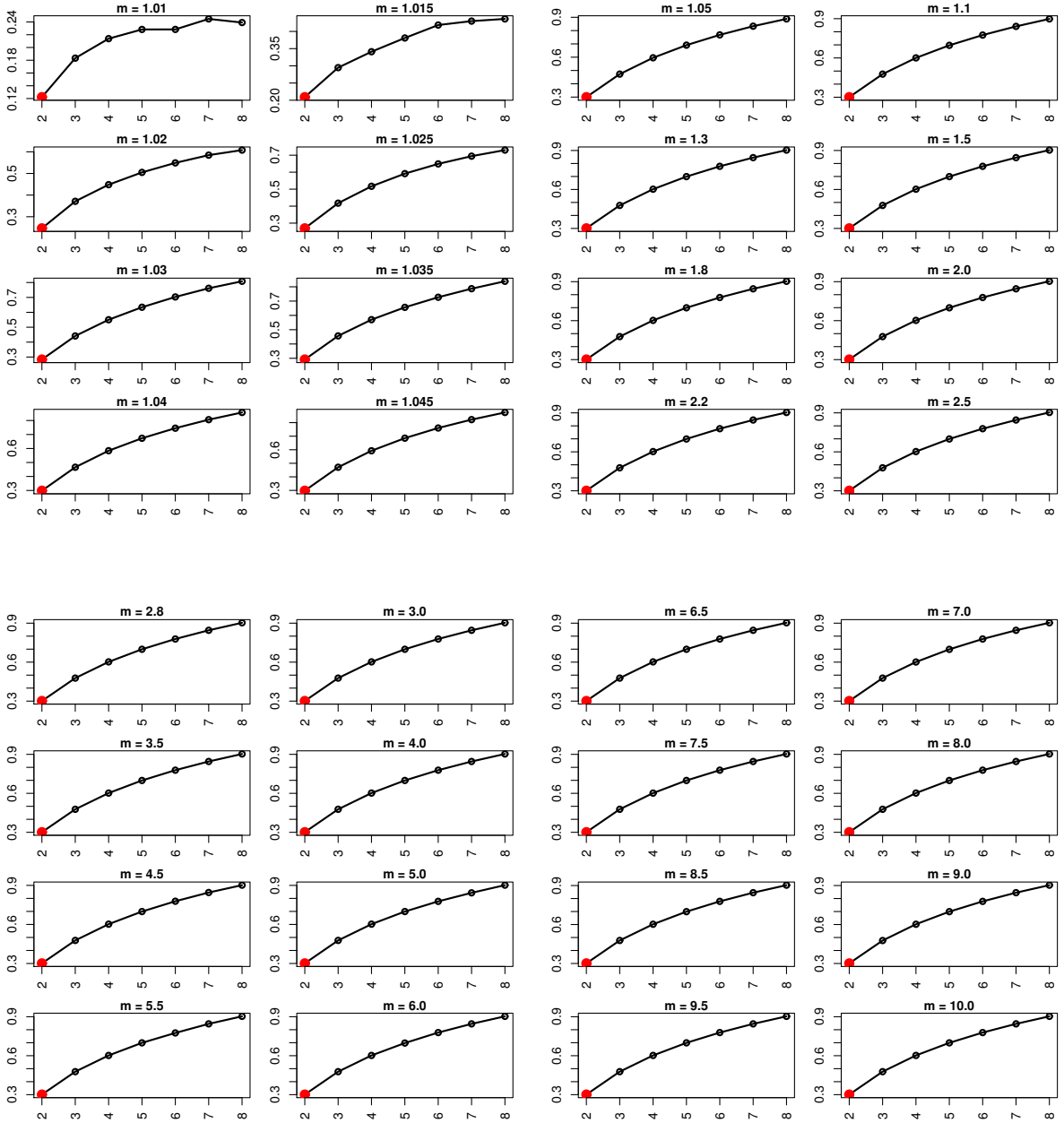


Tabela 23 – WAP

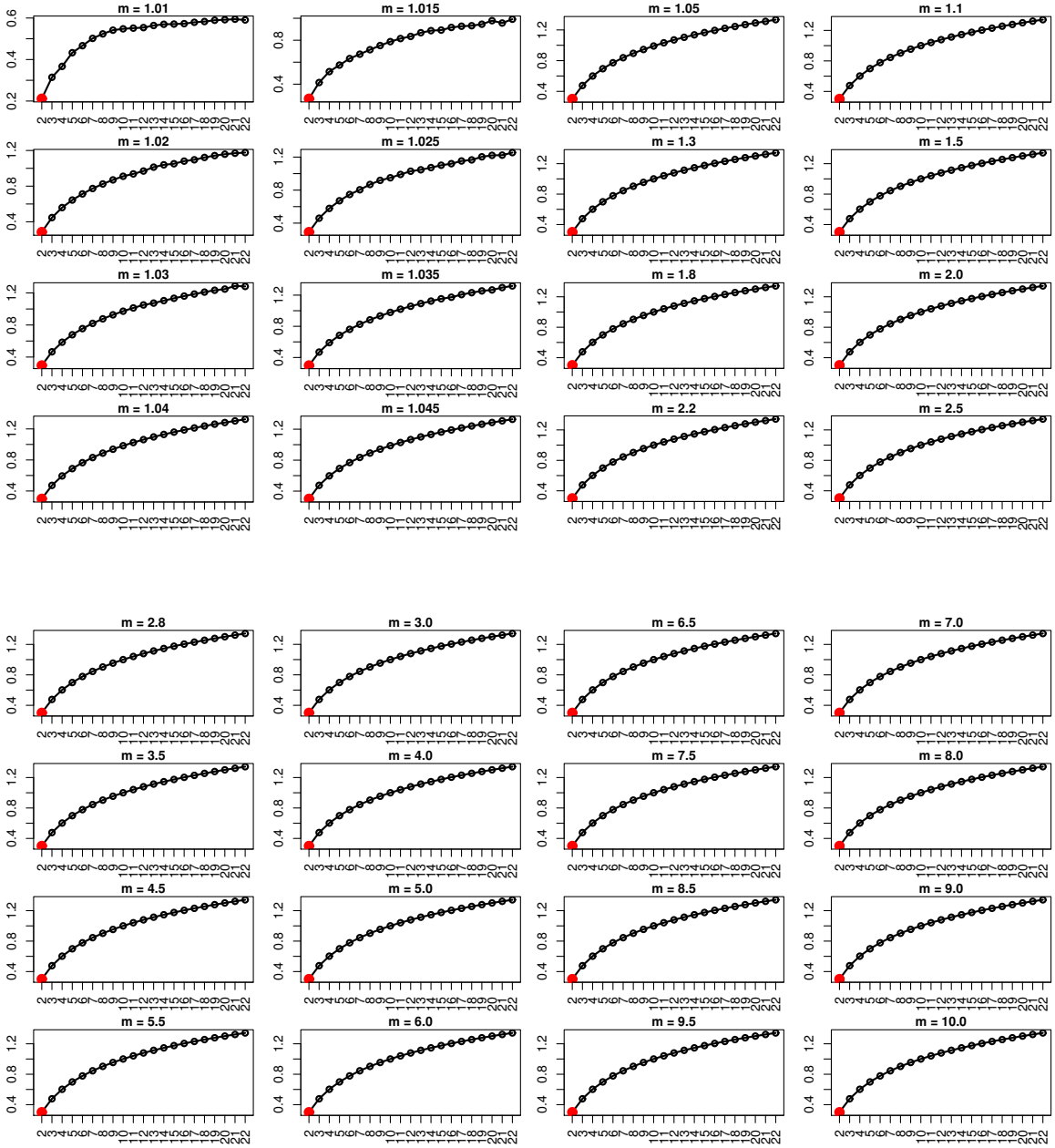


Tabela 24 – NSF

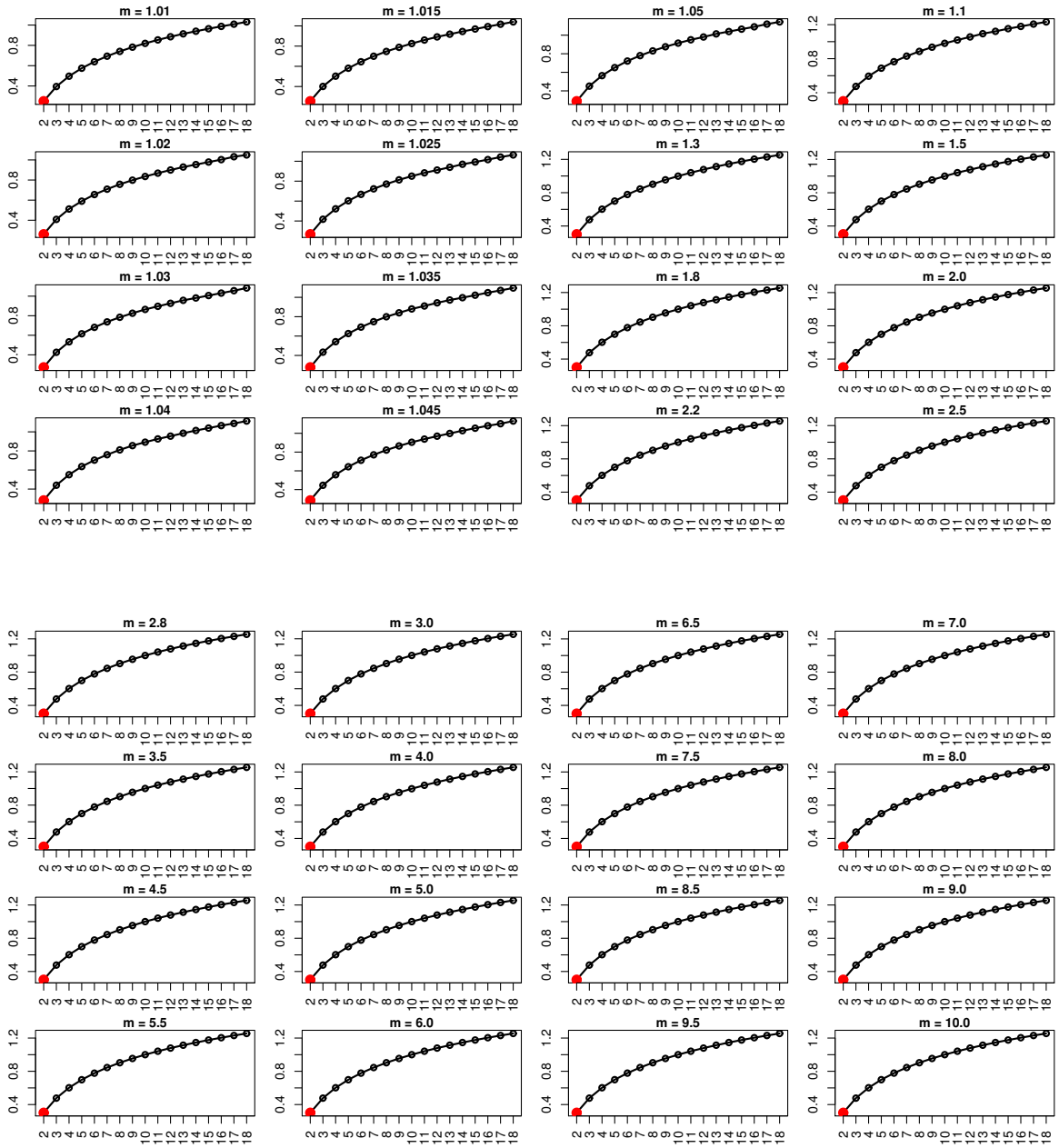


Tabela 25 – Irish-Sentiment

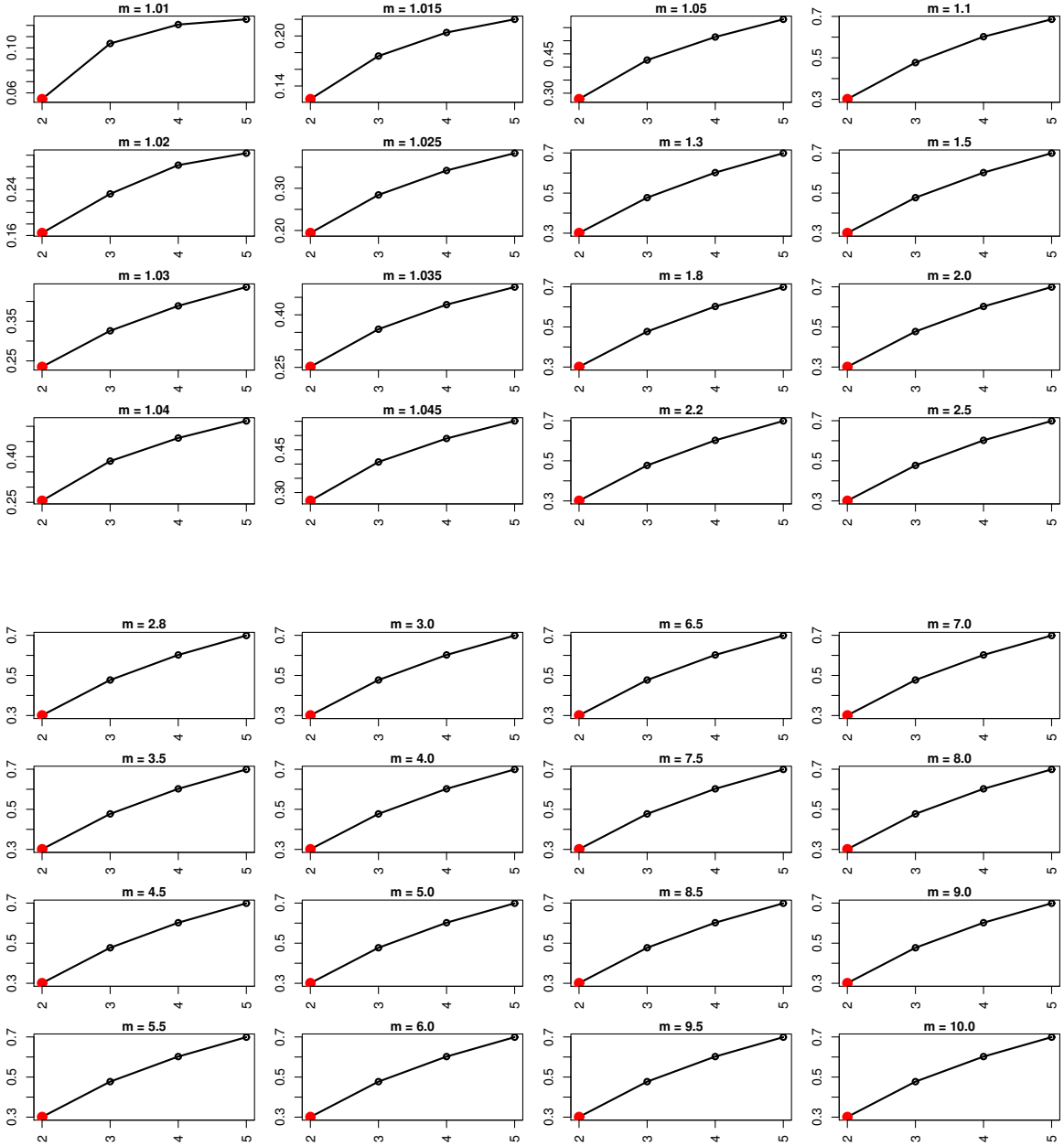


Tabela 26 – 20Newsgroups

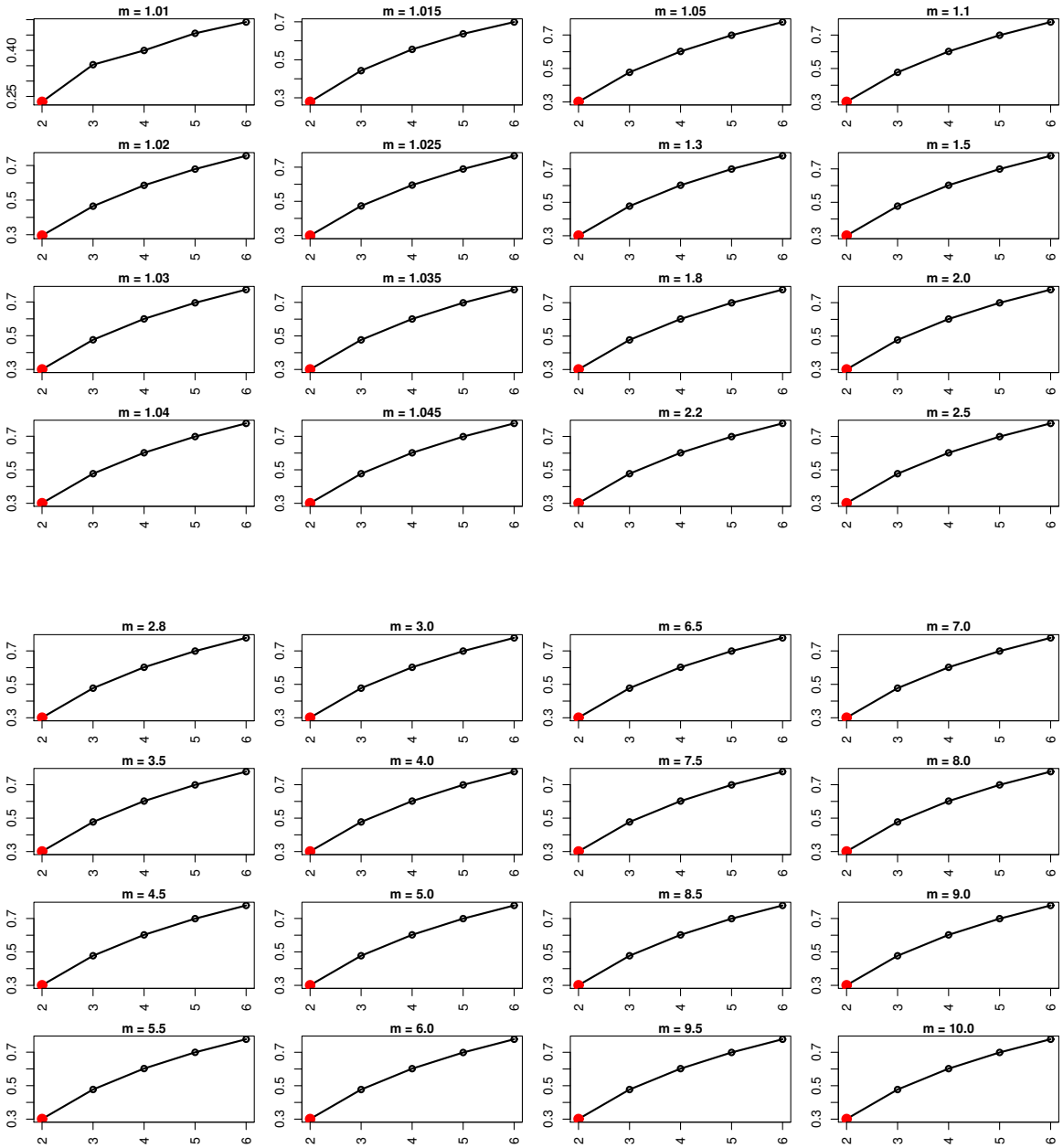


Tabela 27 – La1s

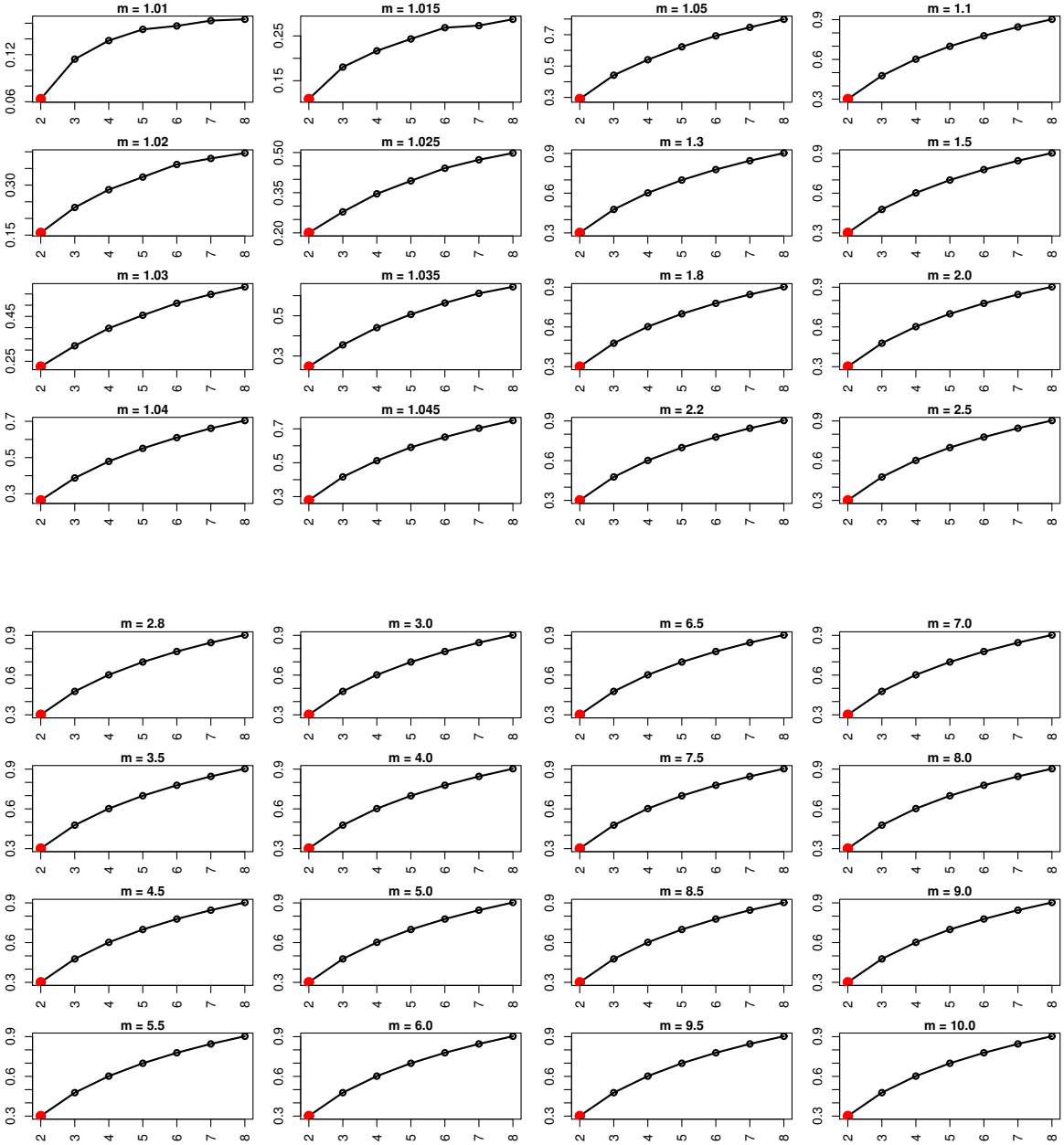
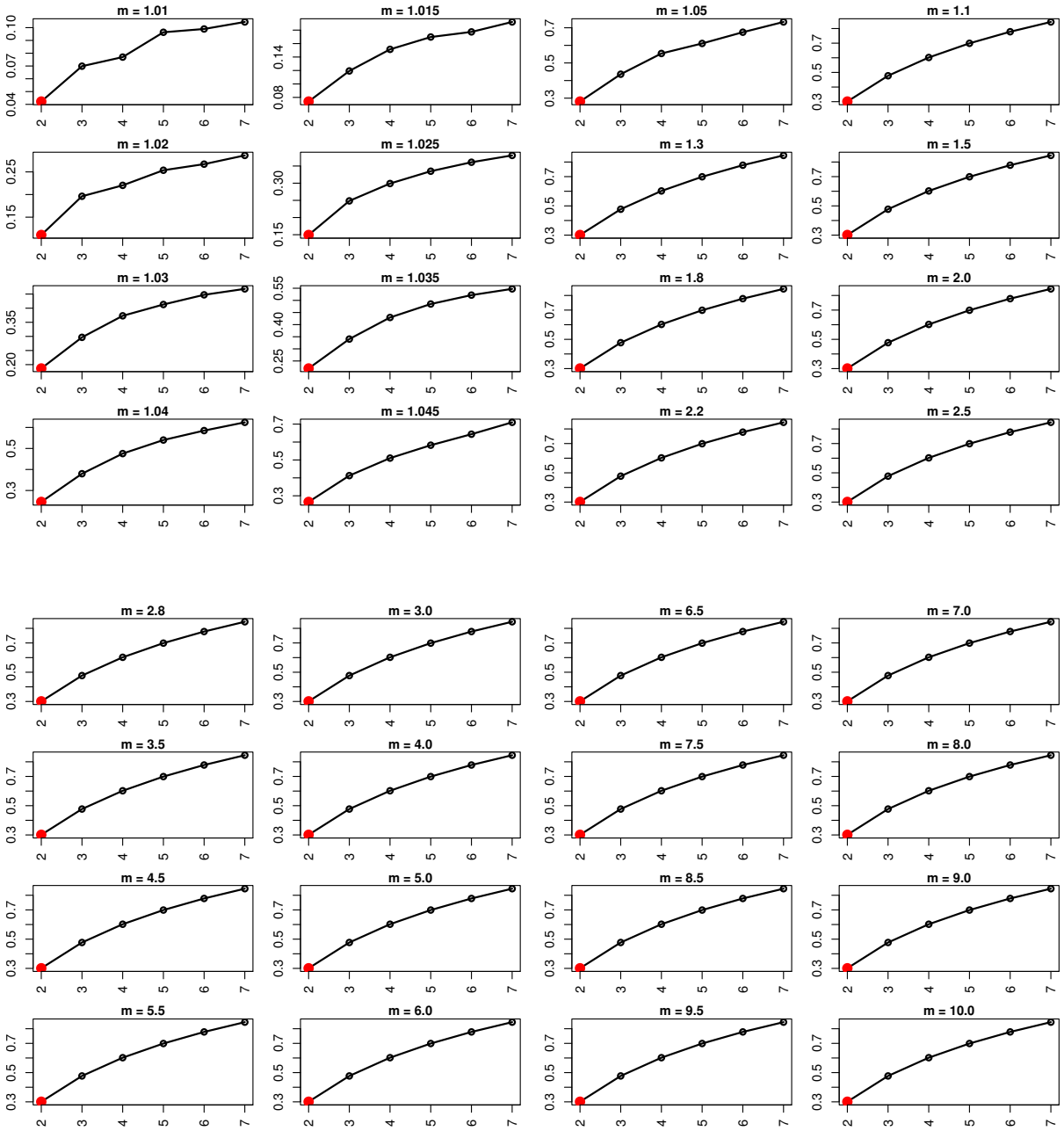


Tabela 28 – Reviews



ANEXO C – MPC

Tabela 29 – NewYorkTimes

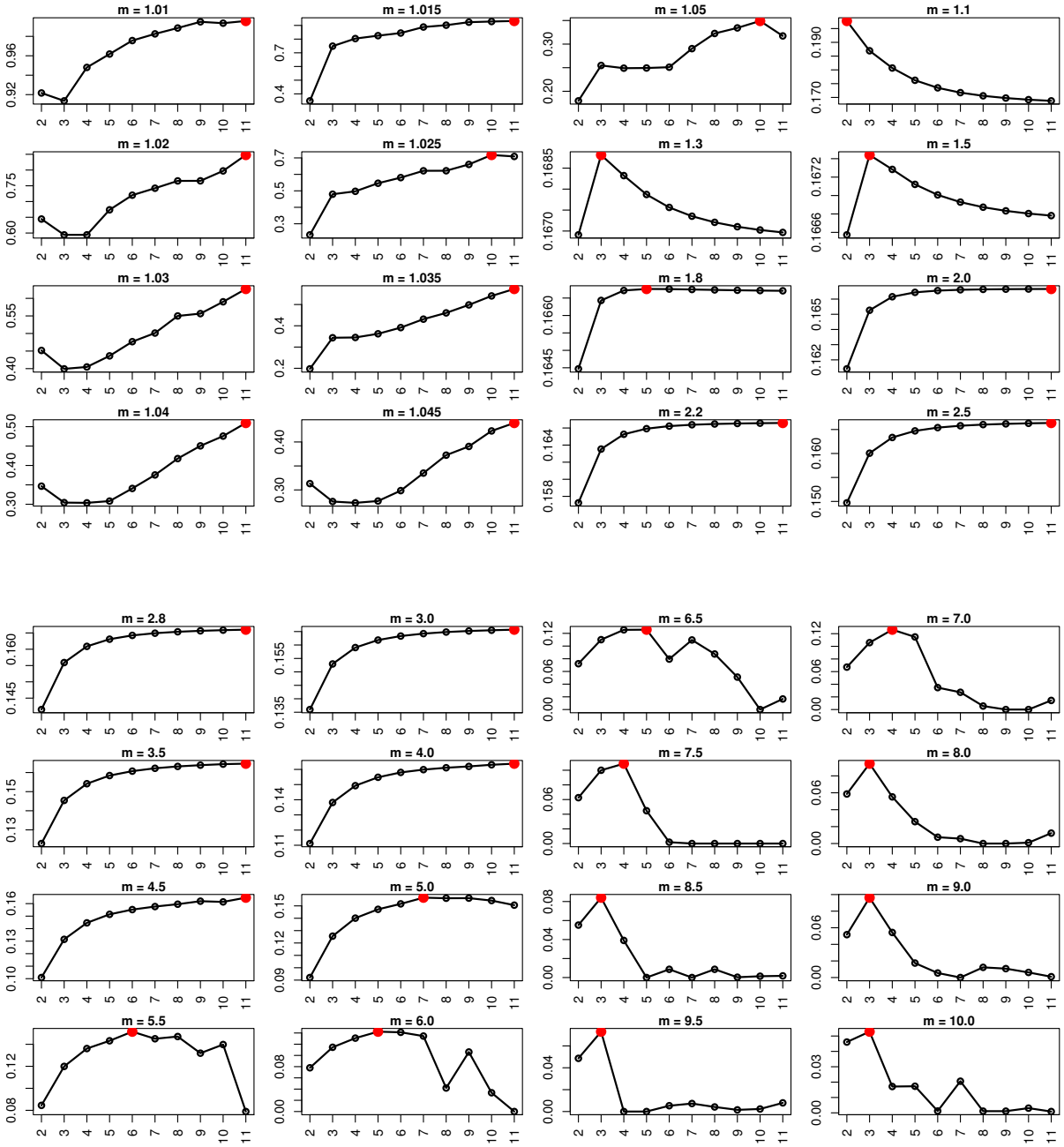


Tabela 30 – IAarticles

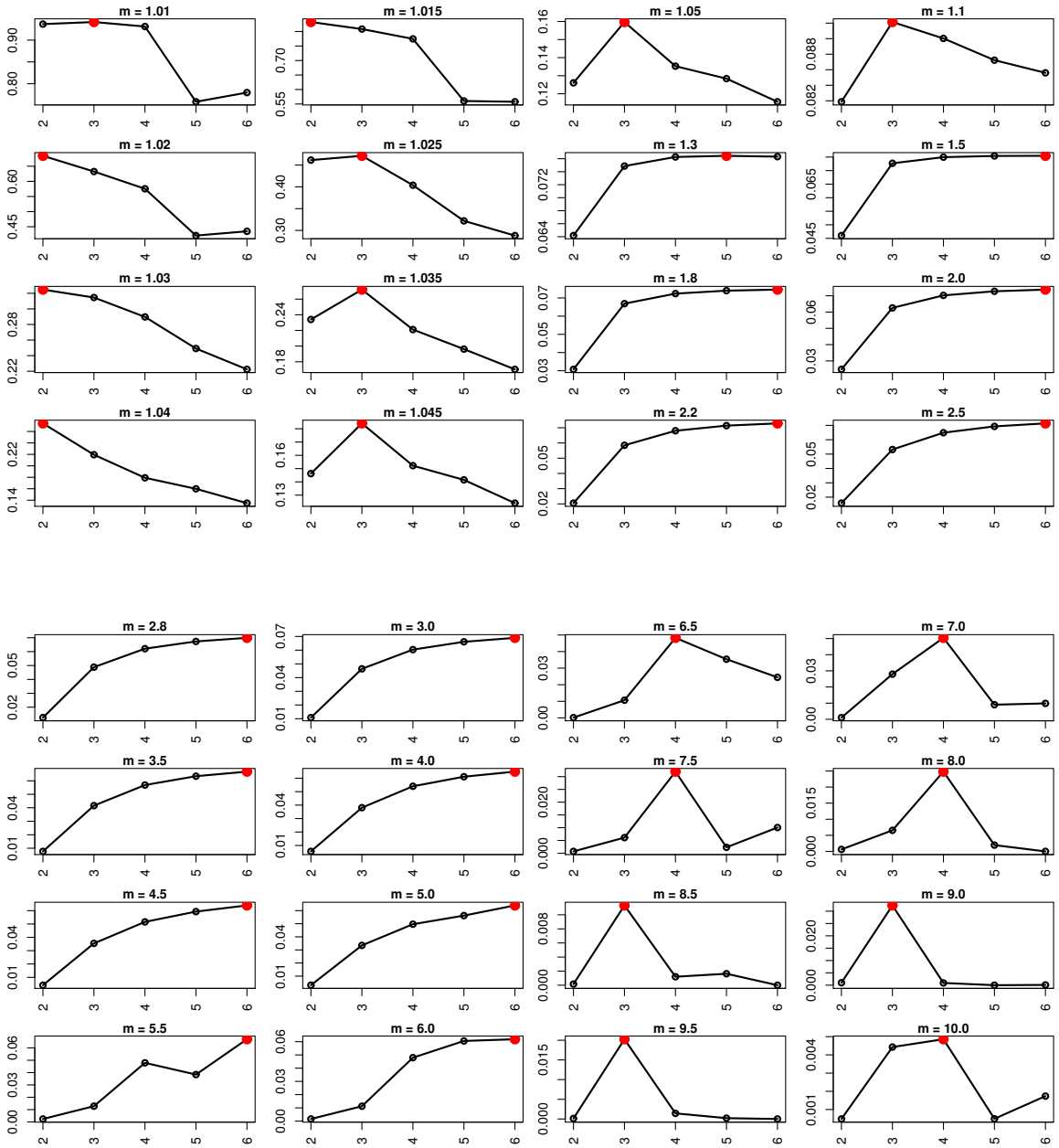


Tabela 31 – Opinosis

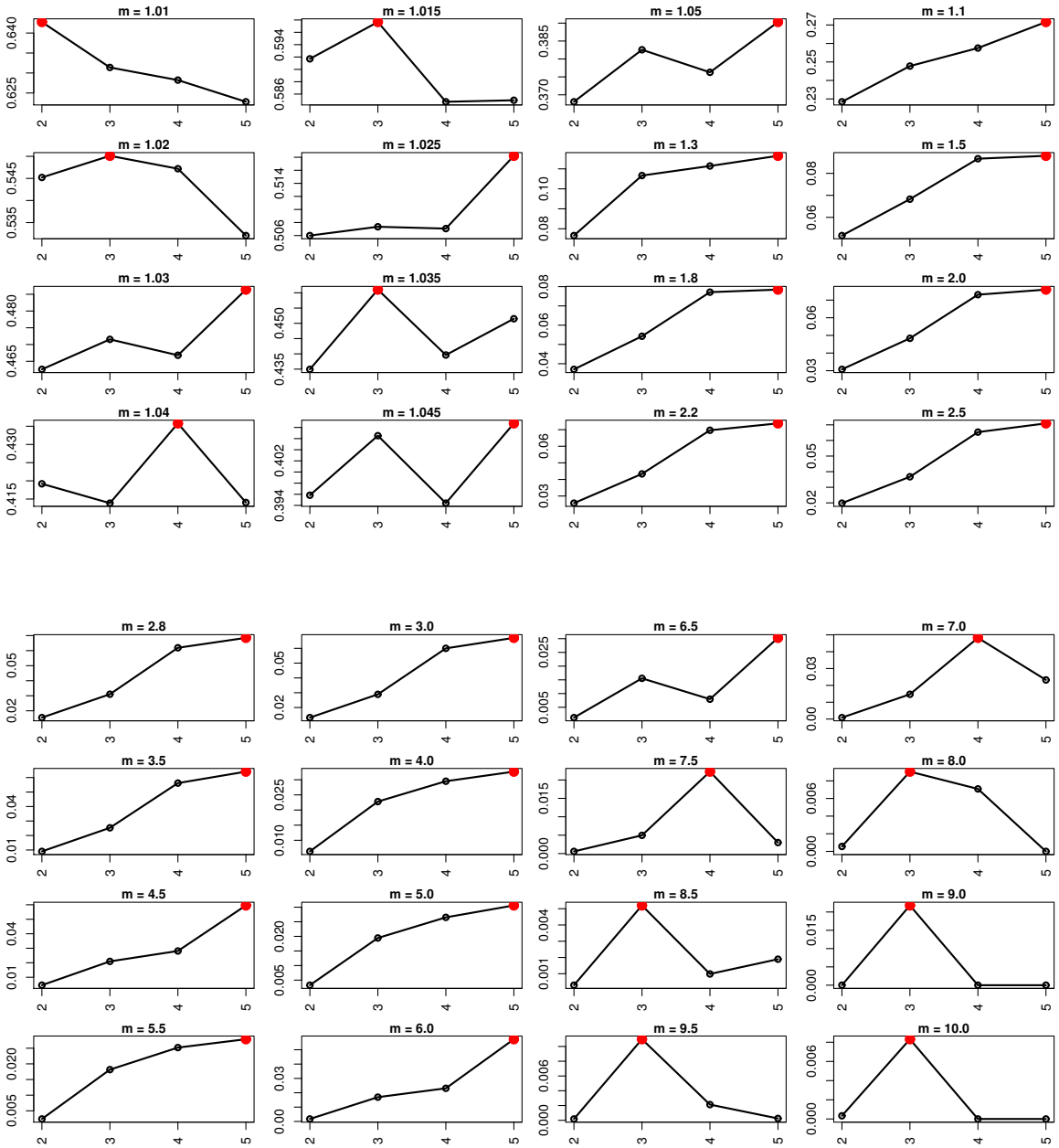


Tabela 32 – CSTR

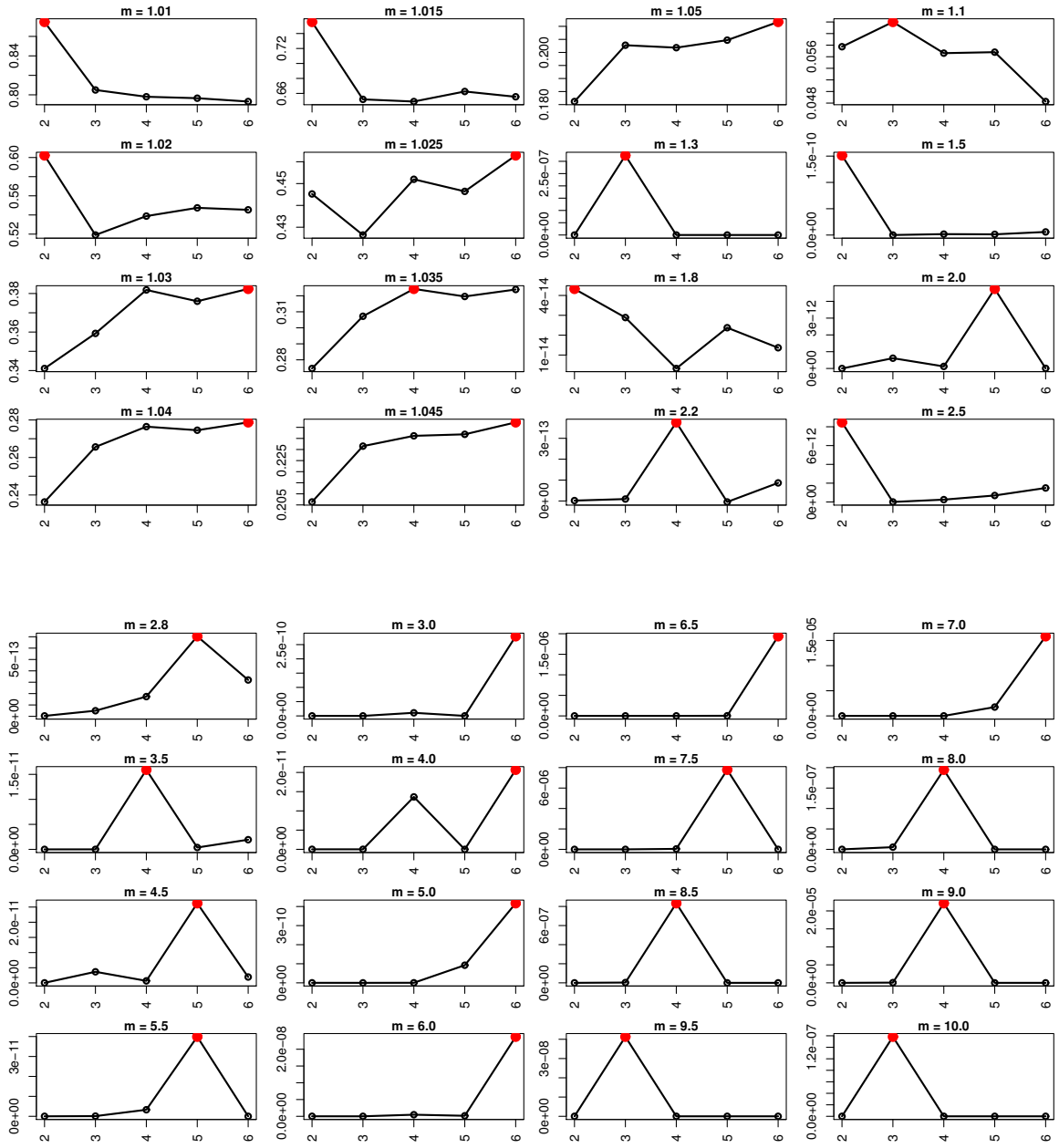


Tabela 33 – SyskillWebert

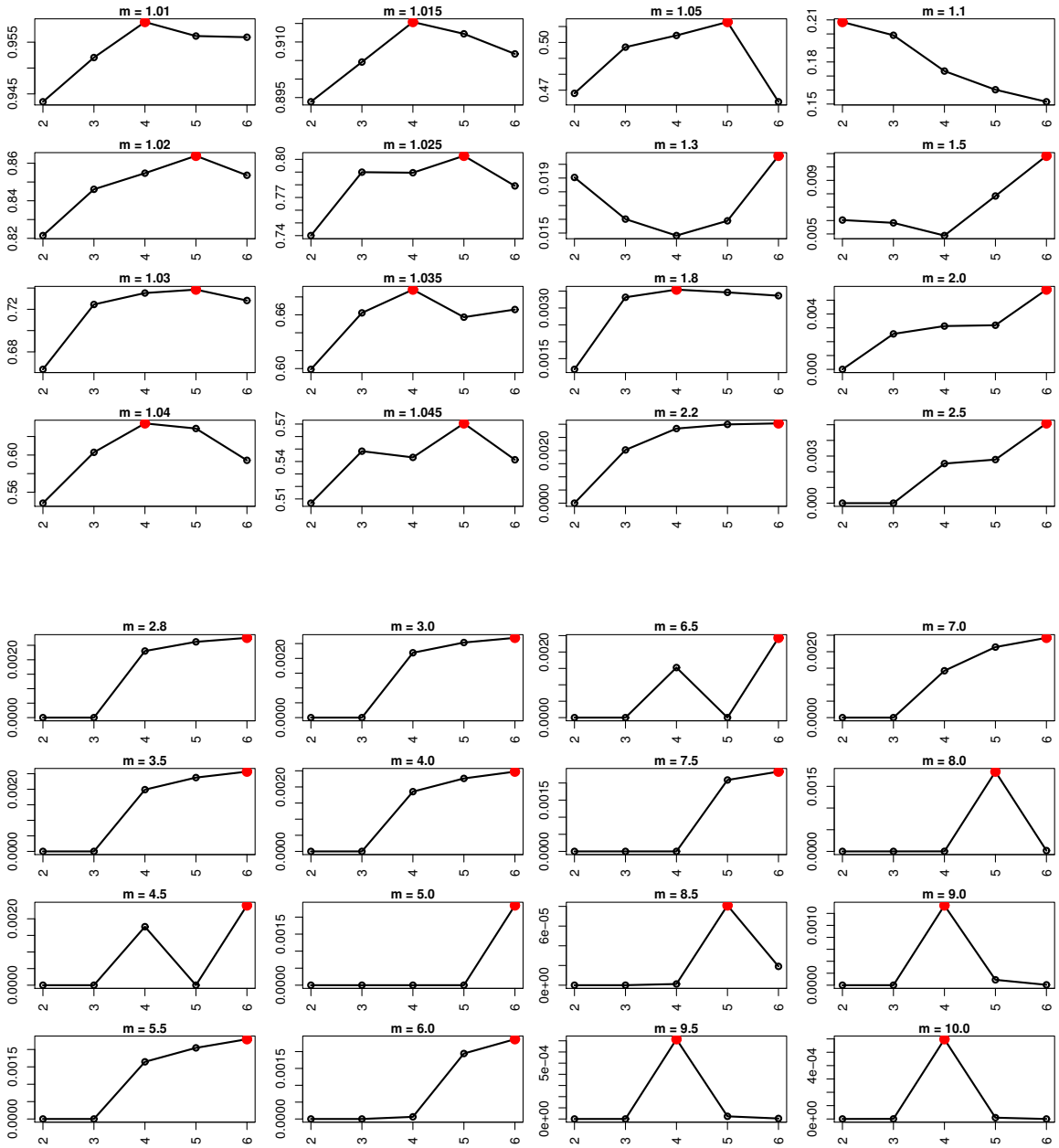


Tabela 34 – Hitech

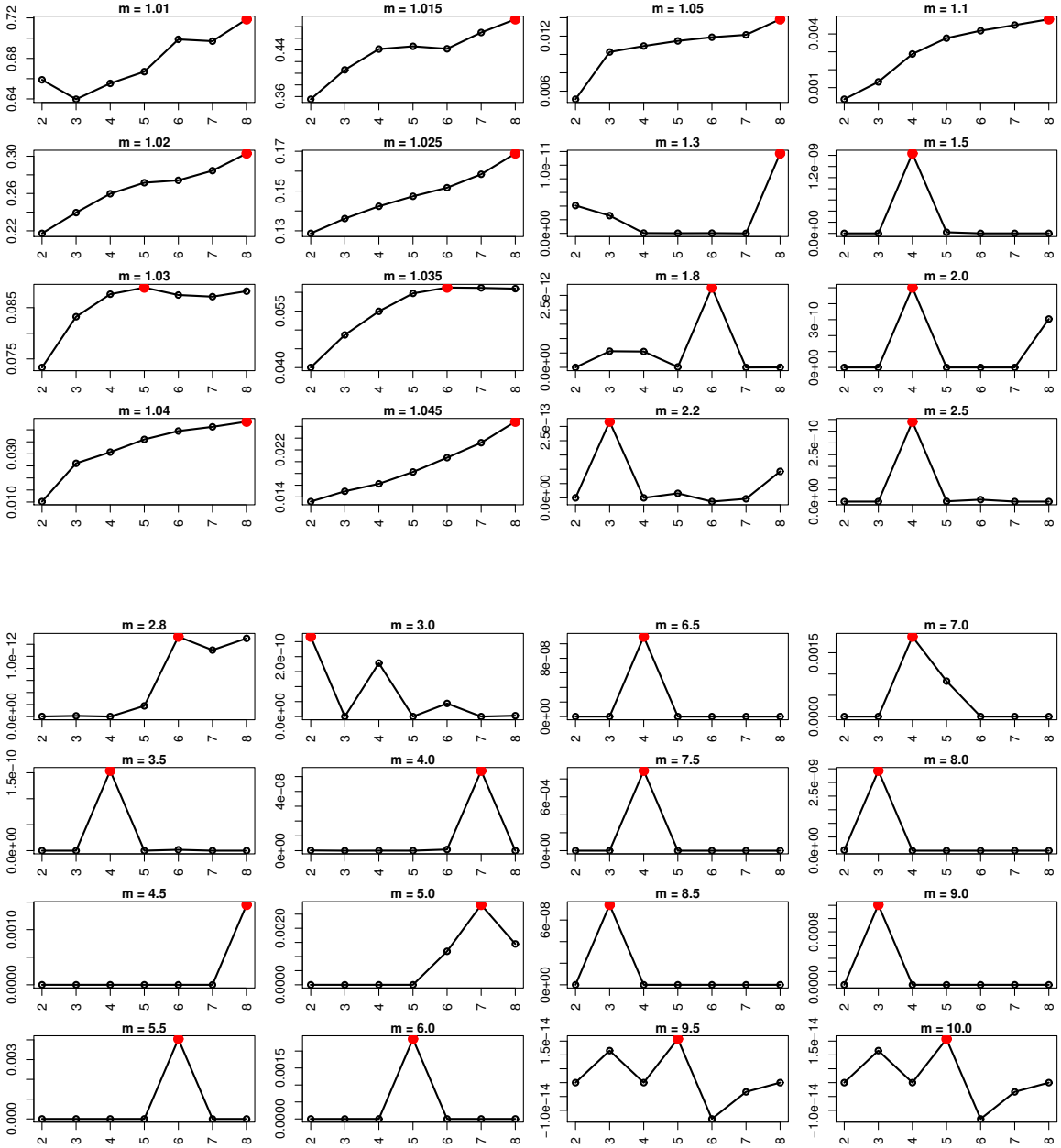


Tabela 35 – WAP

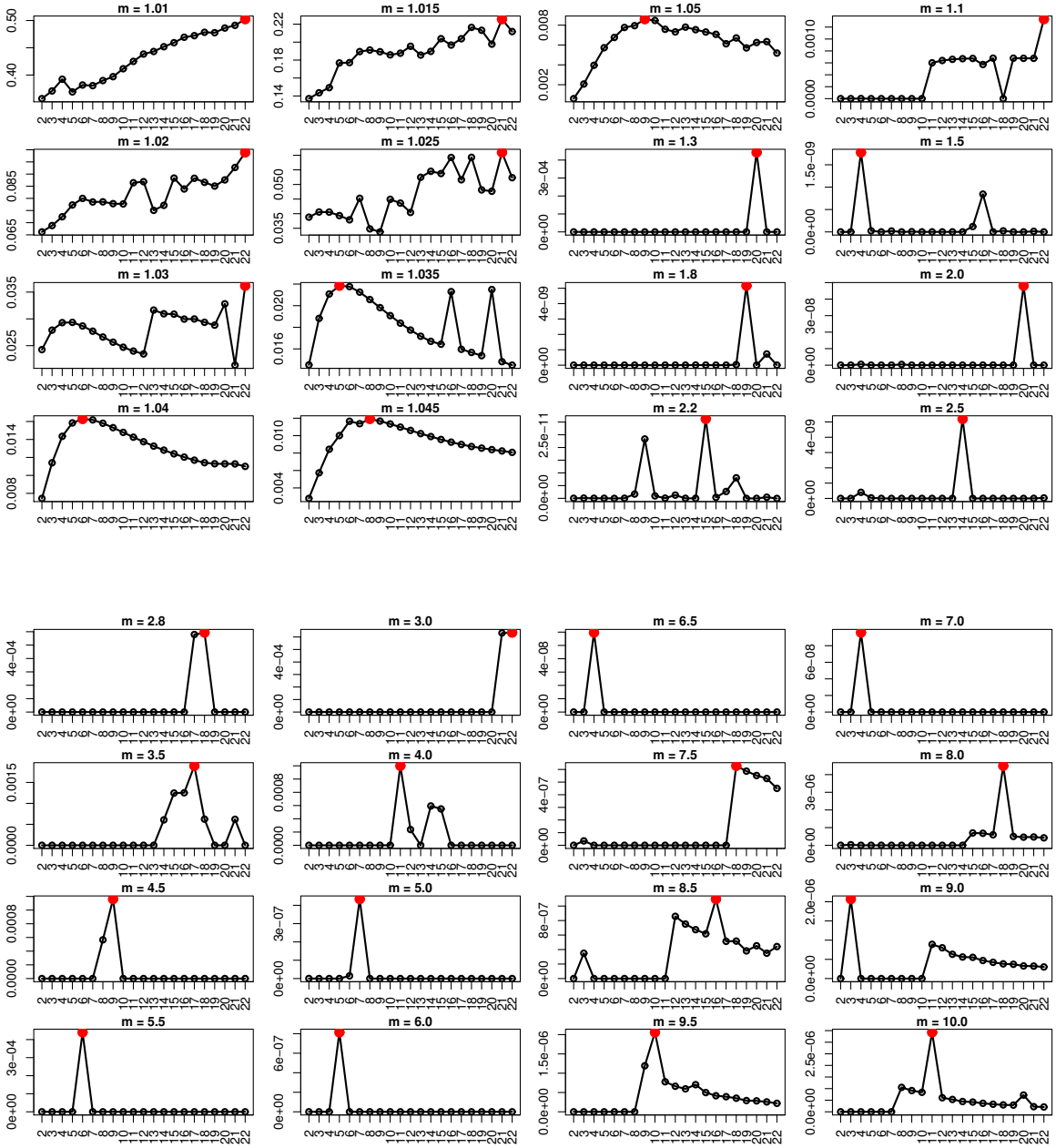


Tabela 36 – NSF

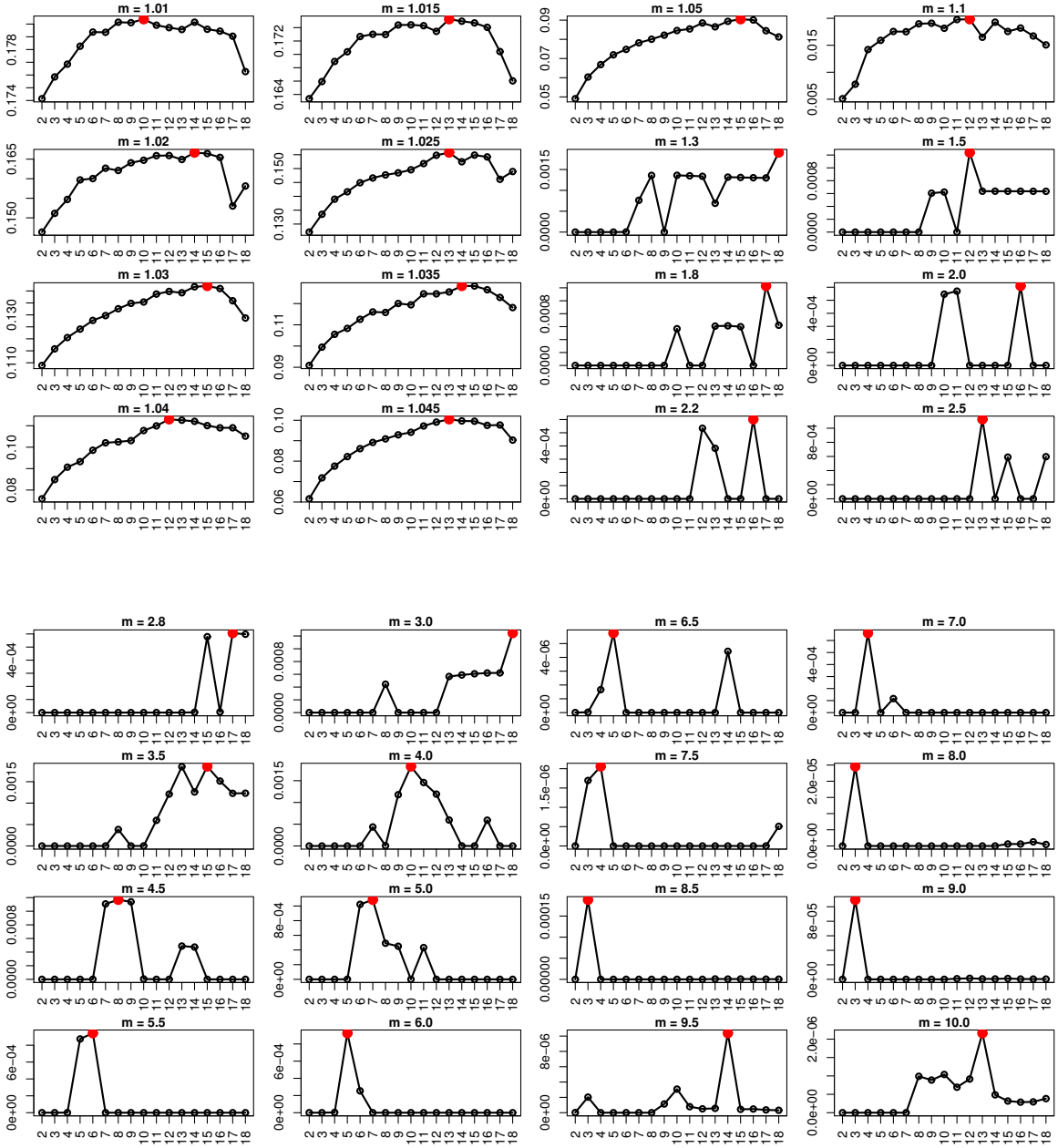


Tabela 37 – Irish-Sentiment

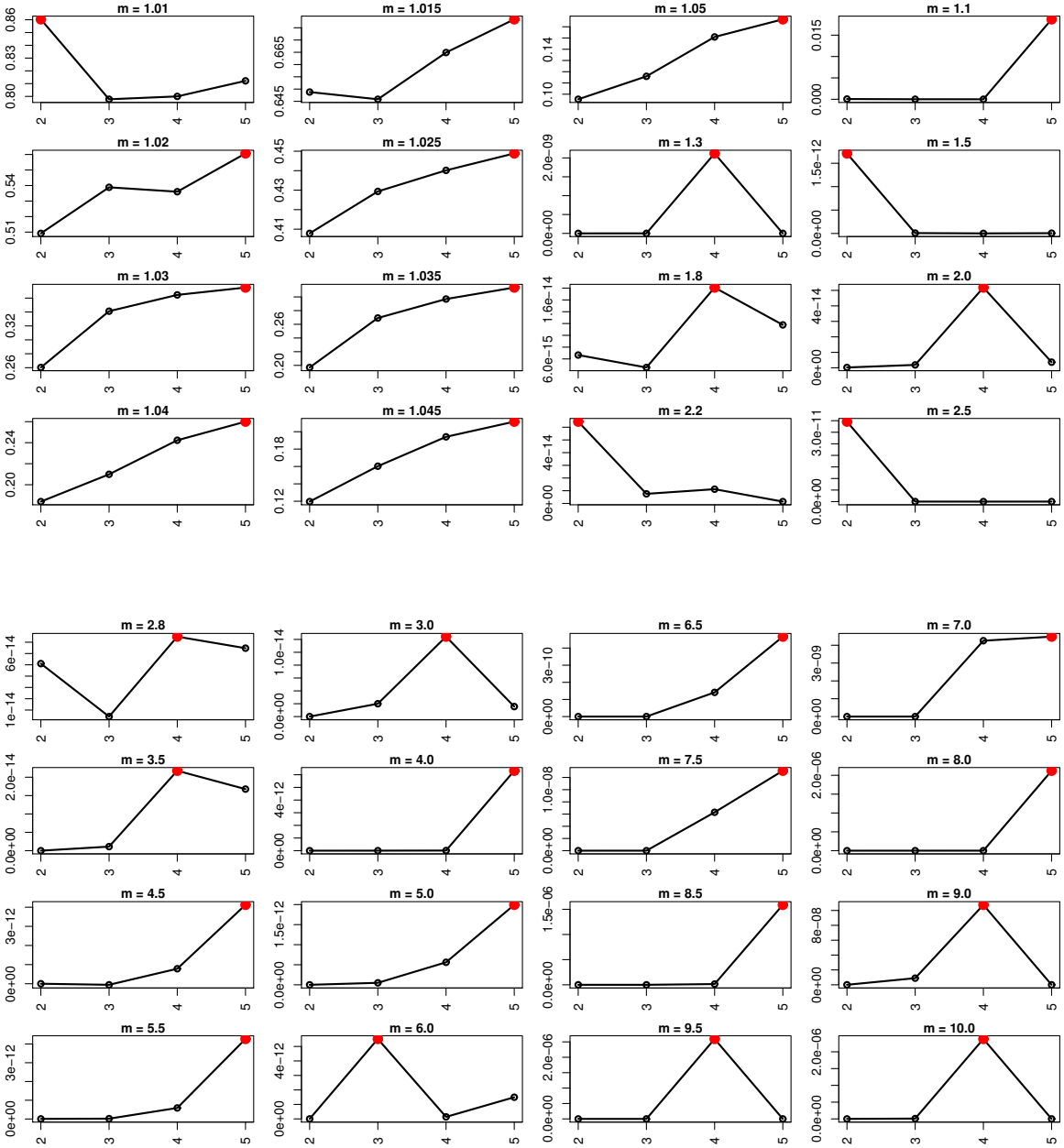


Tabela 38 – 20Newsgroups

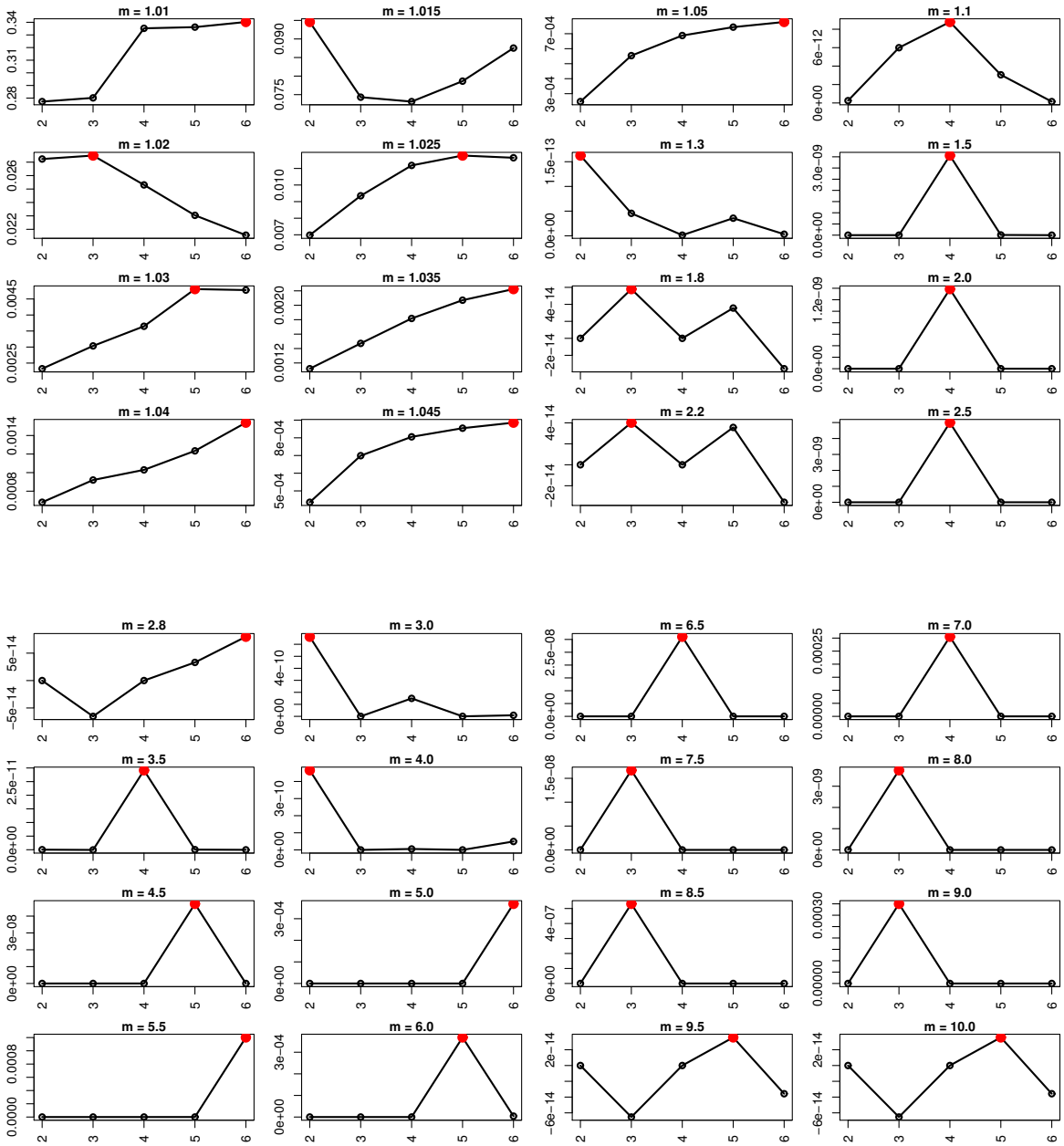


Tabela 39 – La1s

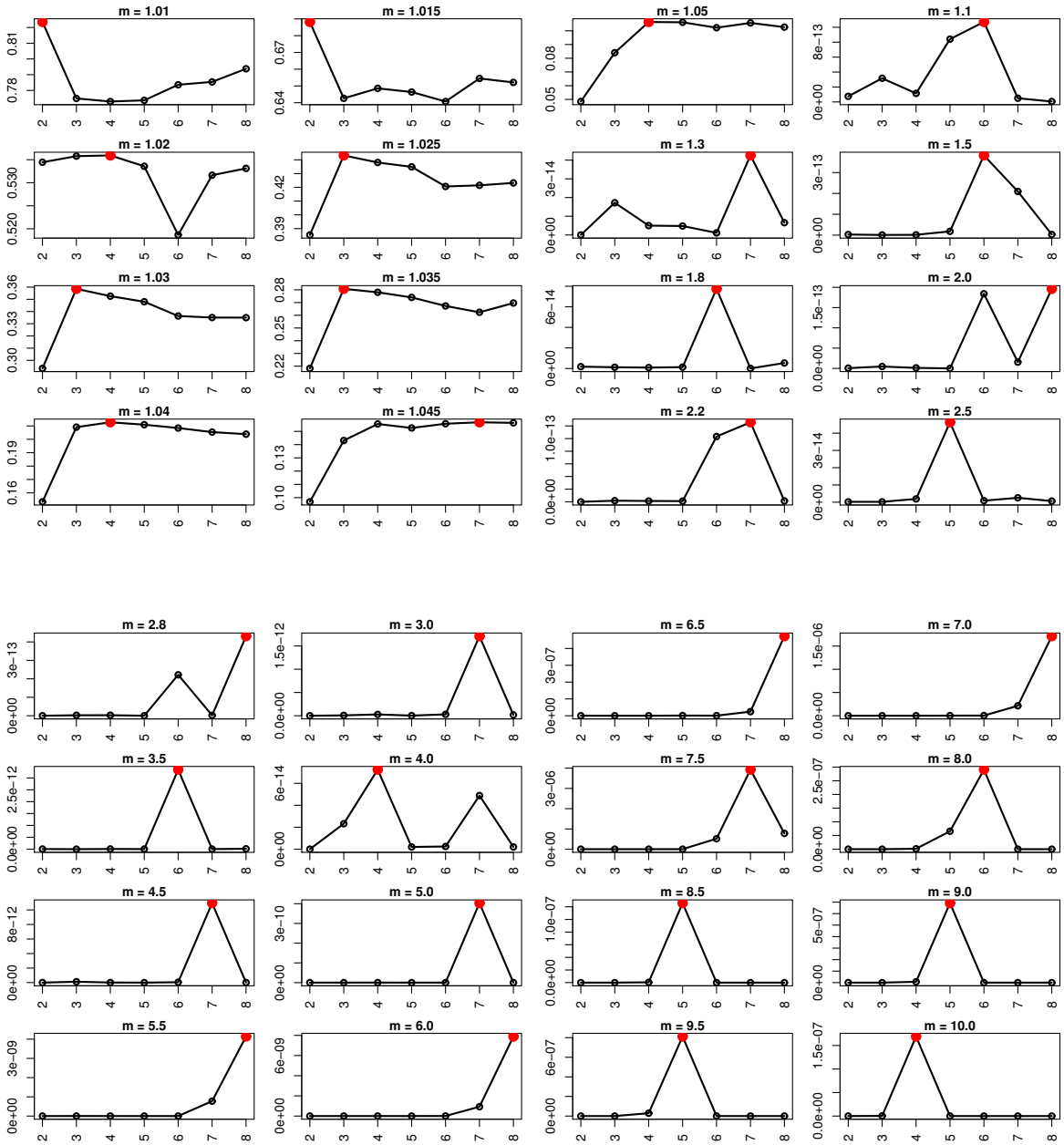
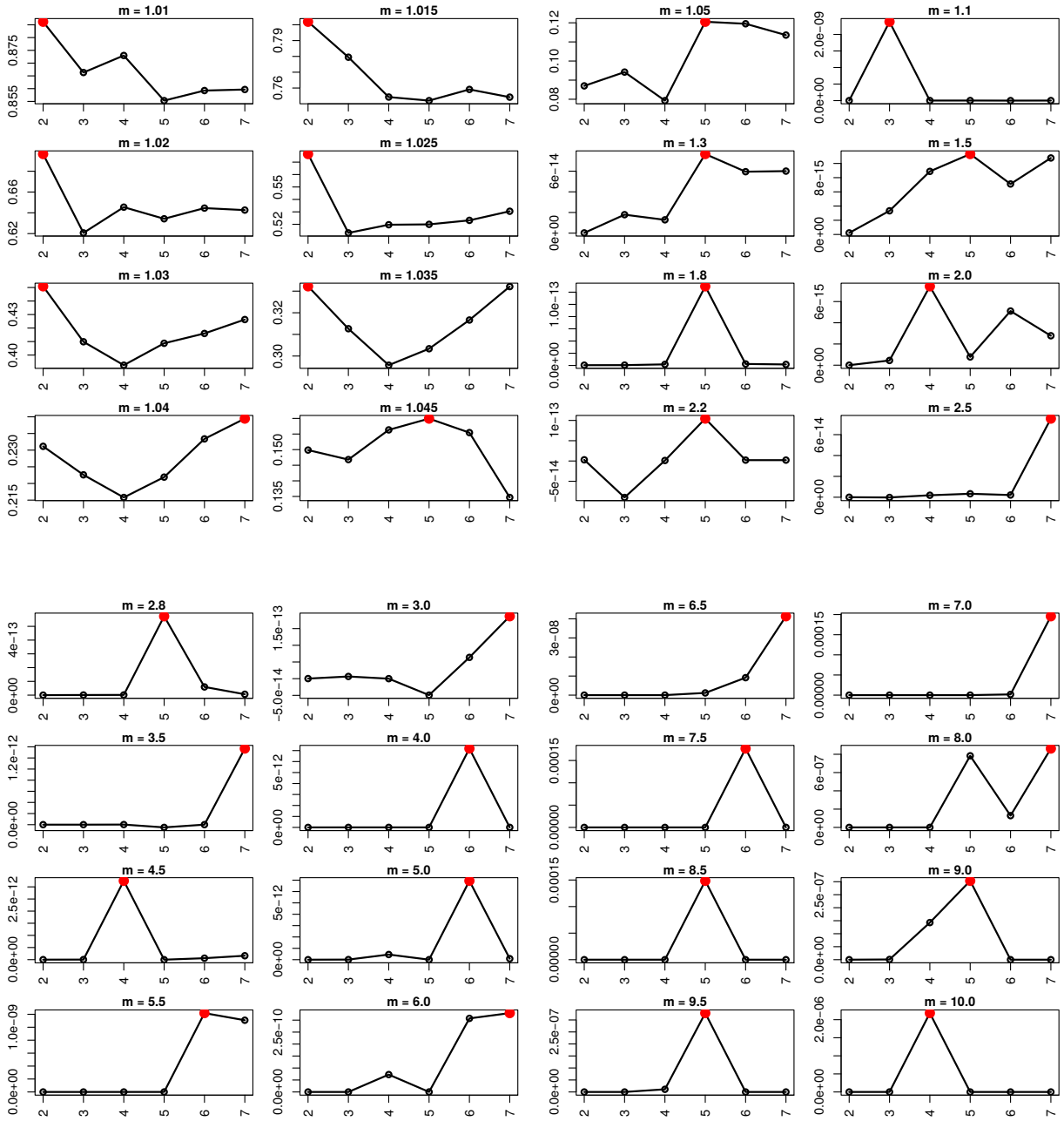


Tabela 40 – Reviews



ANEXO D – KYI

Tabela 41 – NewYorkTimes

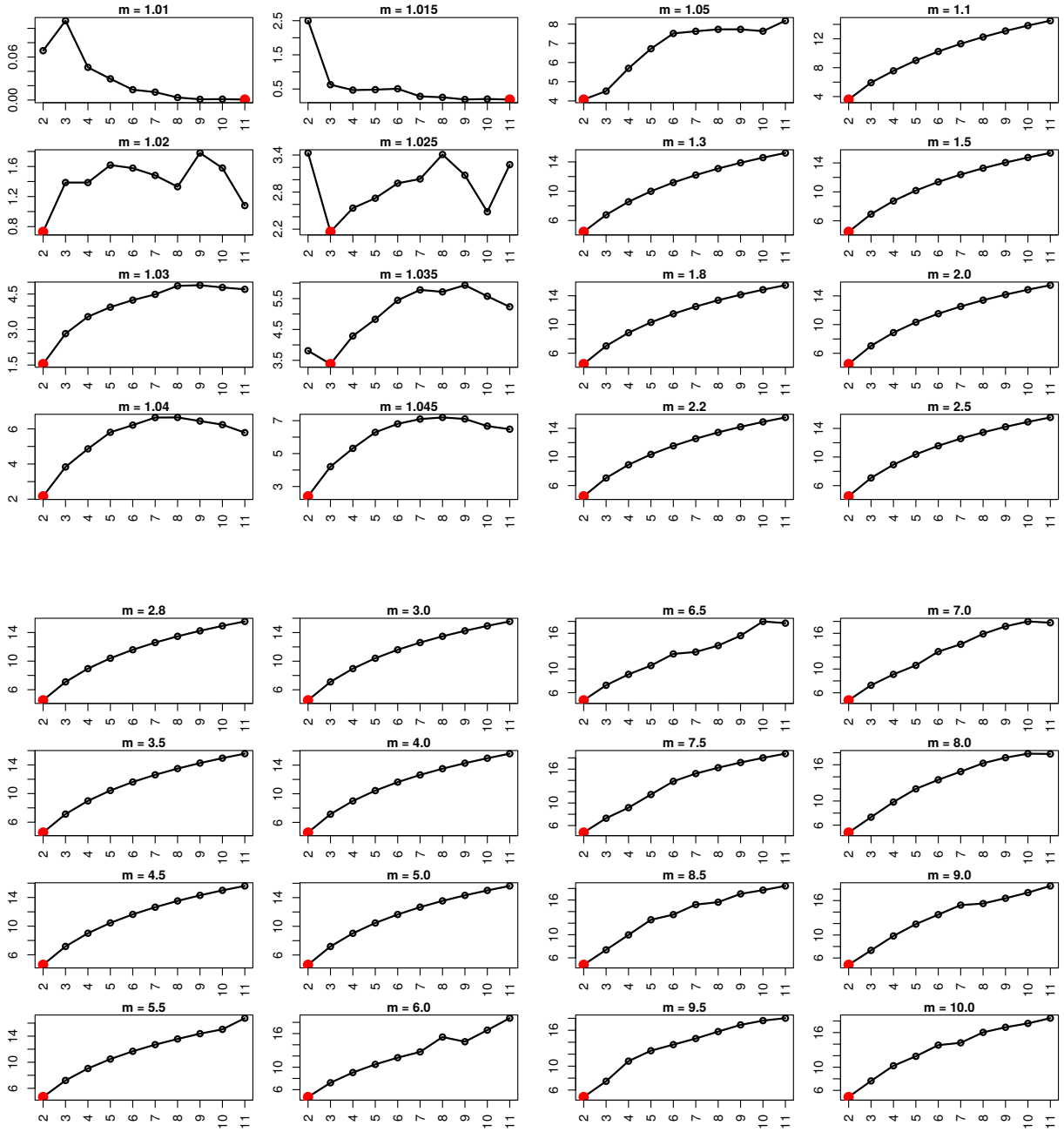


Tabela 42 – IAarticles

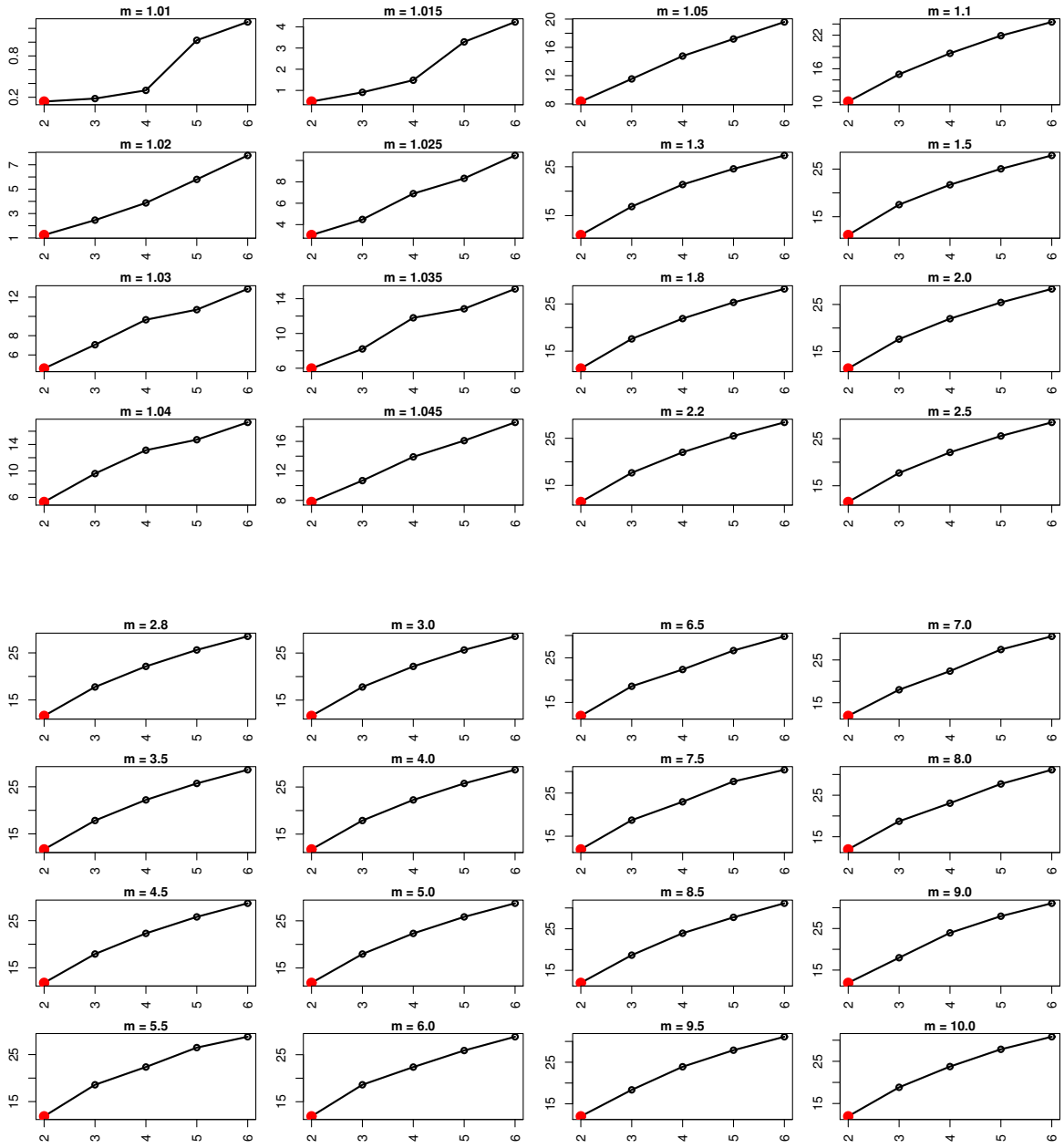


Tabela 43 – Opinosis

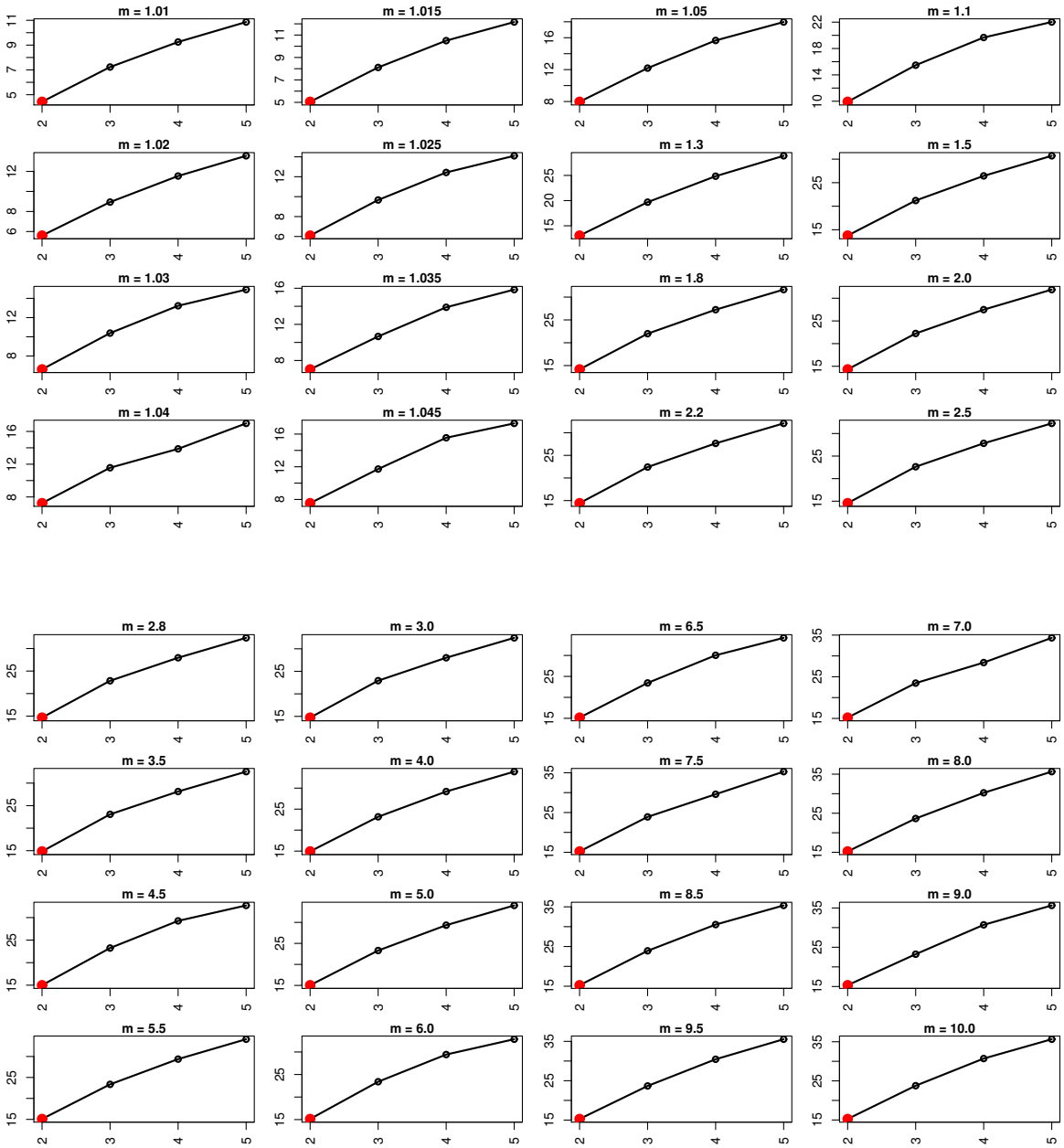


Tabela 44 – CSTR

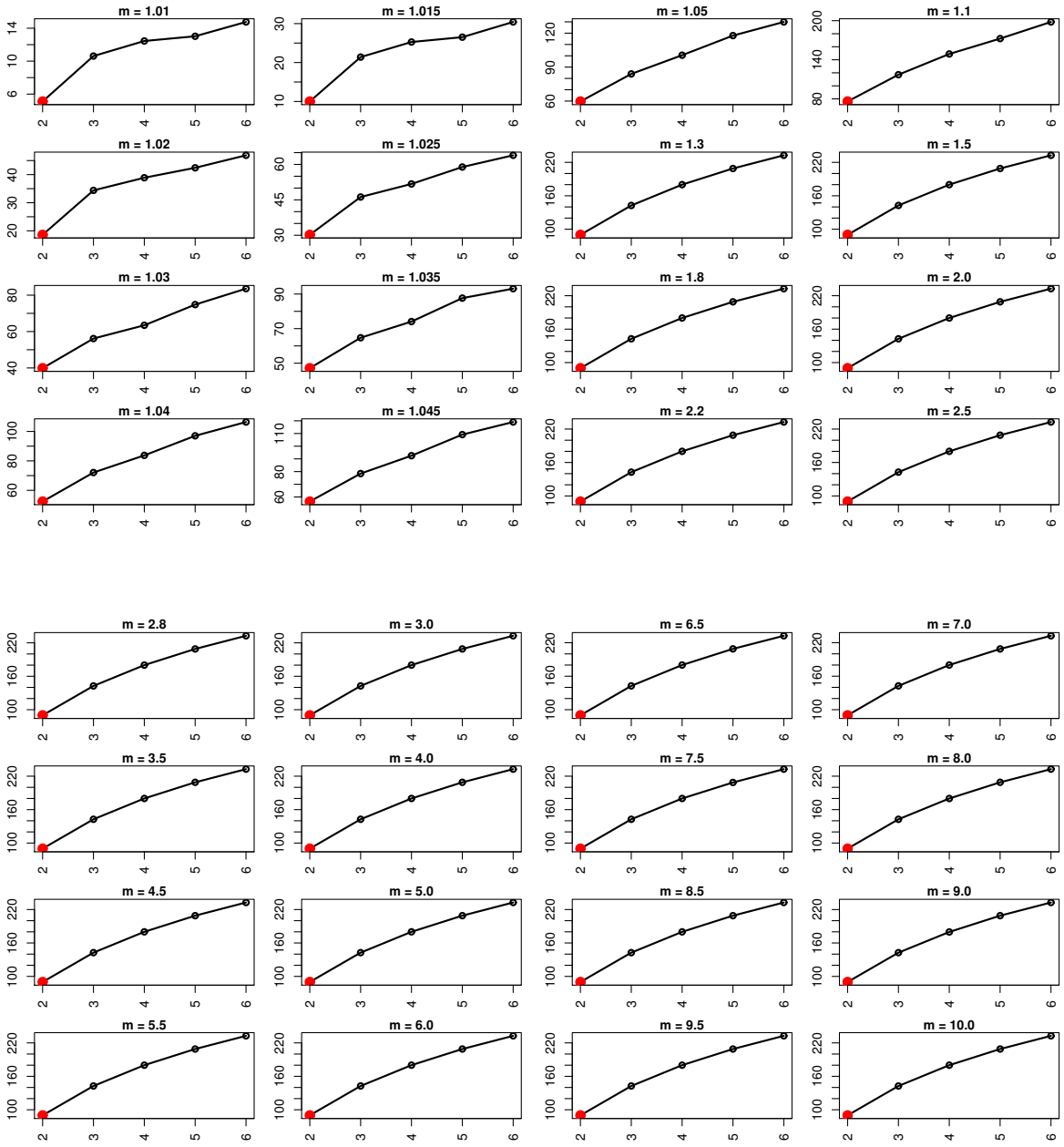


Tabela 45 – SyskillWebert

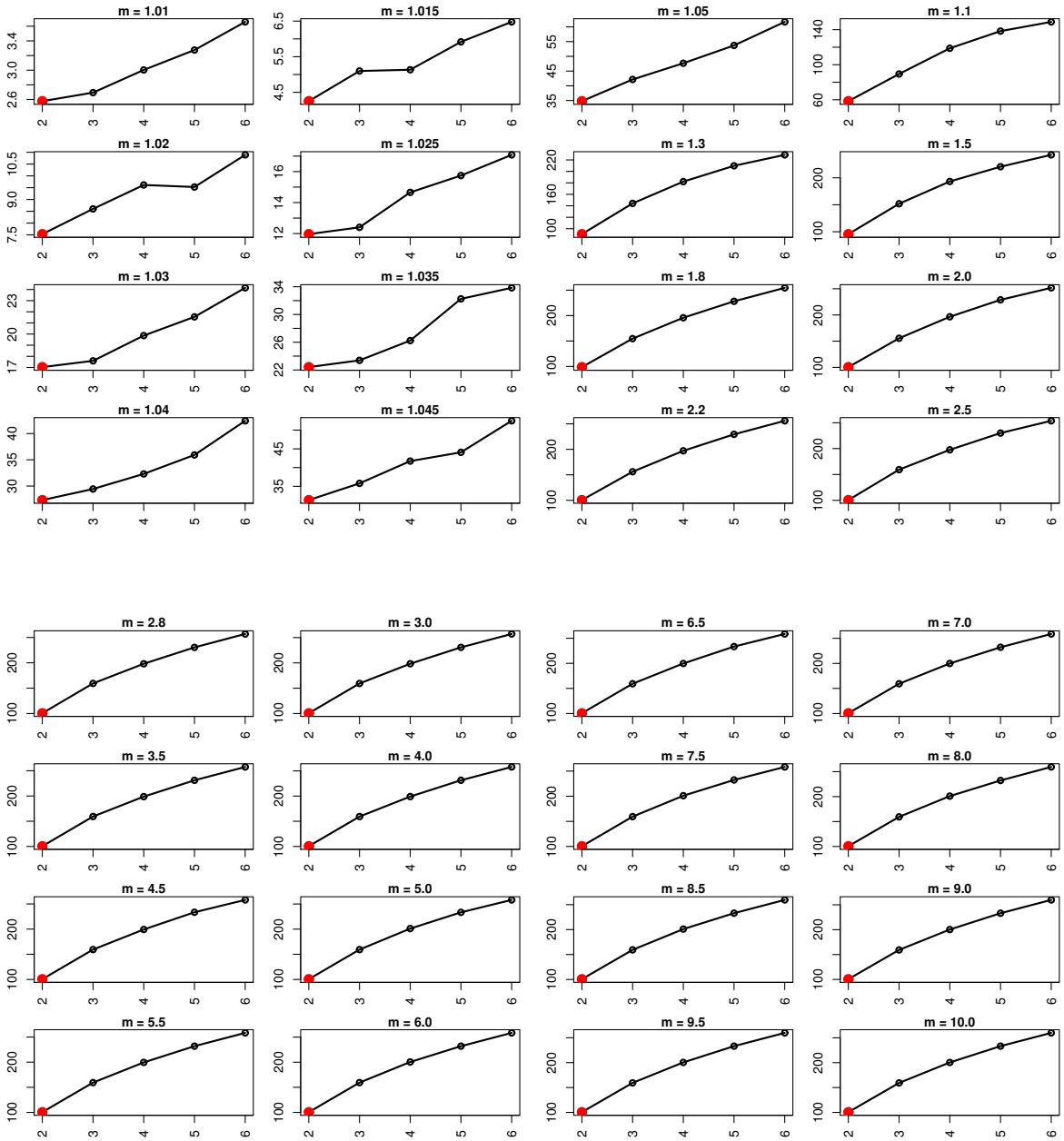


Tabela 46 – Hitech

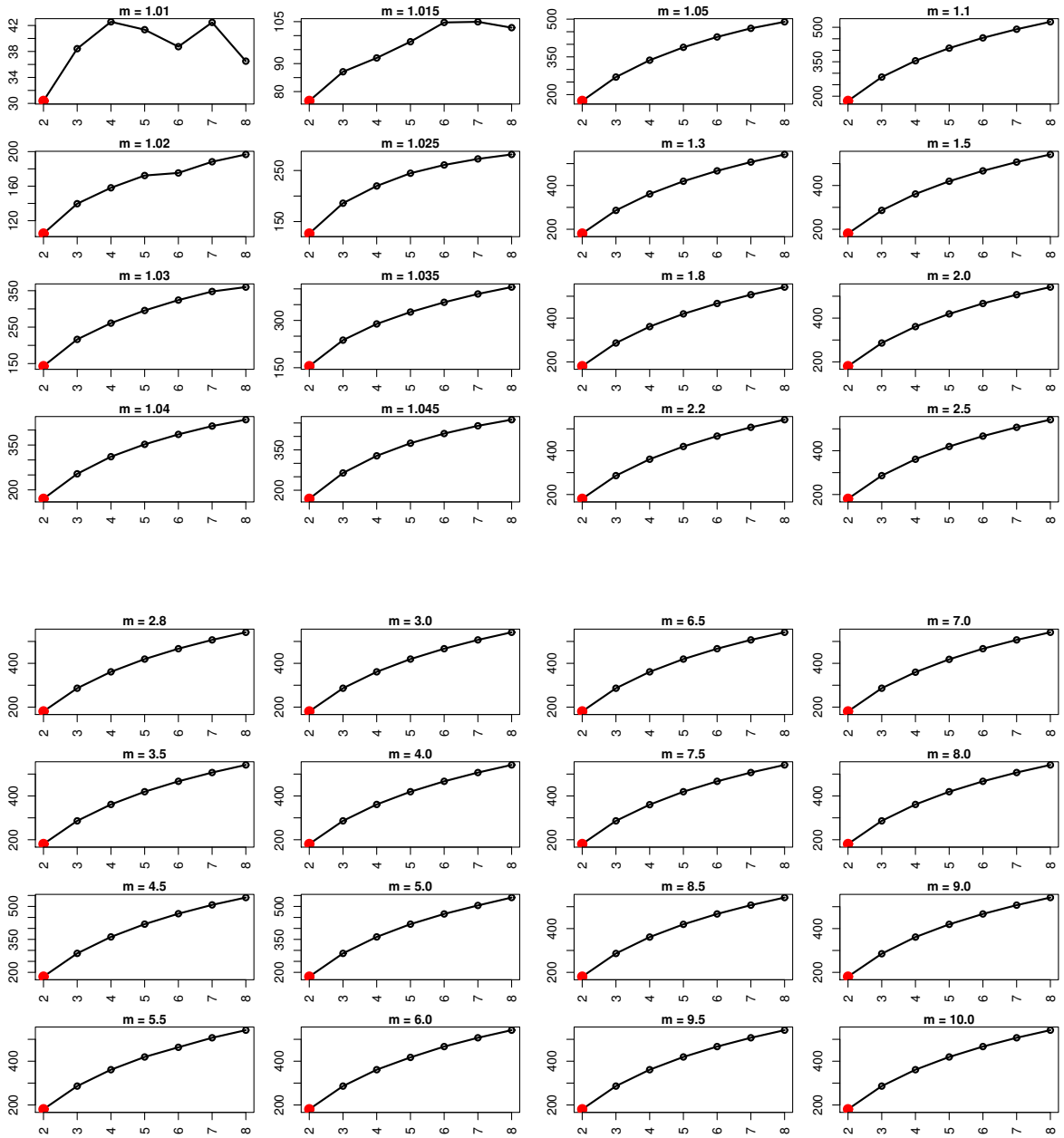


Tabela 47 – WAP

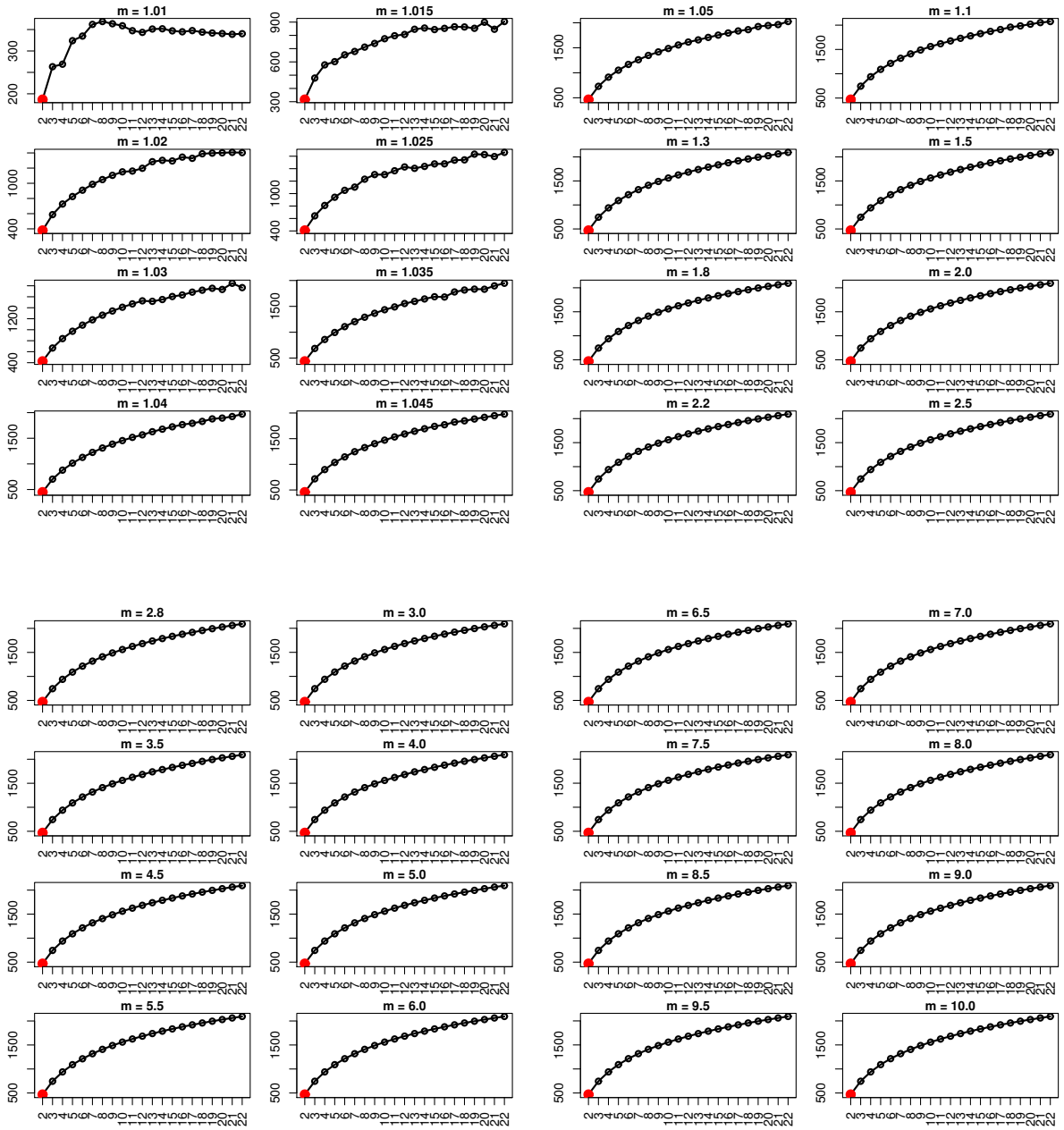


Tabela 48 – NSF

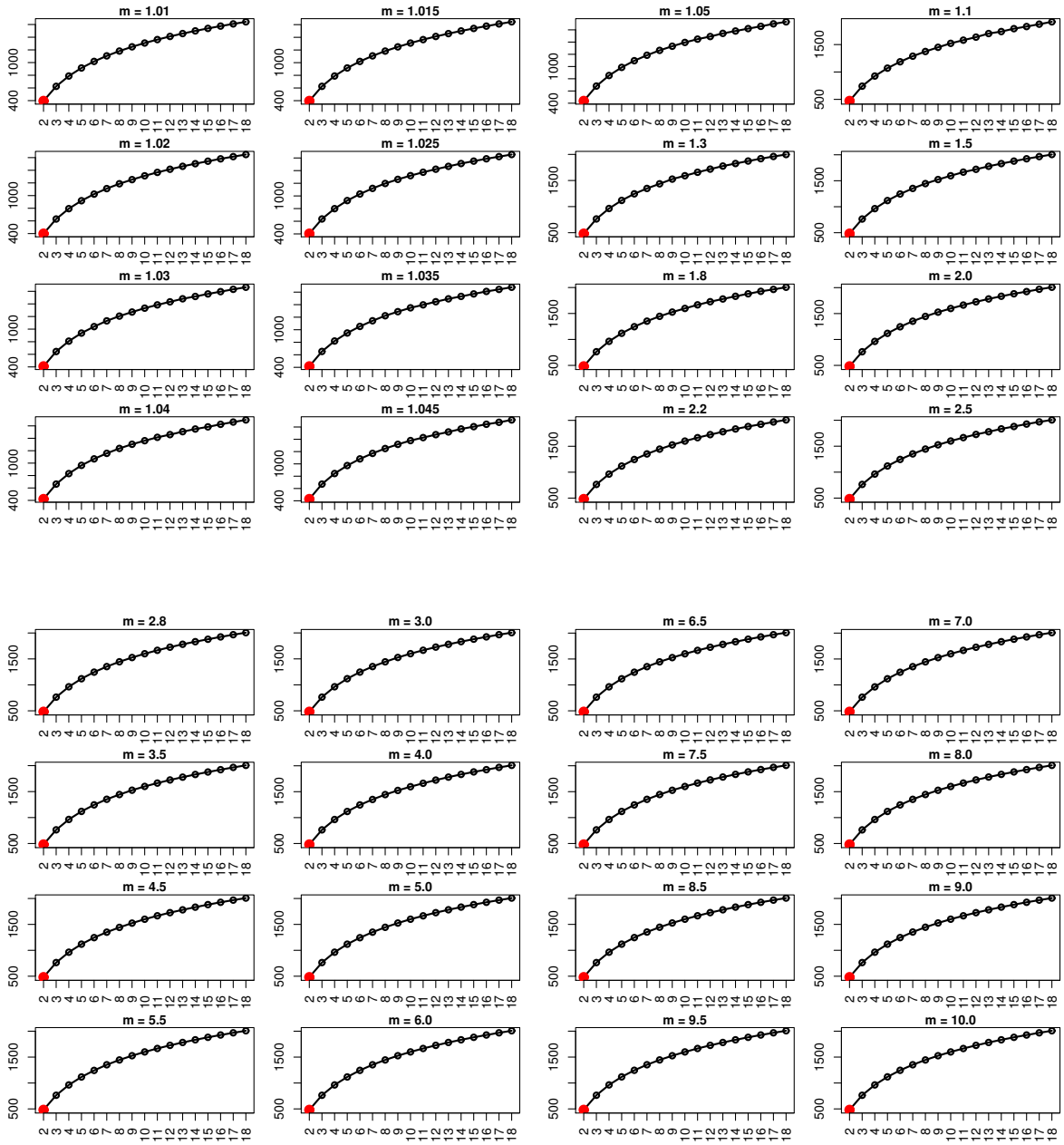


Tabela 49 – Irish-Sentiment

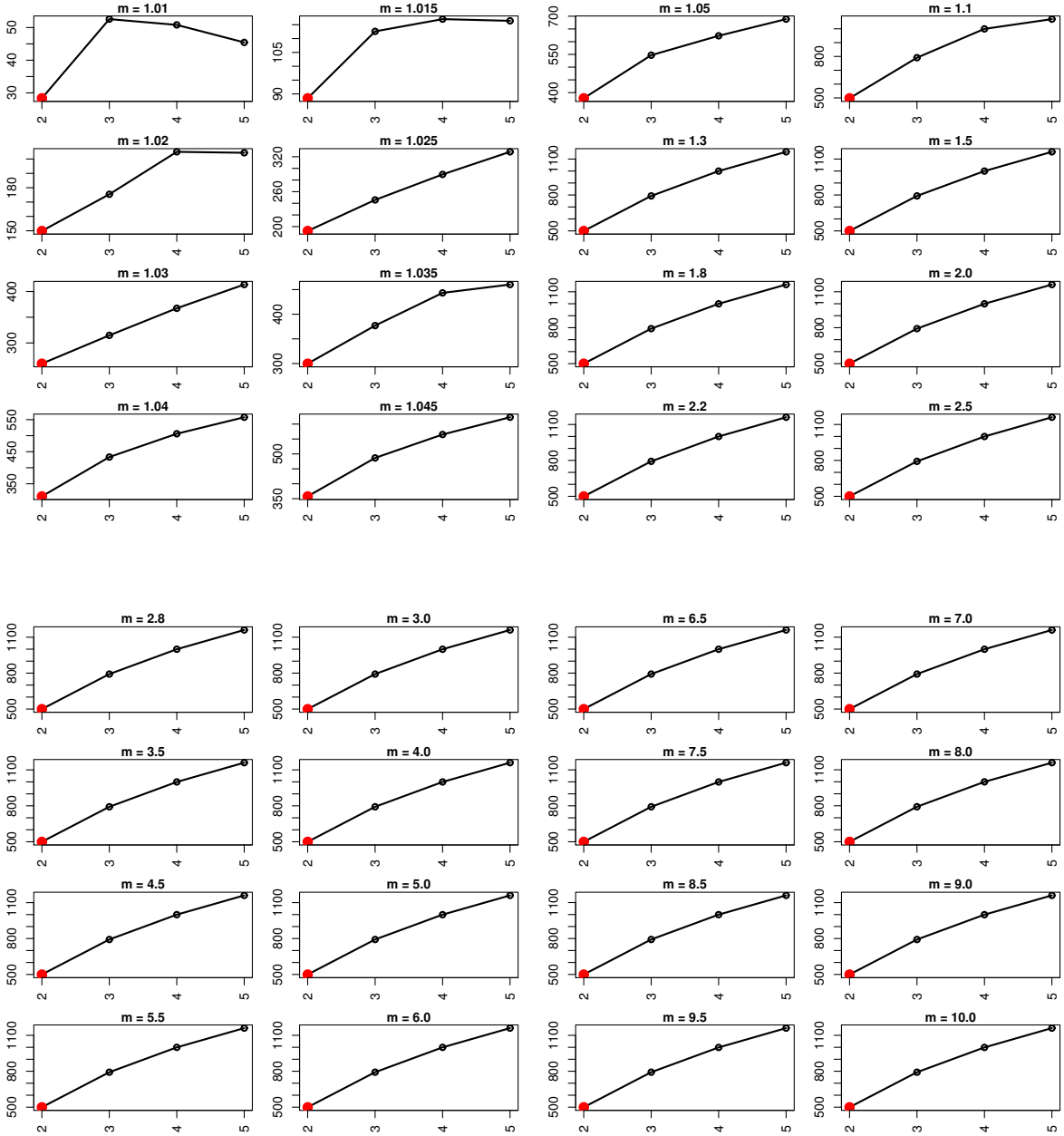


Tabela 50 – 20Newsgroups

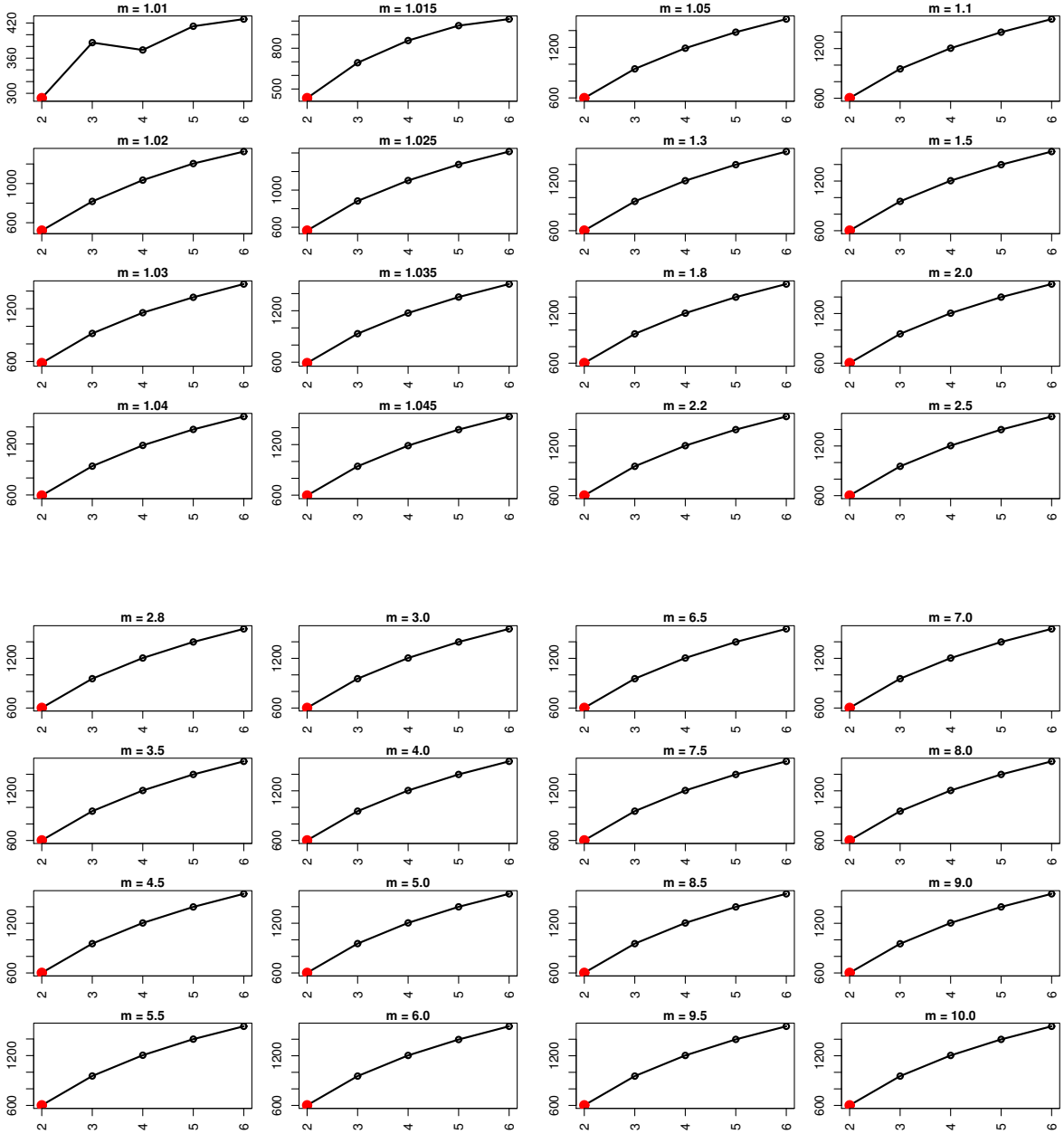


Tabela 51 – La1s

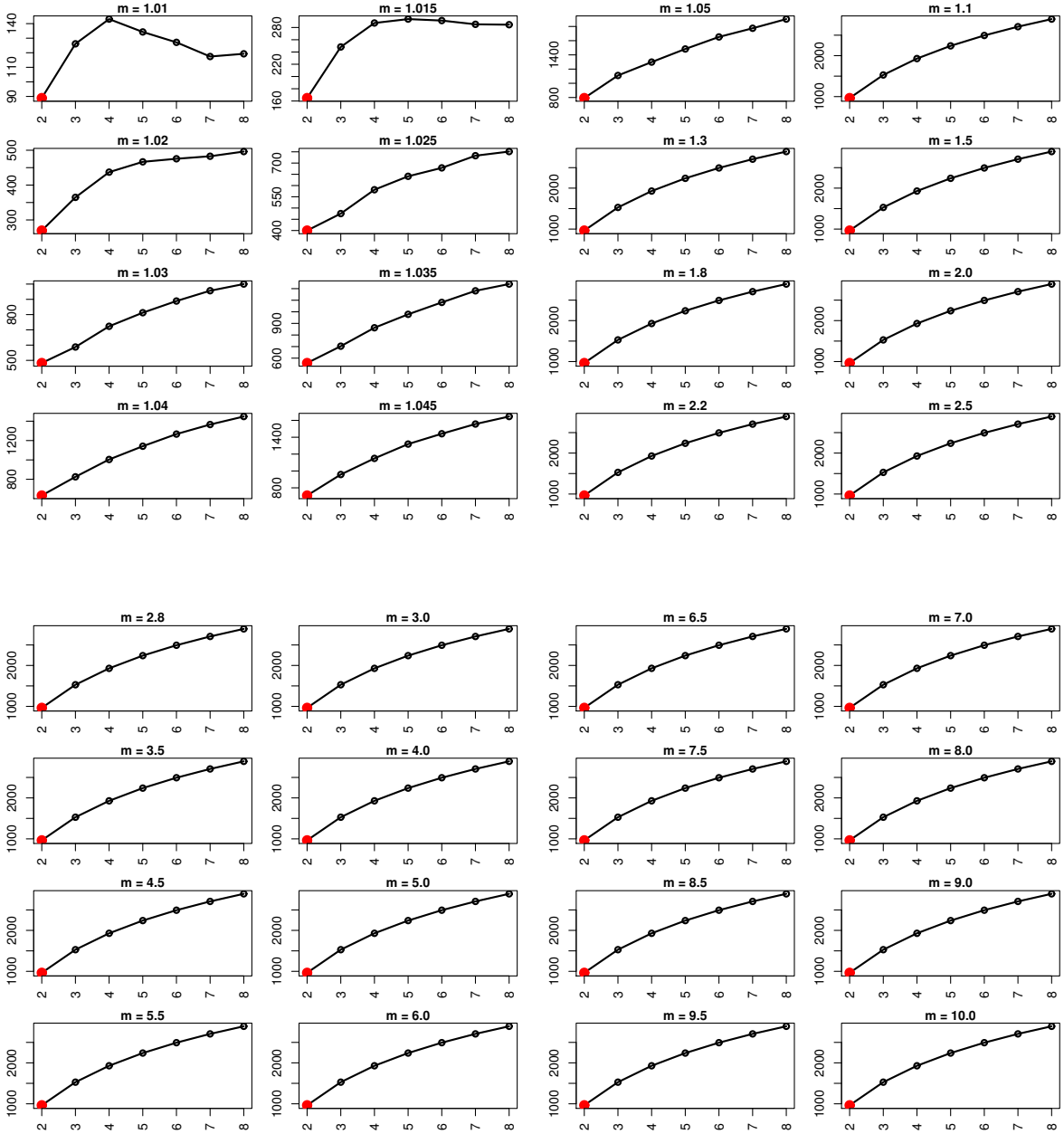
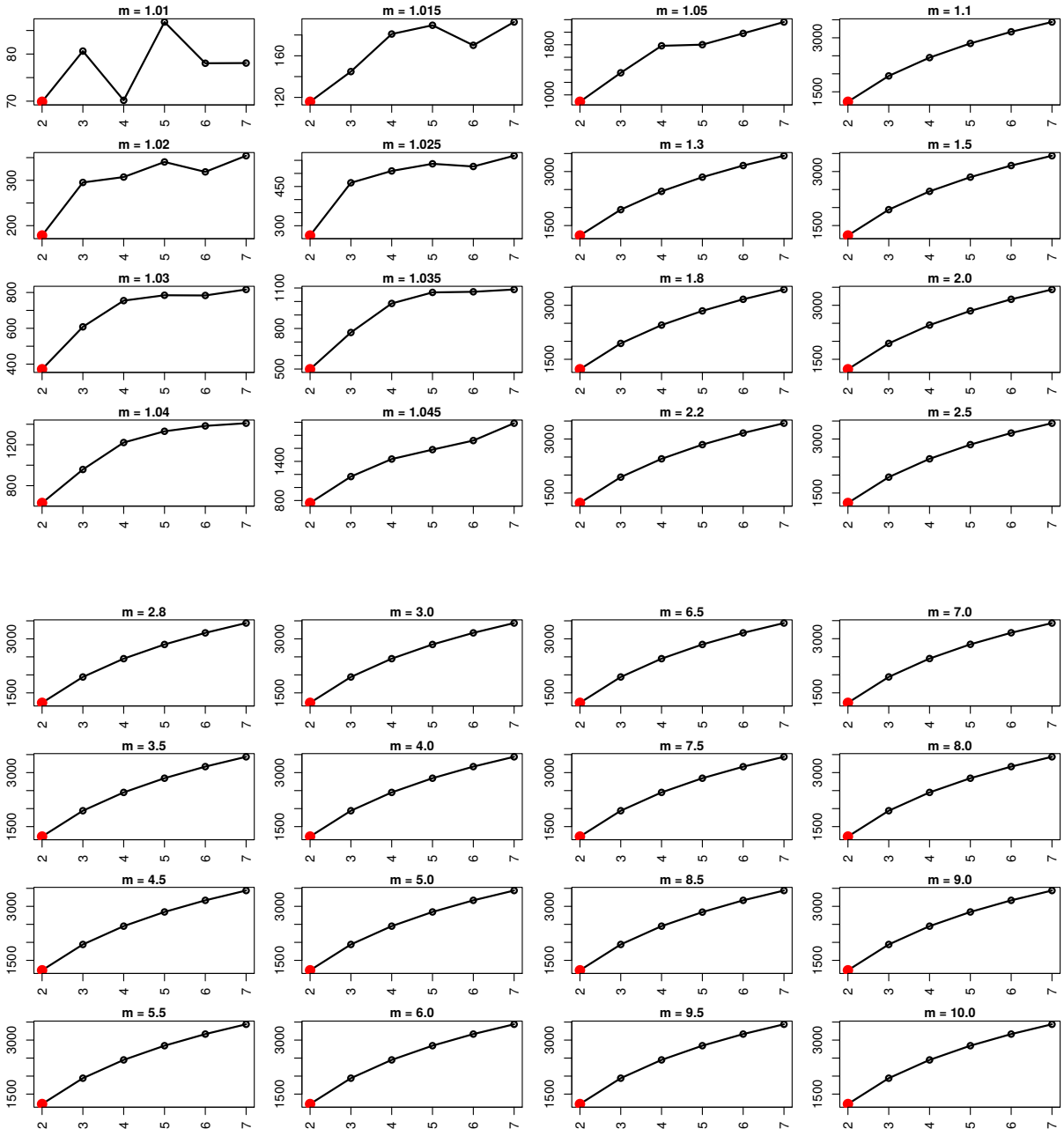


Tabela 52 – Reviews



ANEXO E – P

Tabela 53 – NewYorkTimes

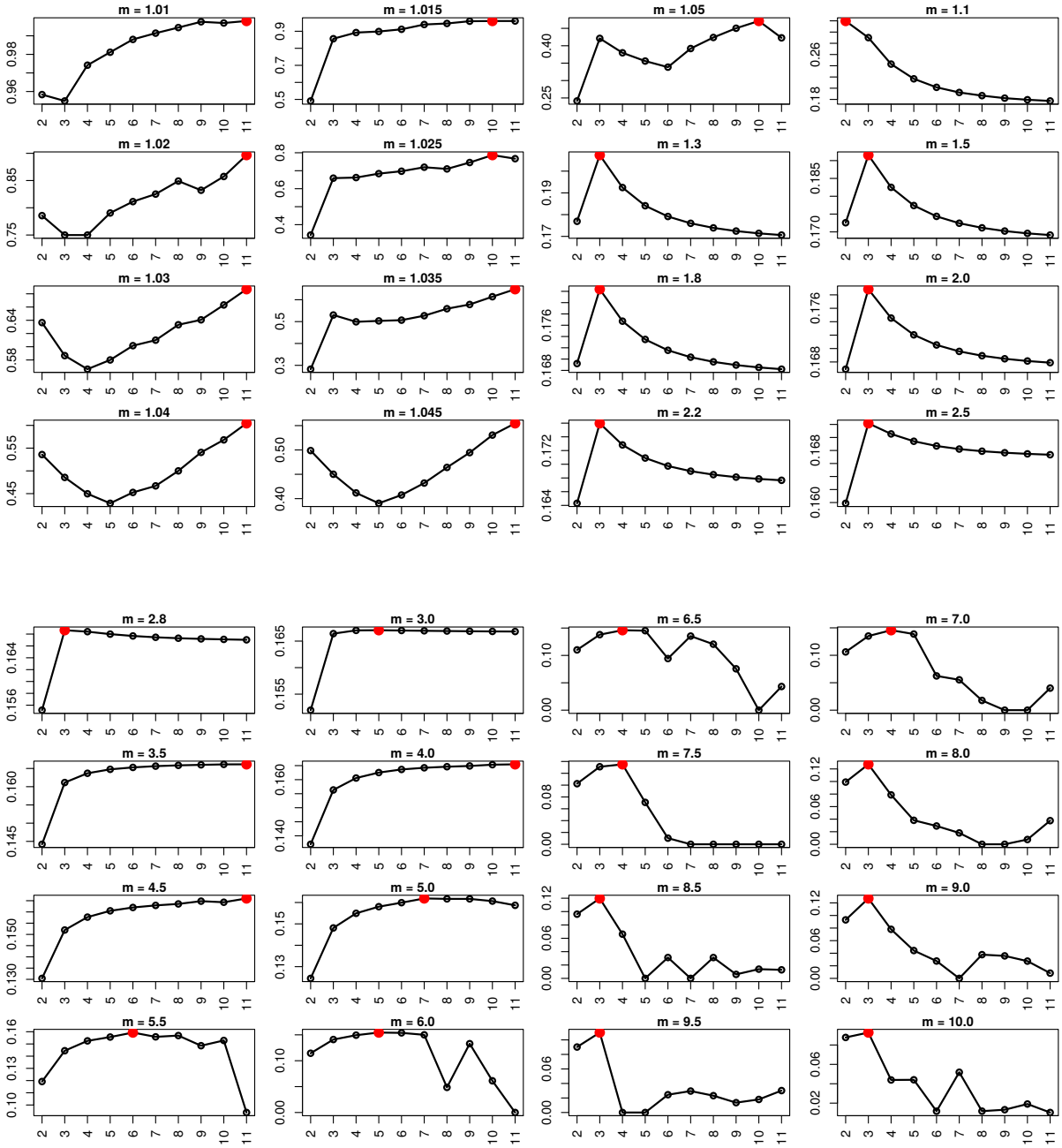


Tabela 54 – IAarticles

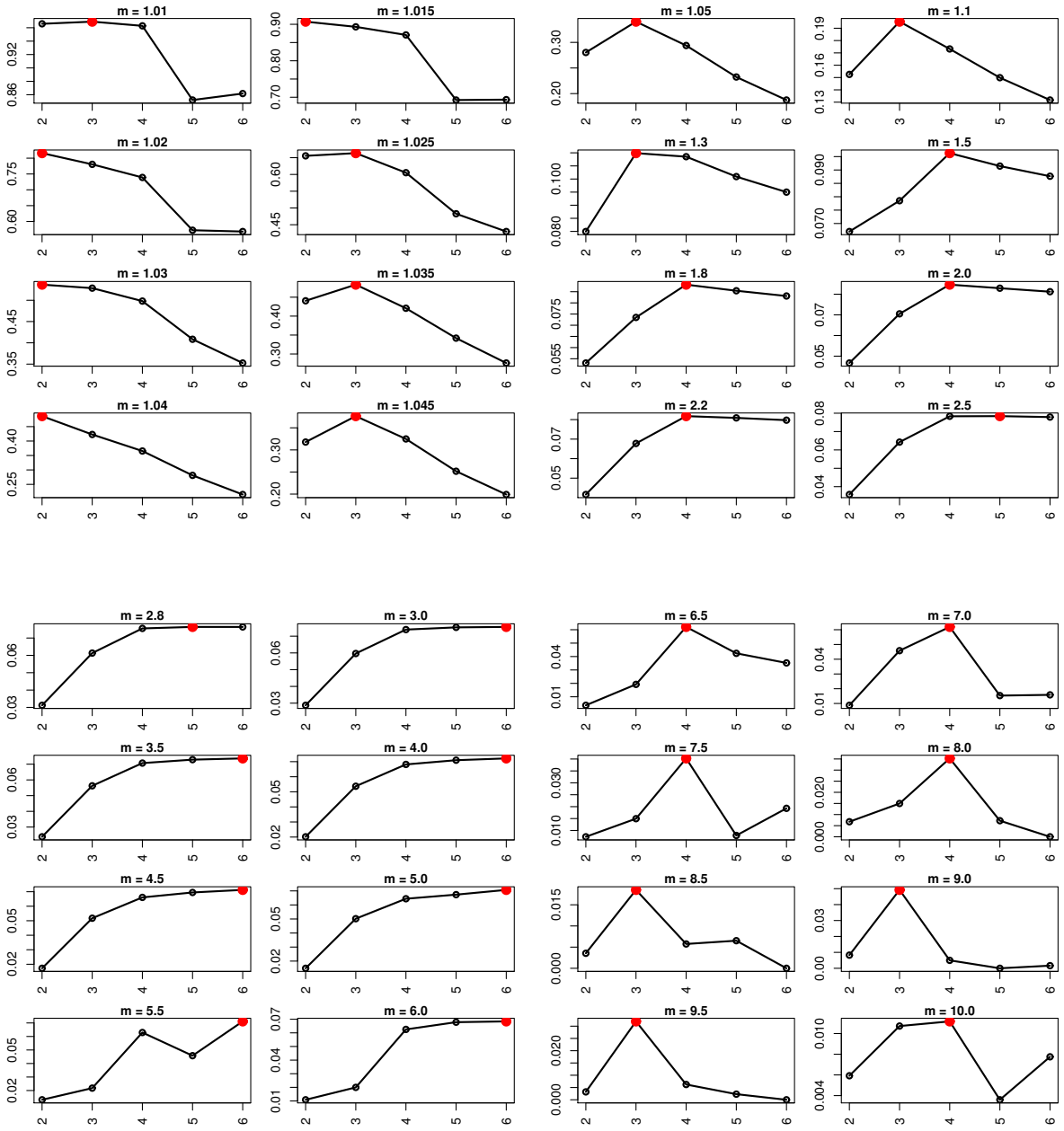


Tabela 55 – Opinosis

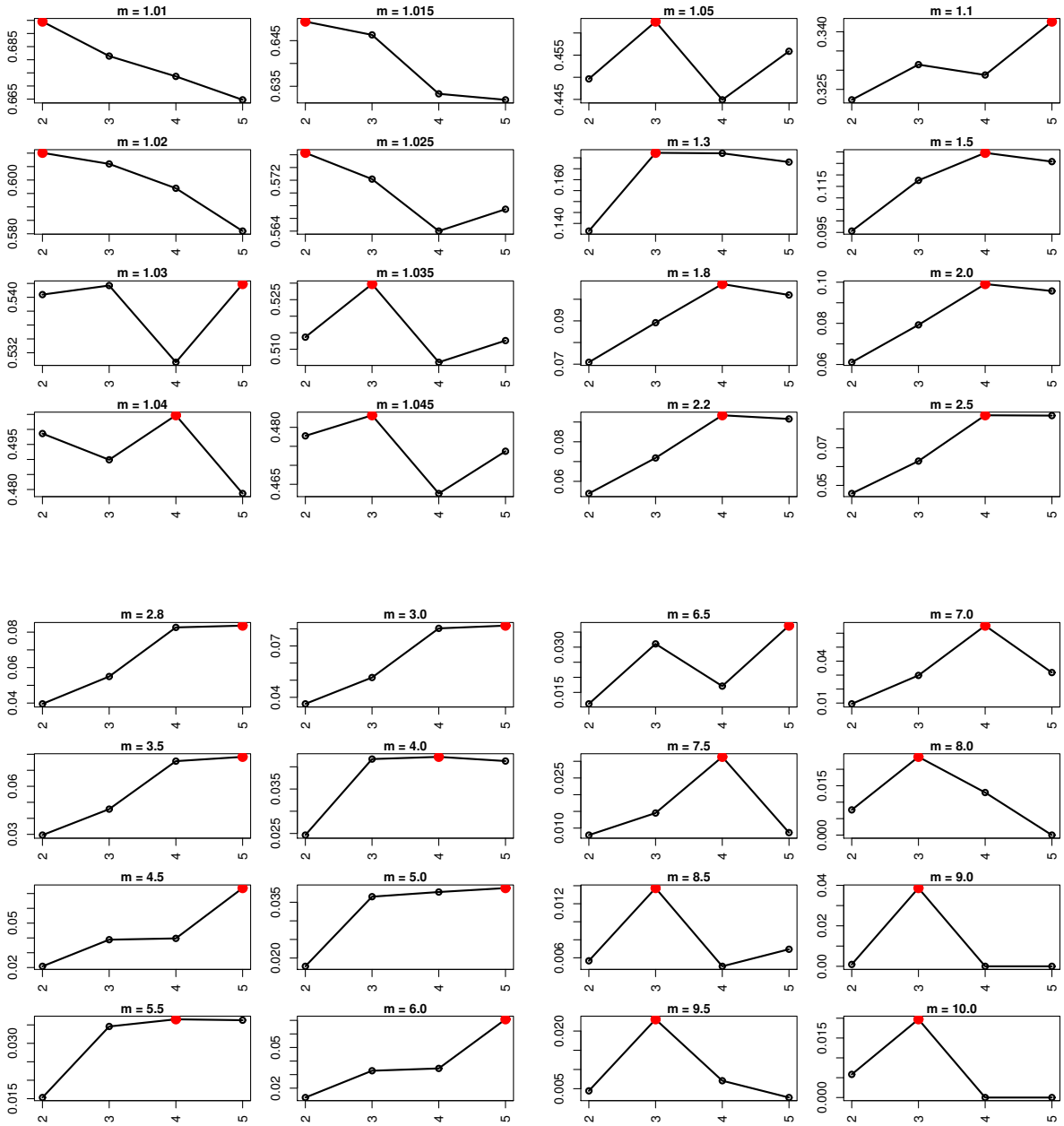


Tabela 56 – CSTR

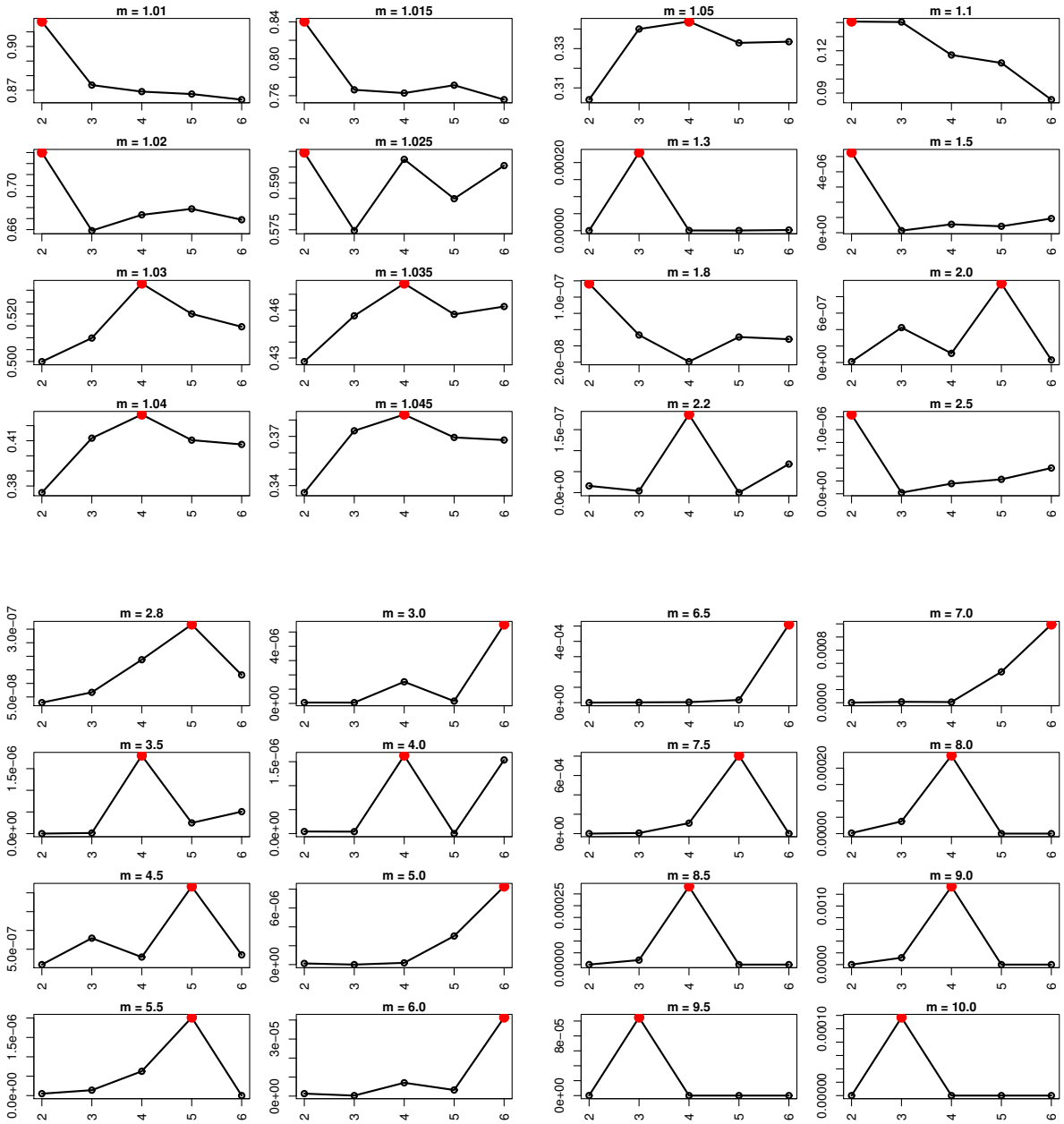


Tabela 57 – SyskillWebert

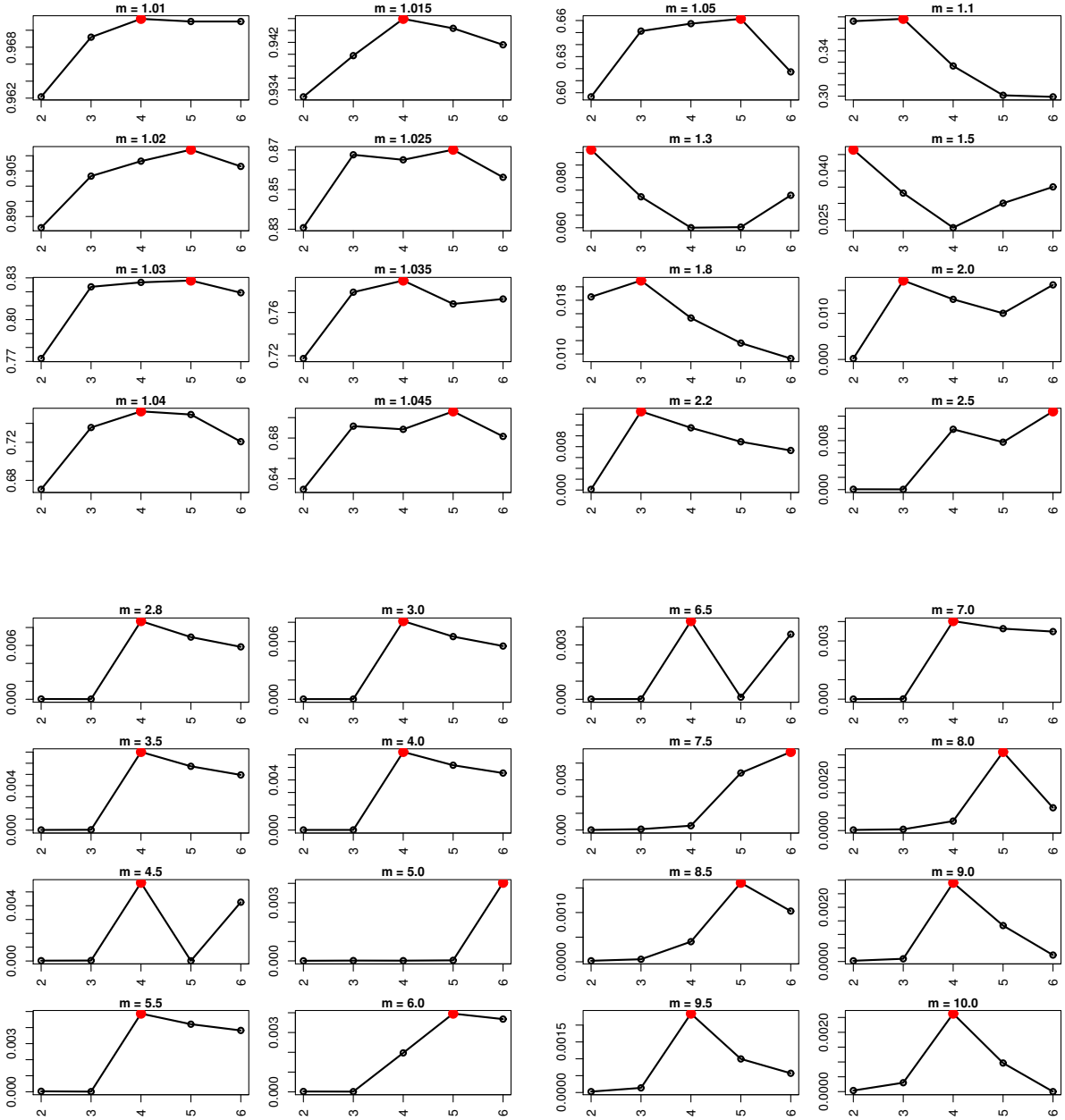


Tabela 58 – Hitech

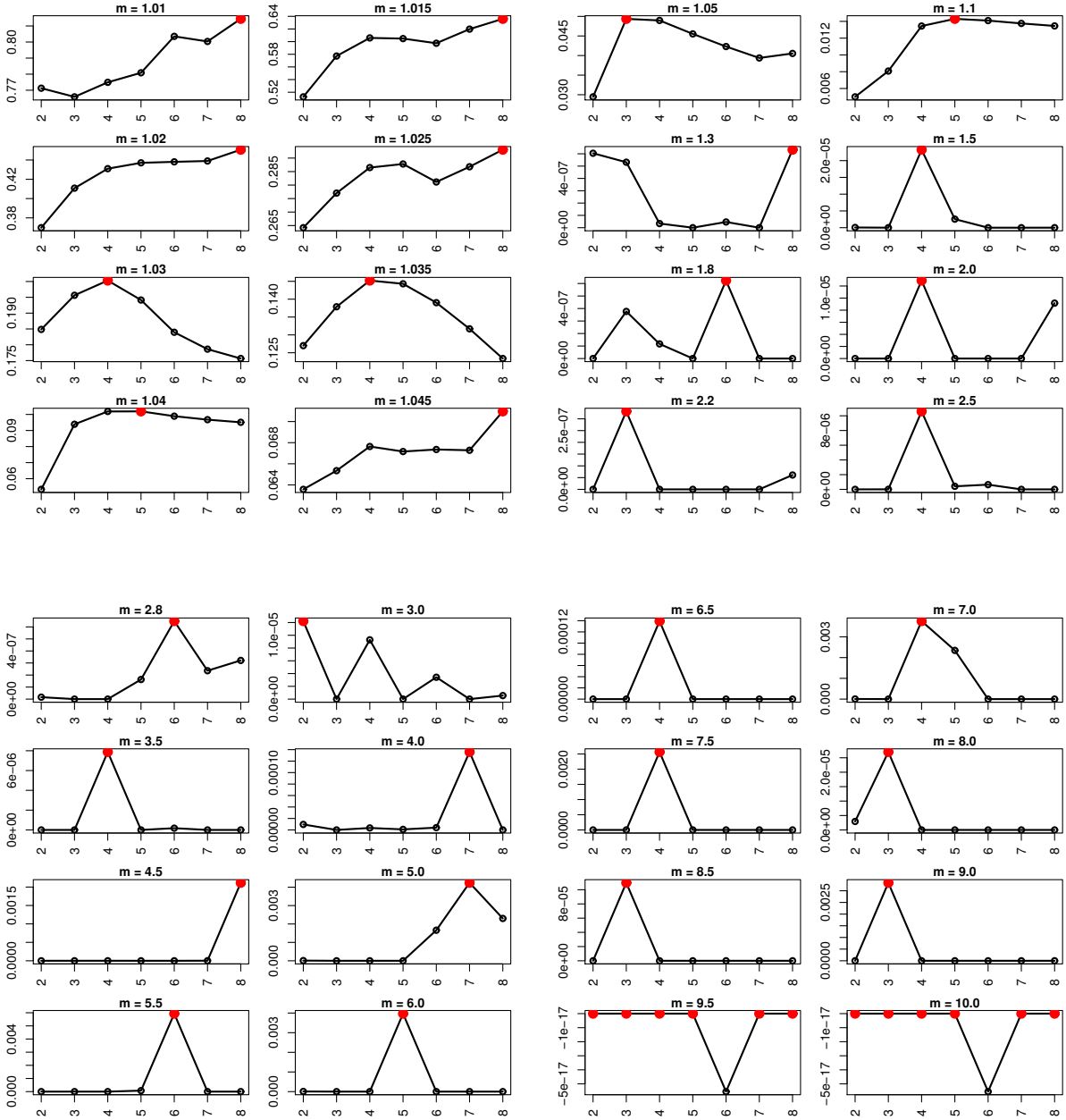


Tabela 59 – WAP

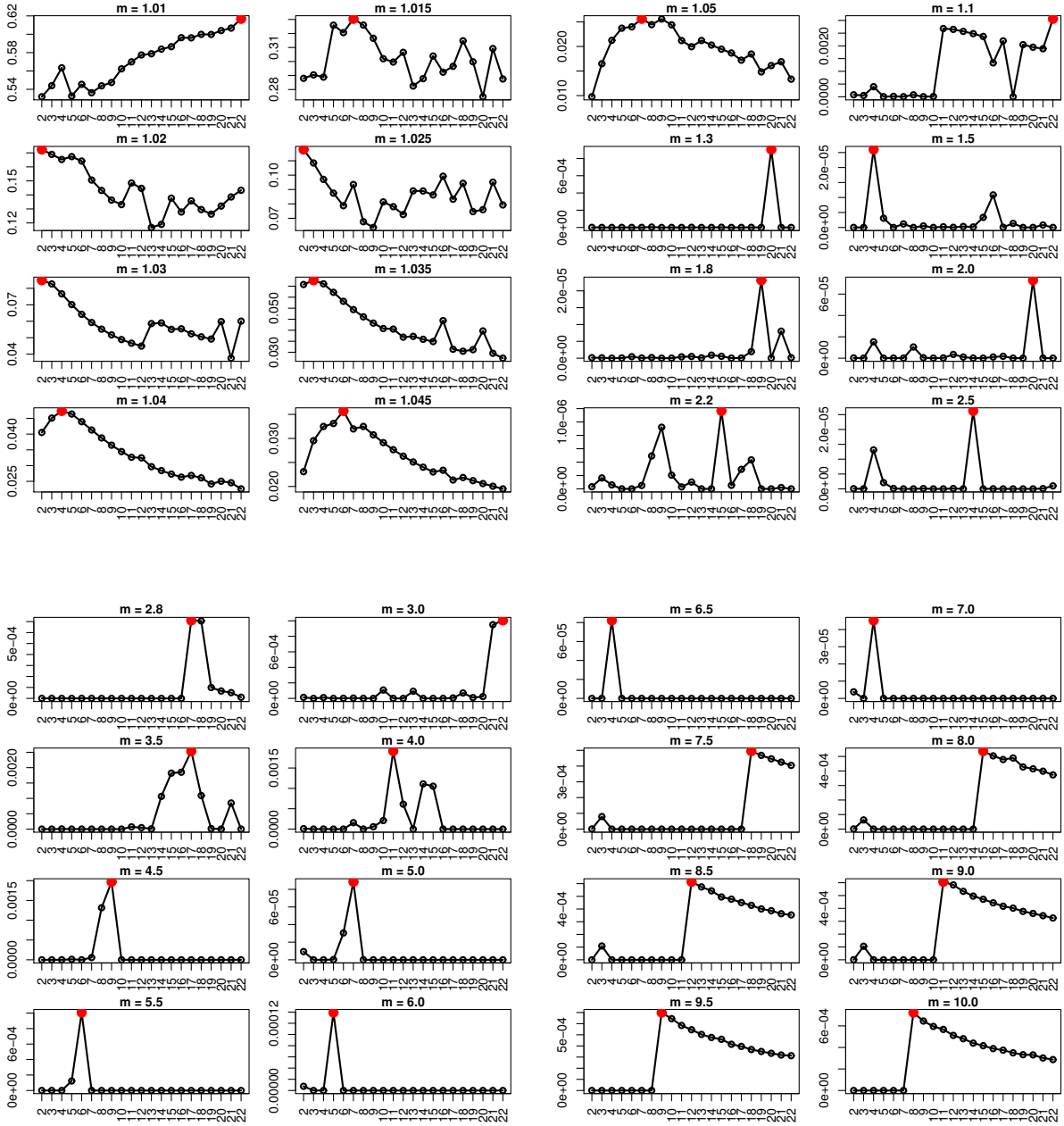


Tabela 60 – NSF

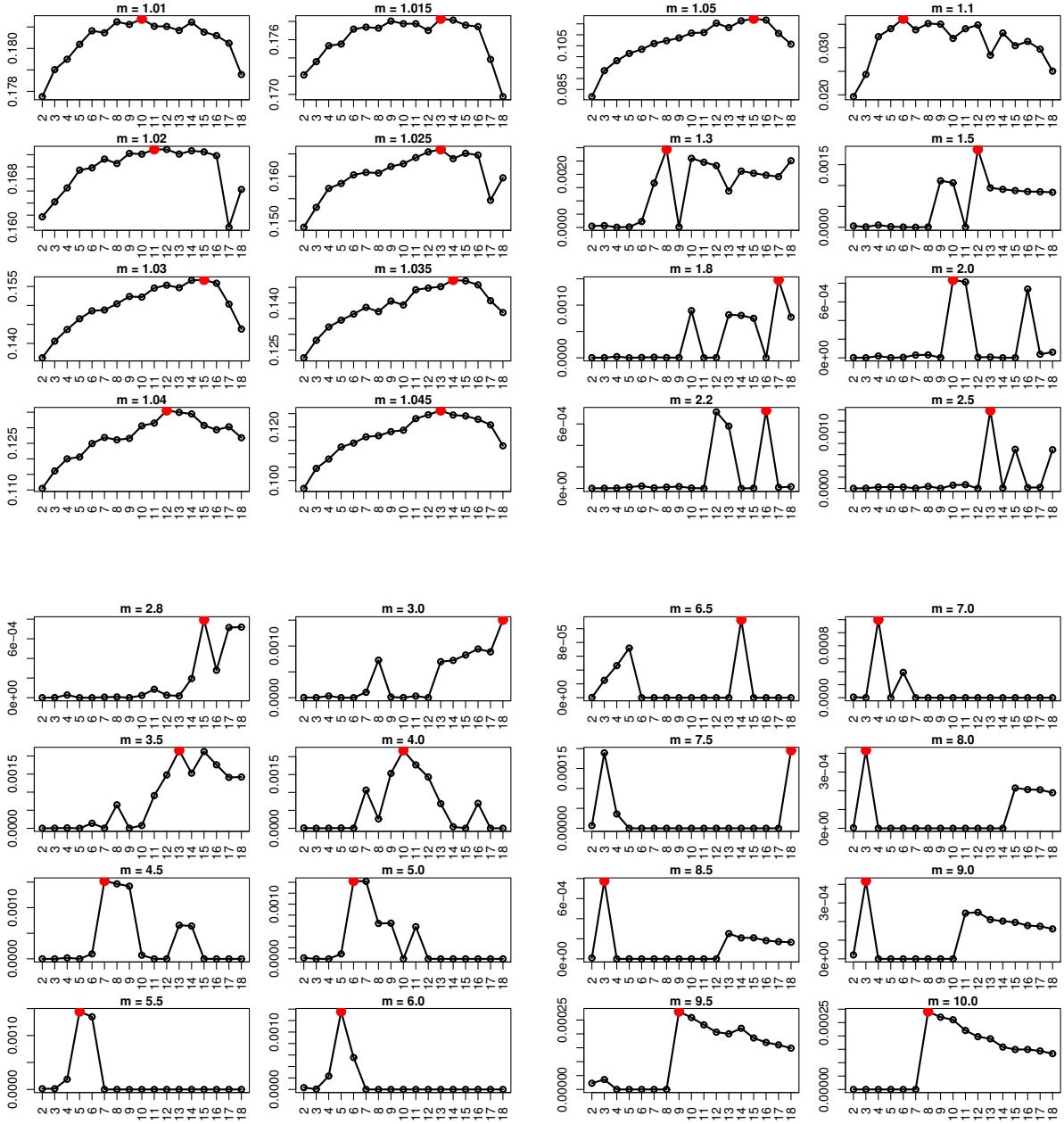


Tabela 61 – Irish-Sentiment

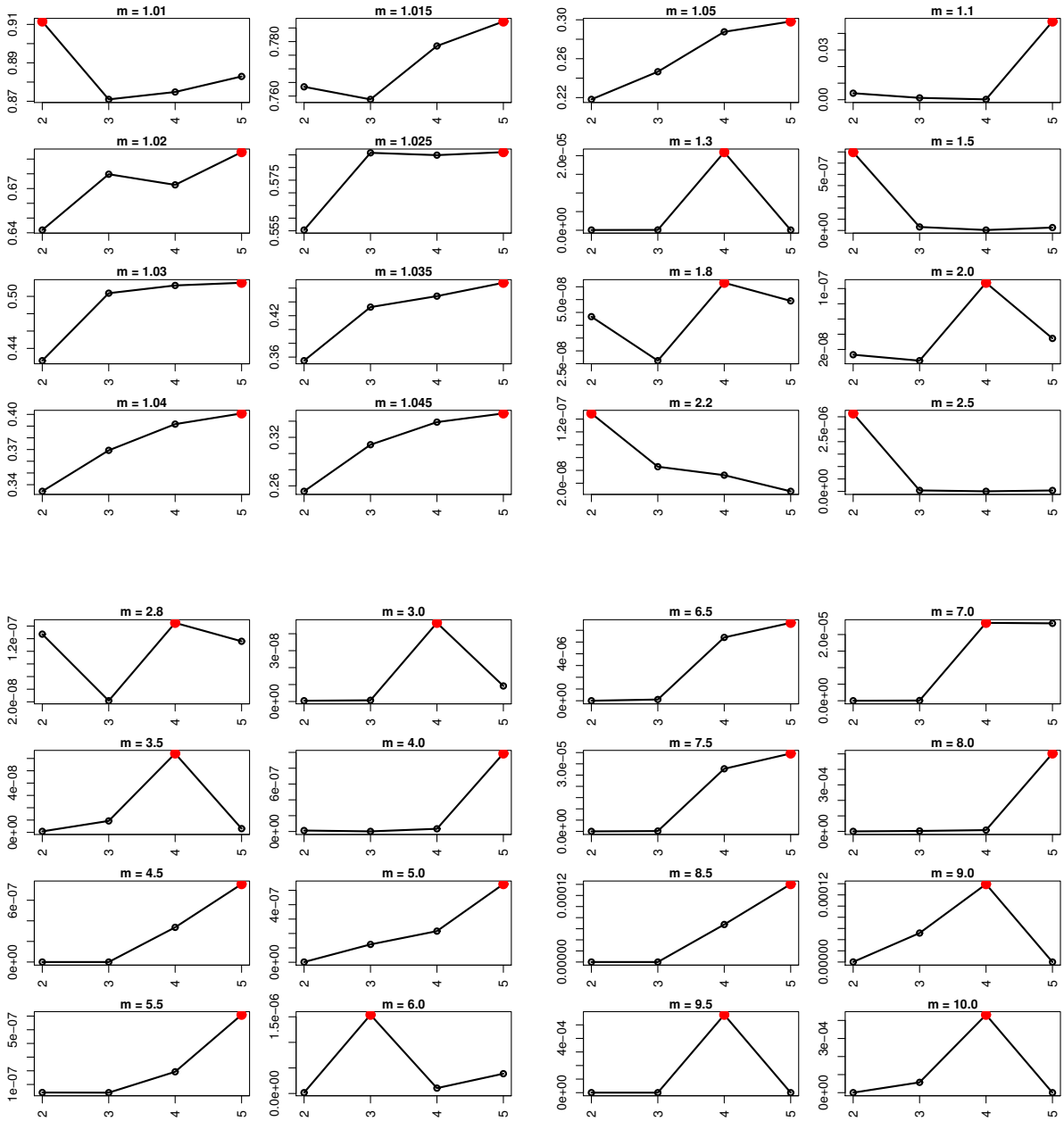


Tabela 62 – 20Newsgroups

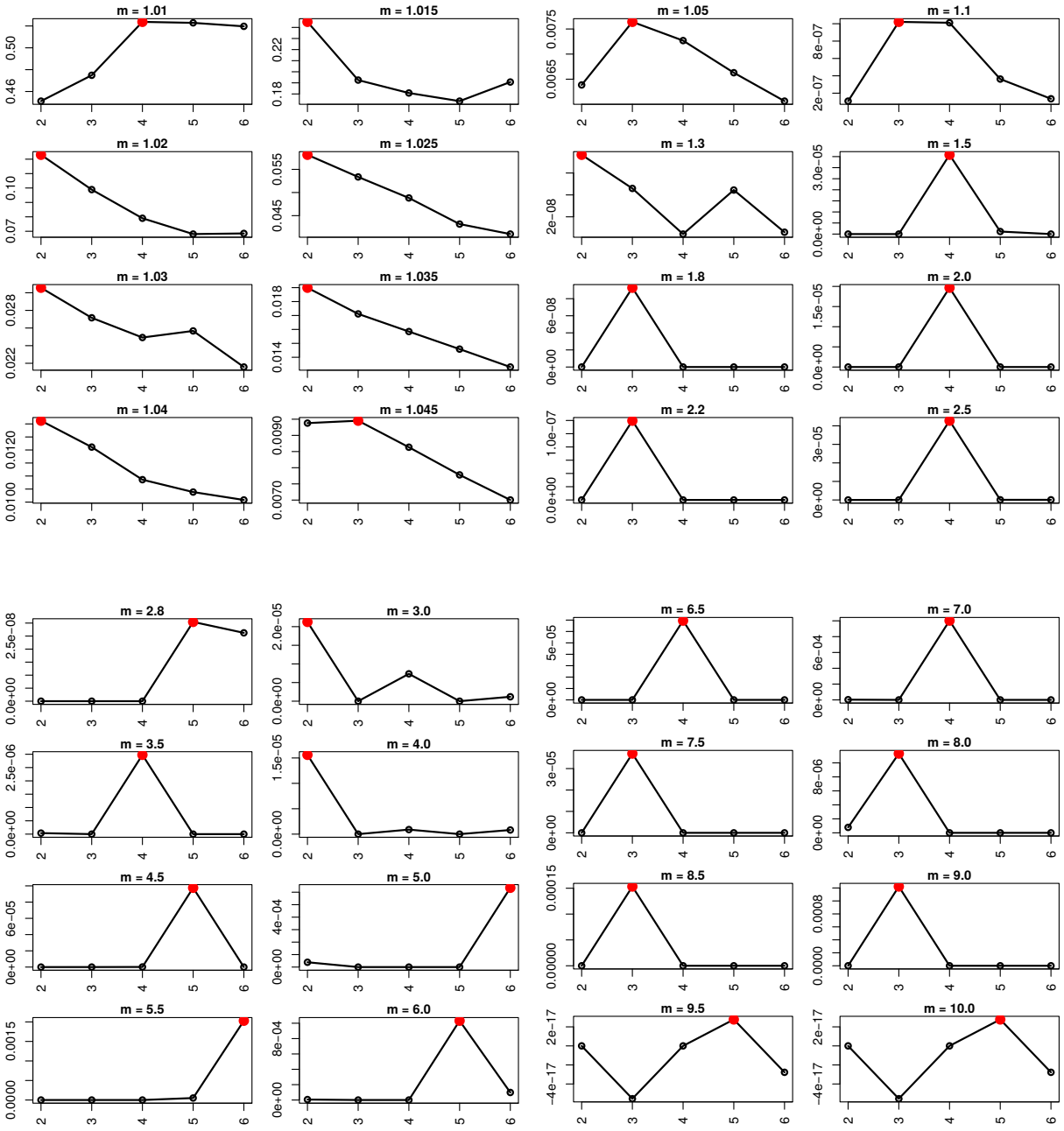


Tabela 63 – La1s

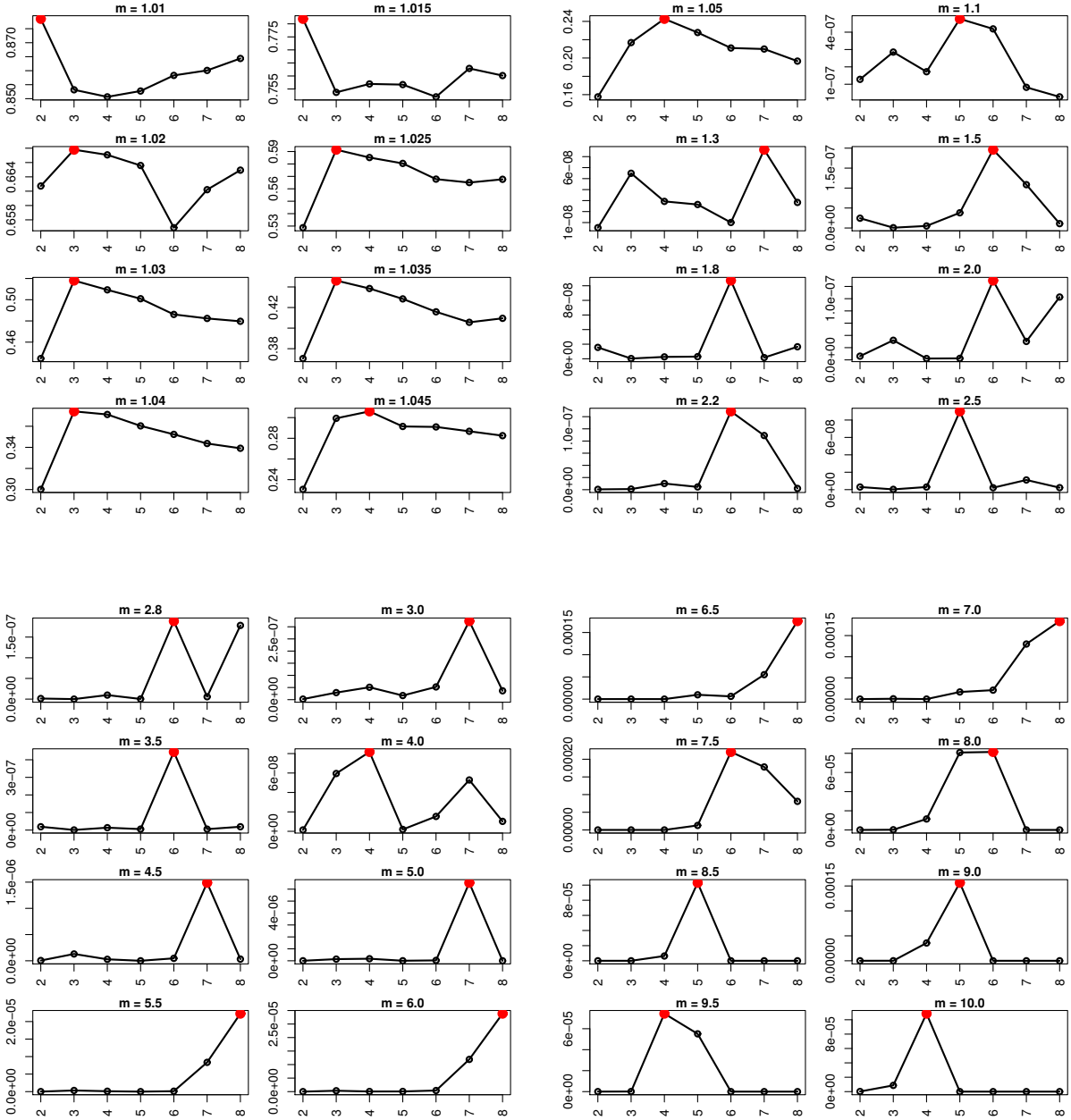
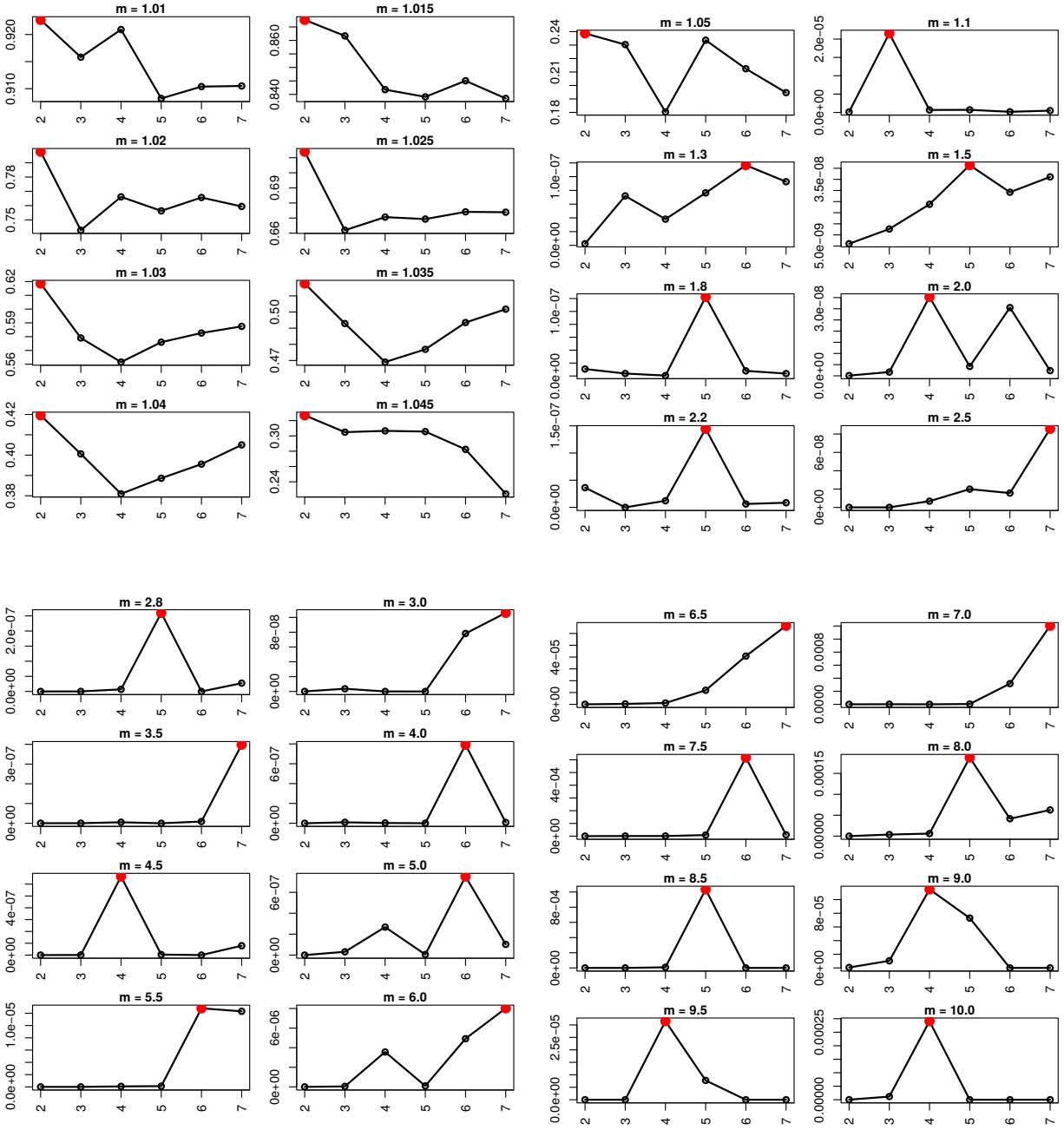


Tabela 64 – Reviews



ANEXO F – MPO

Tabela 65 – NewYorkTimes

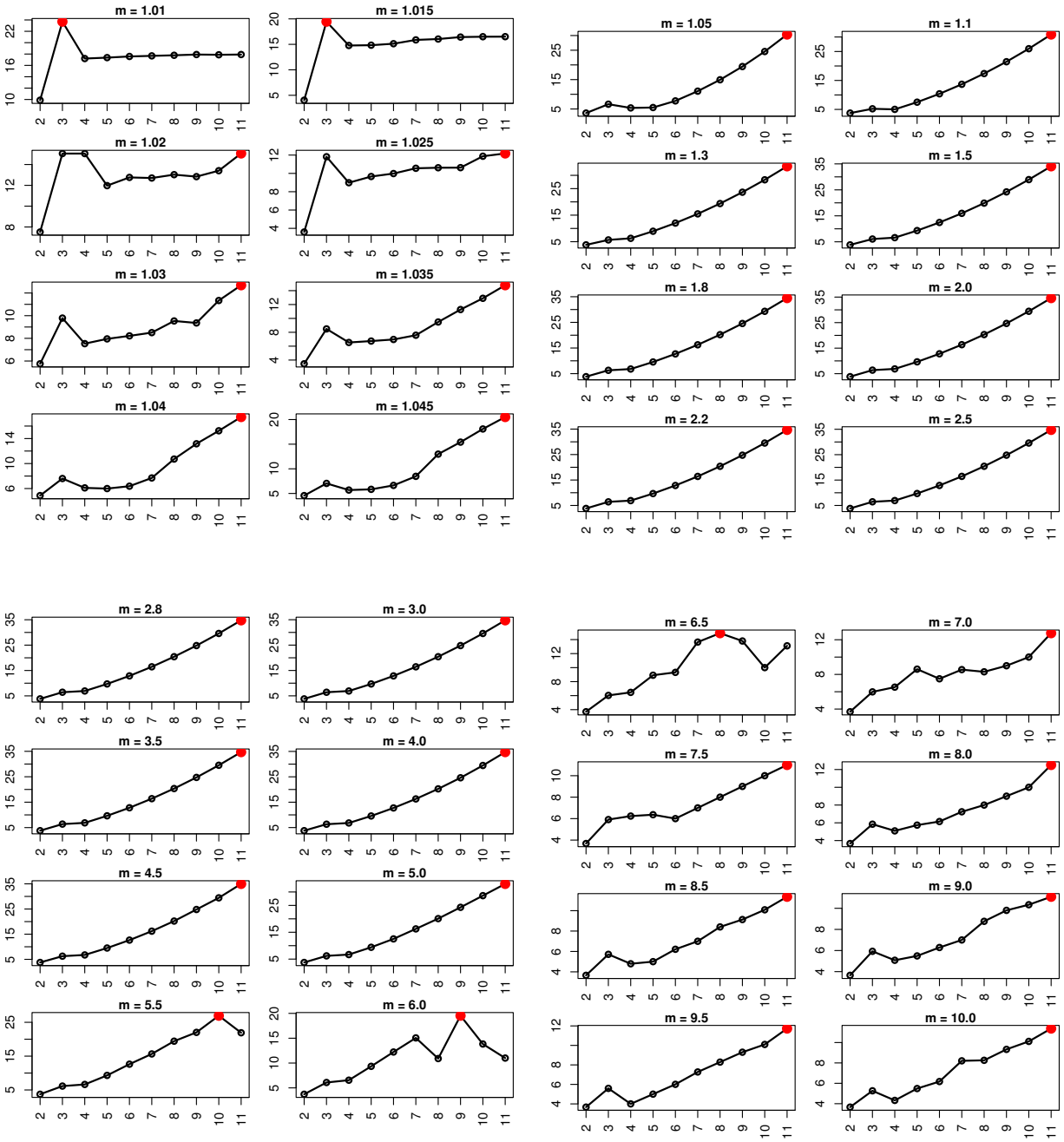


Tabela 66 – IAarticles

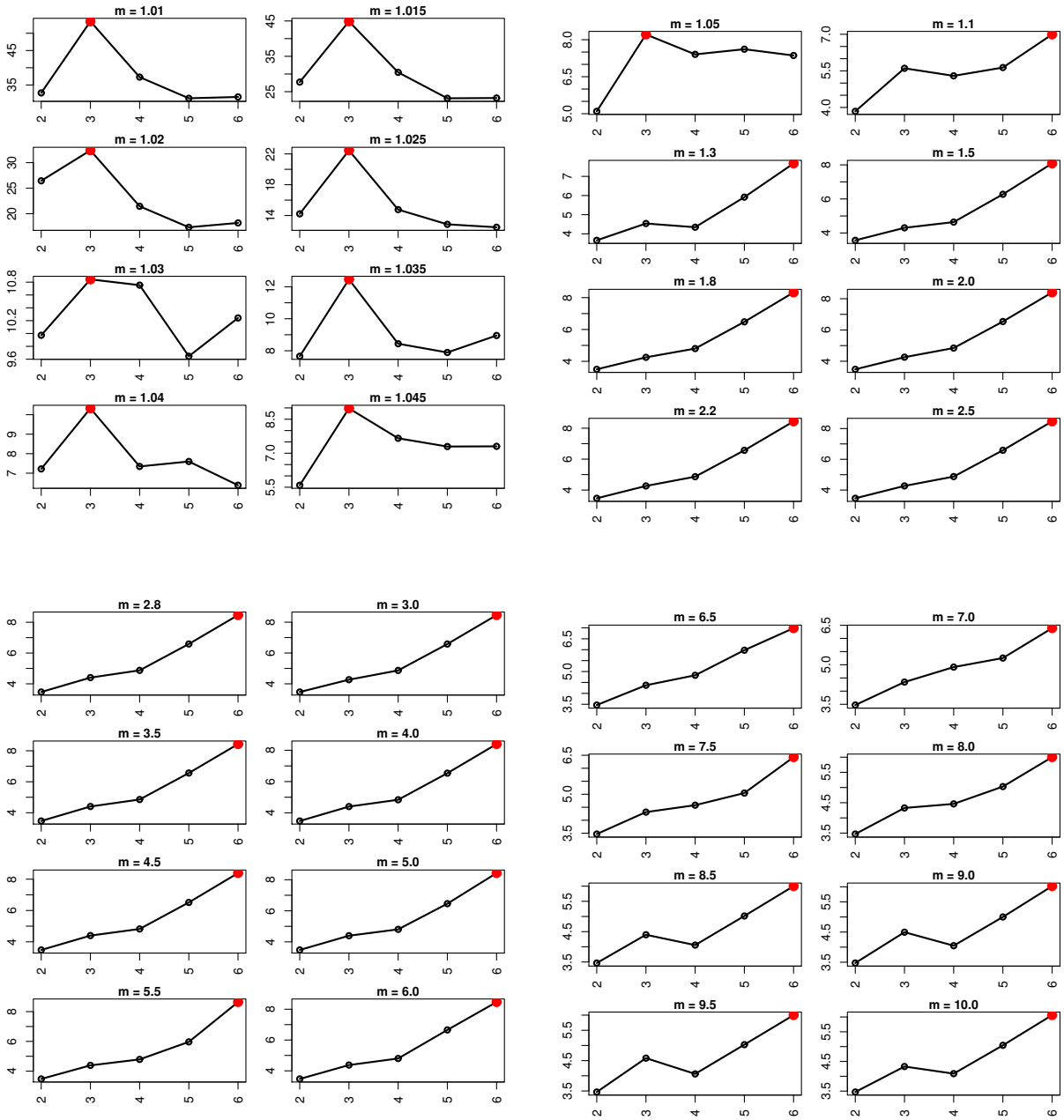
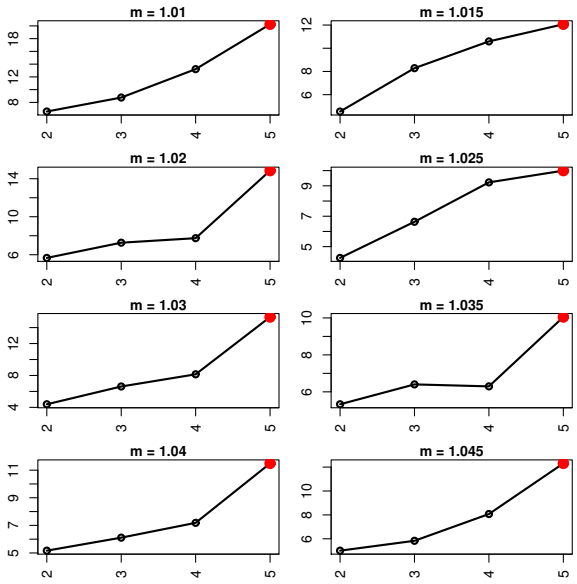
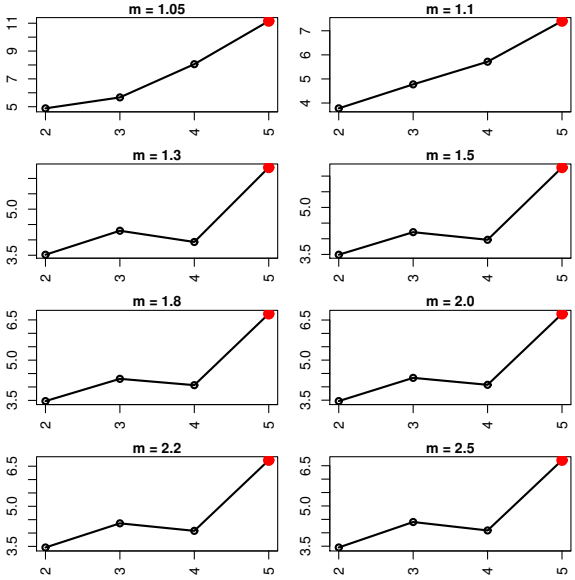
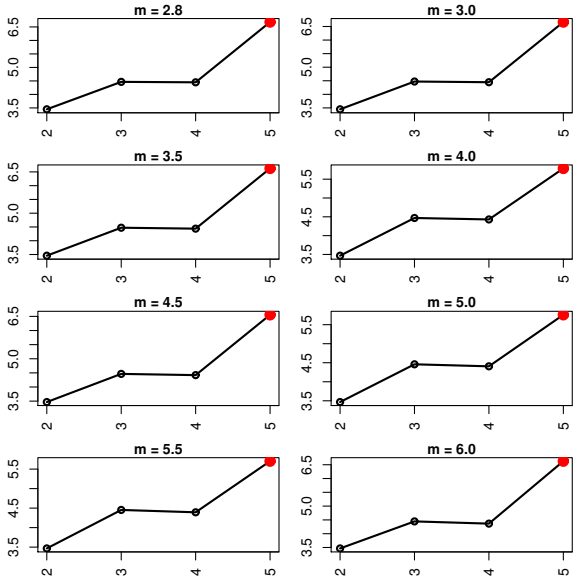


Tabela 67 – Opinosis







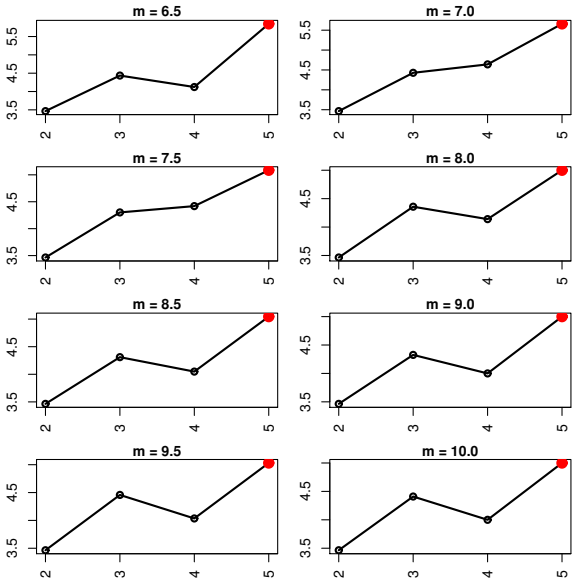


Tabela 68 – CSTR

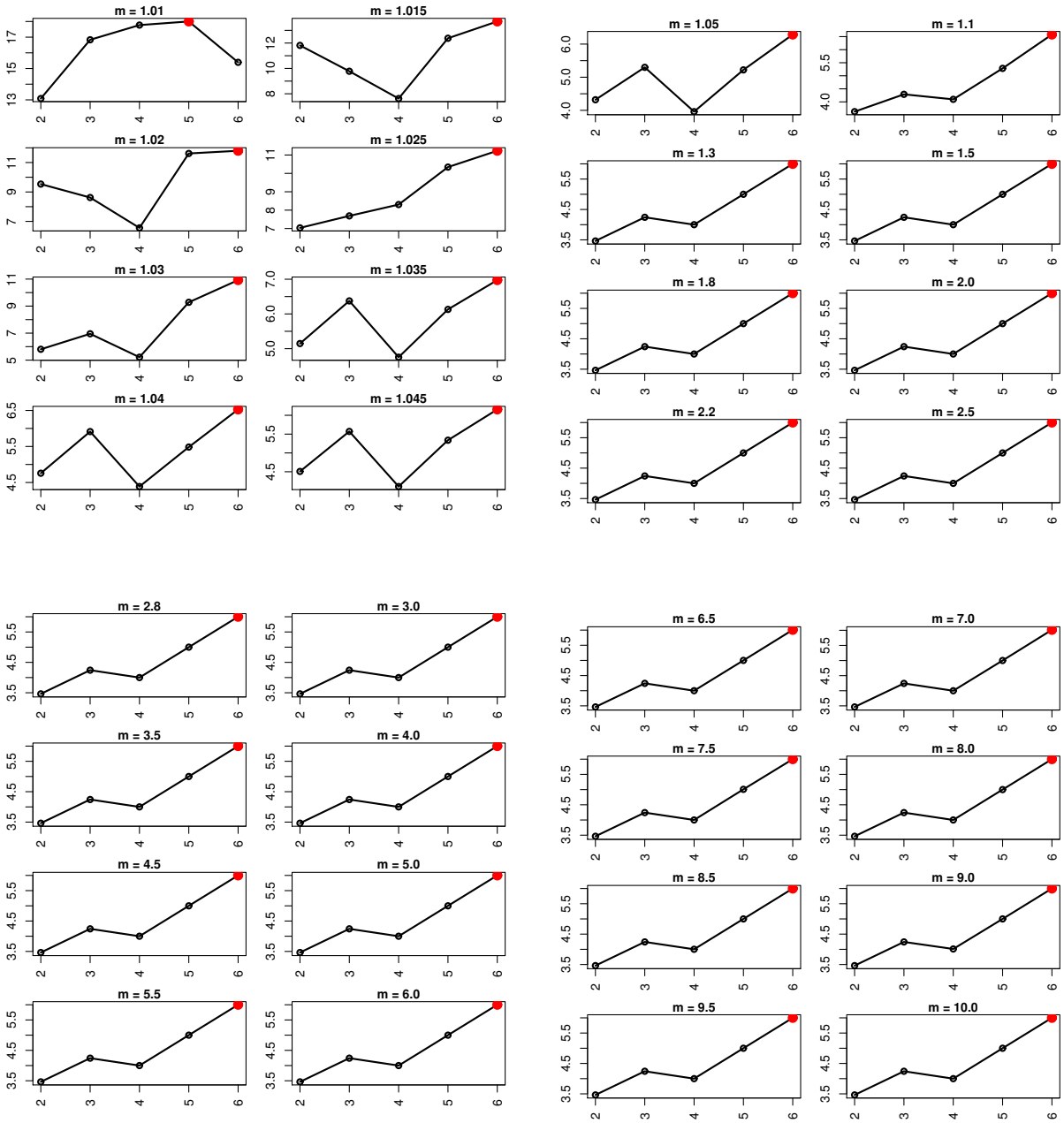
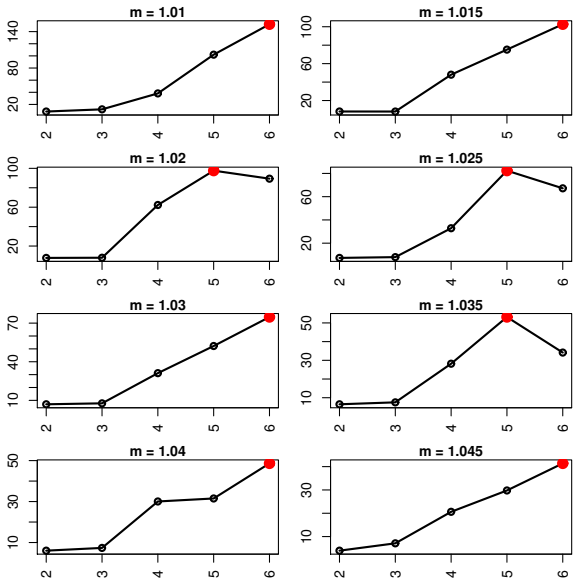
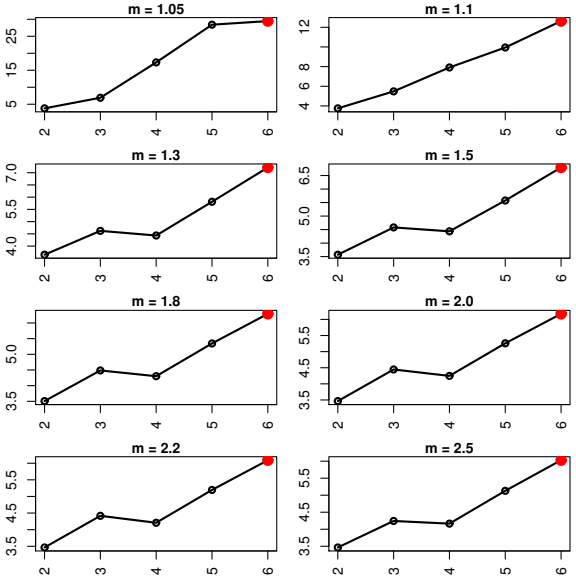
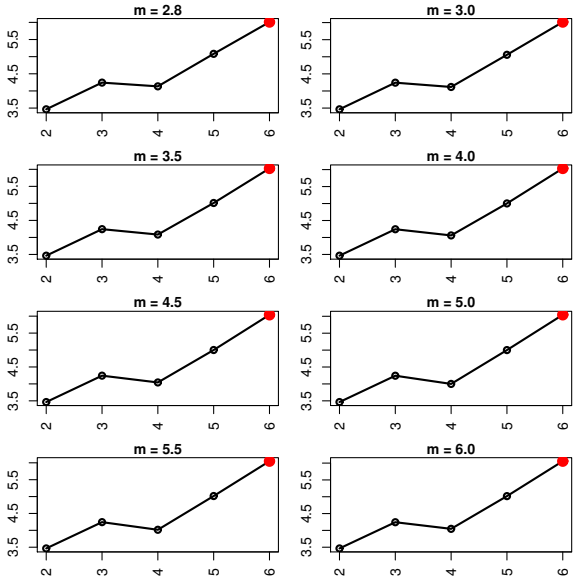


Tabela 69 – SyskillWebert







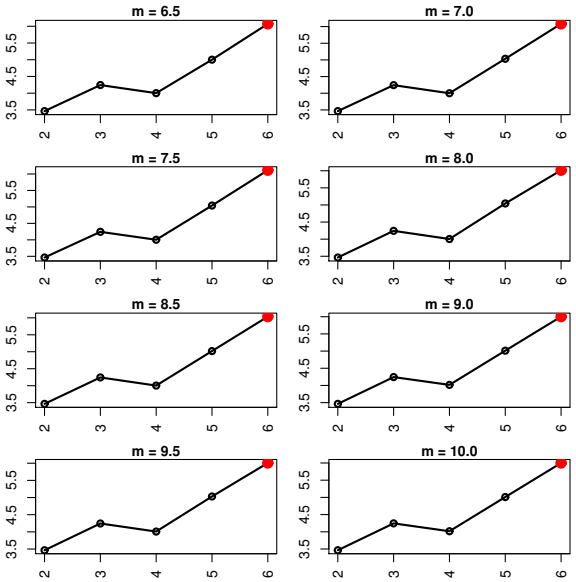


Tabela 70 – Hitech

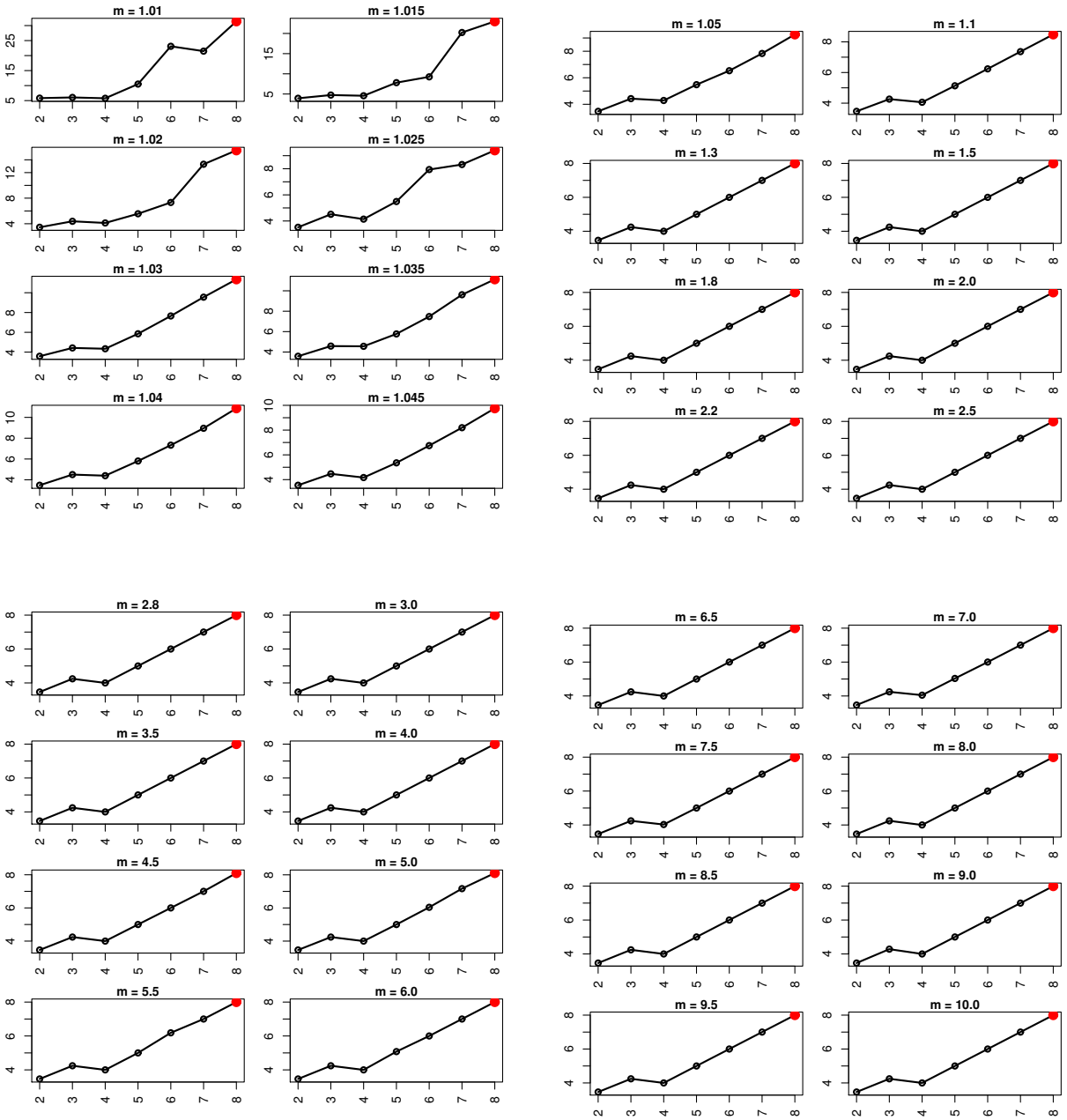


Tabela 71 – WAP

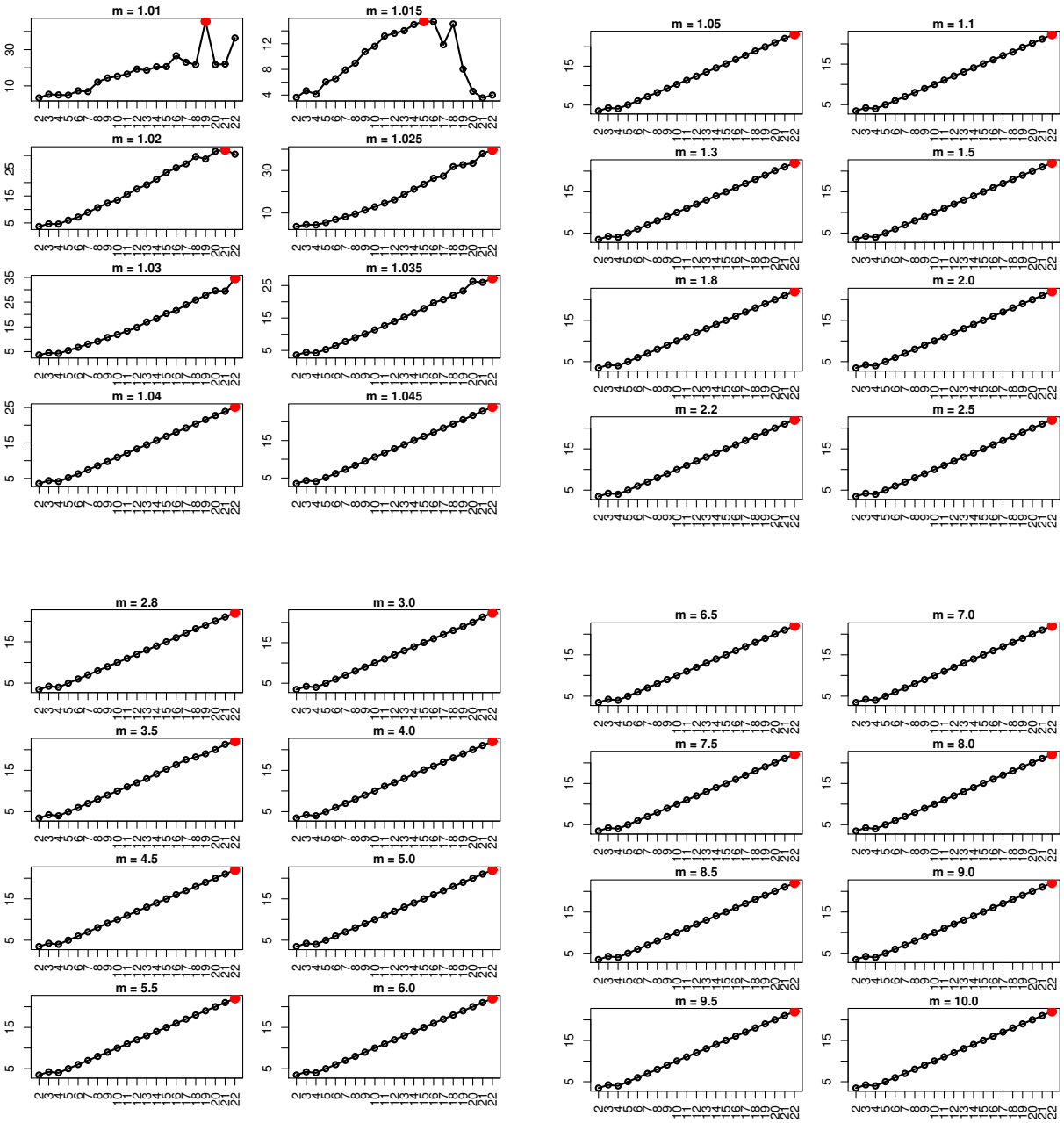


Tabela 72 – NSF

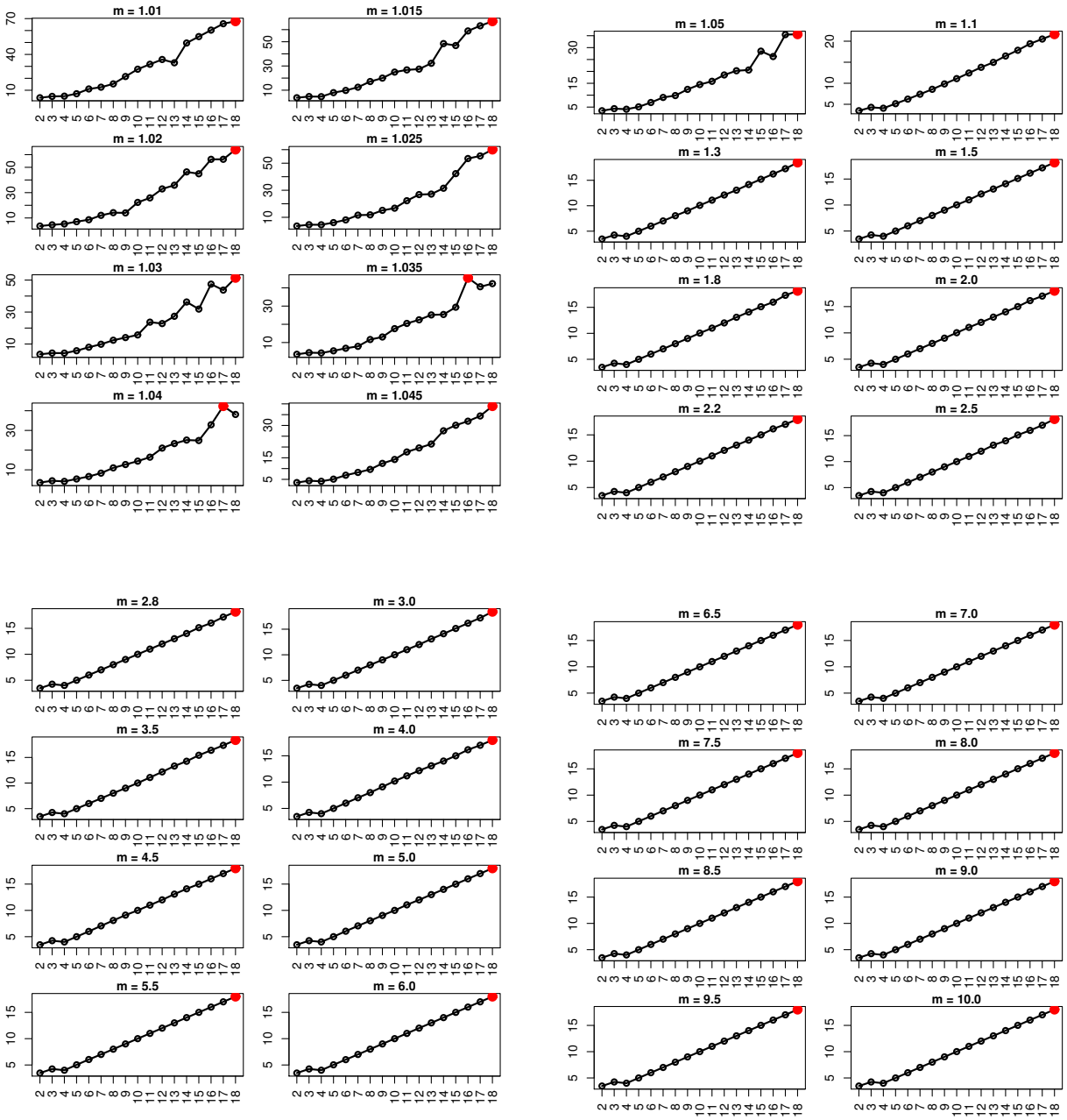


Tabela 73 – Irish-Sentiment

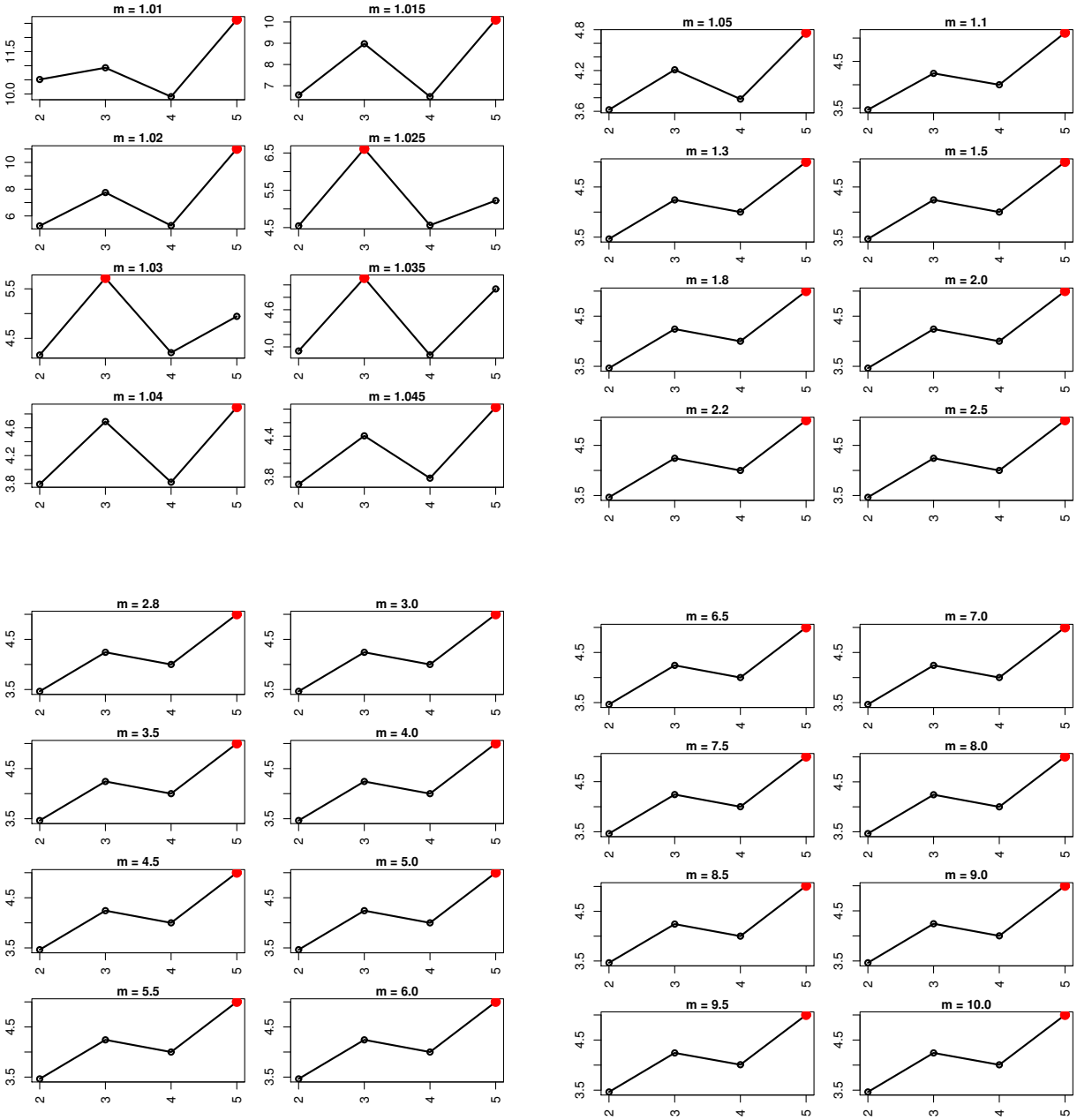


Tabela 74 – 20Newsgroups

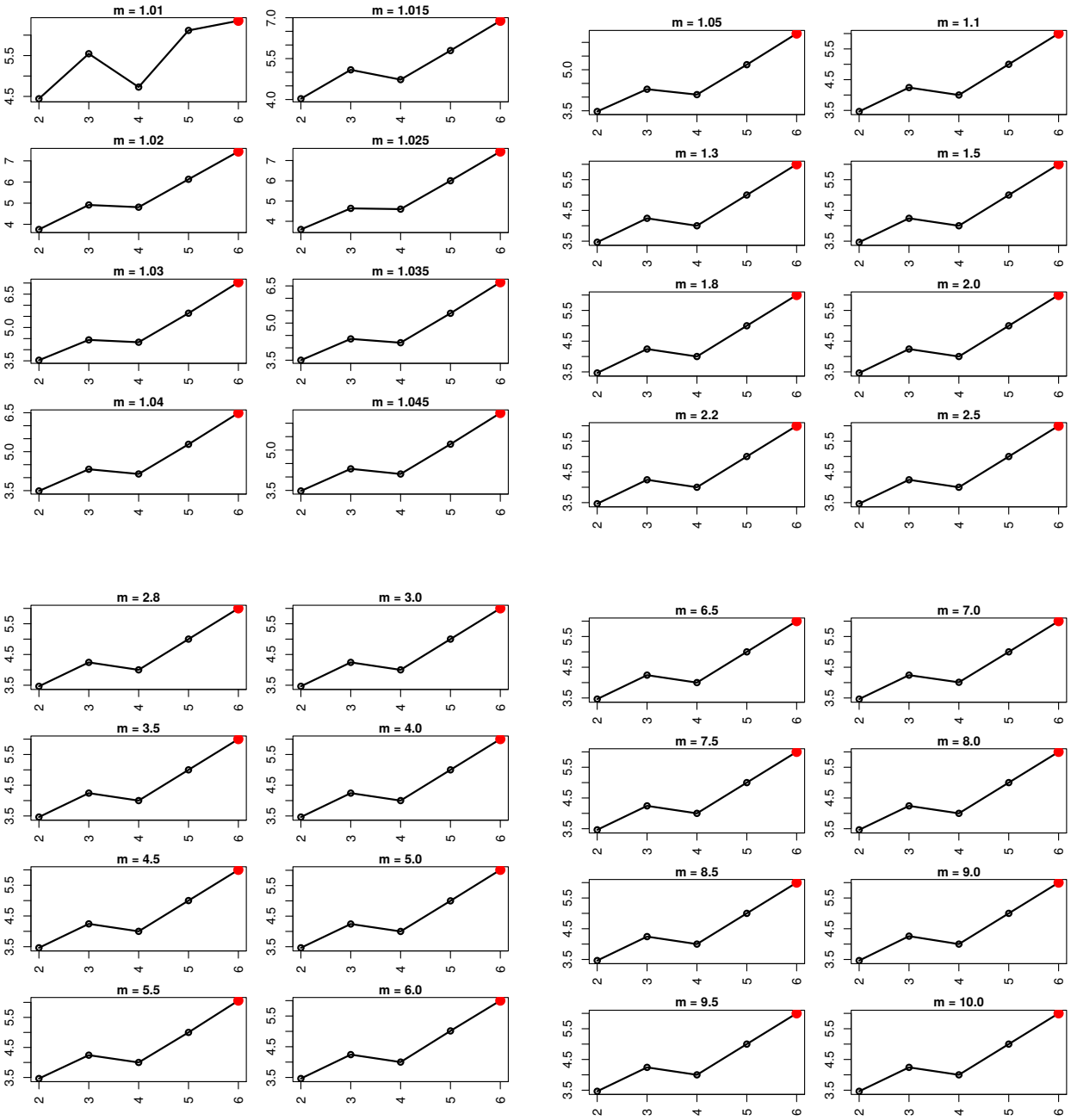


Tabela 75 – La1s

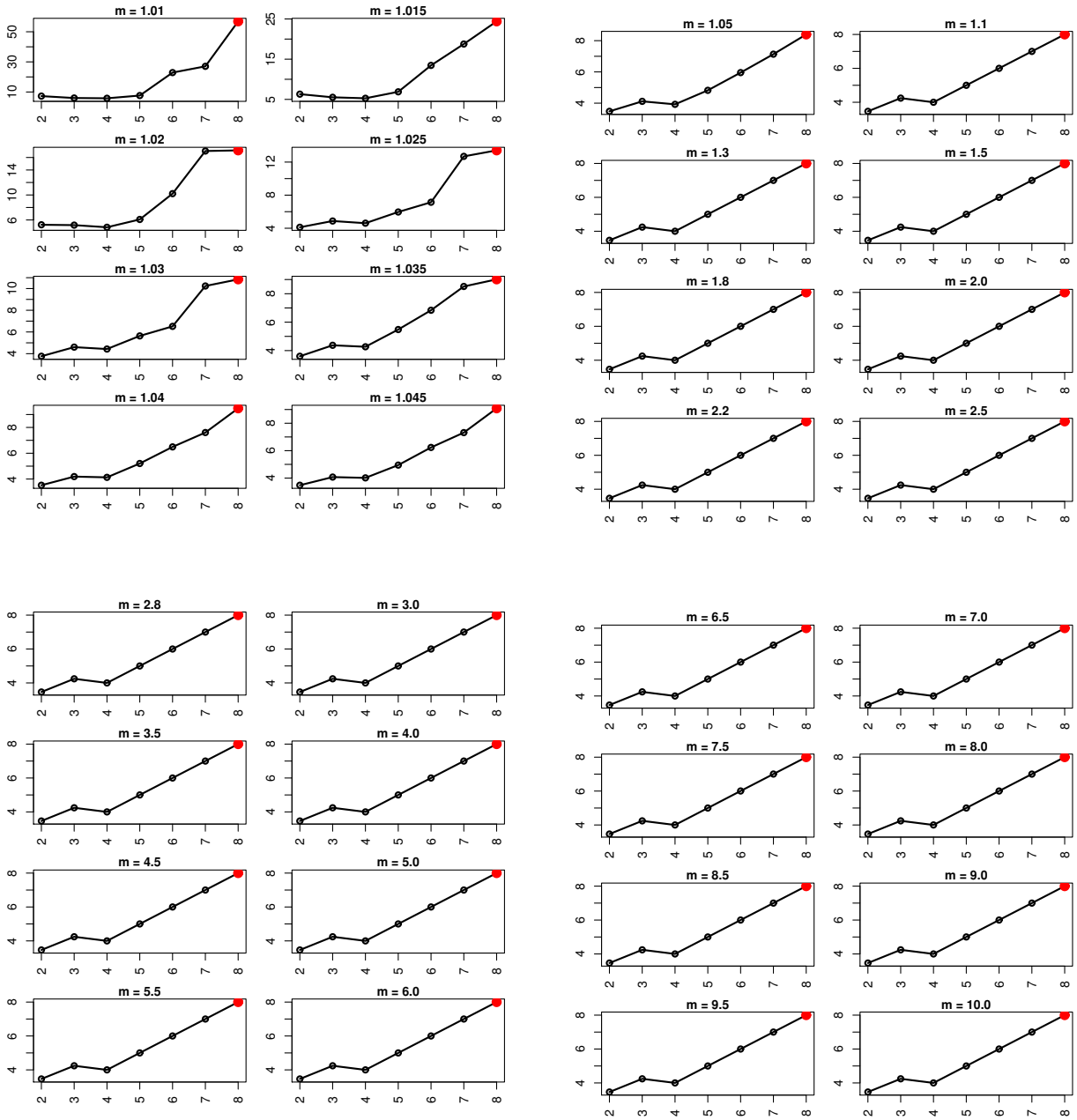
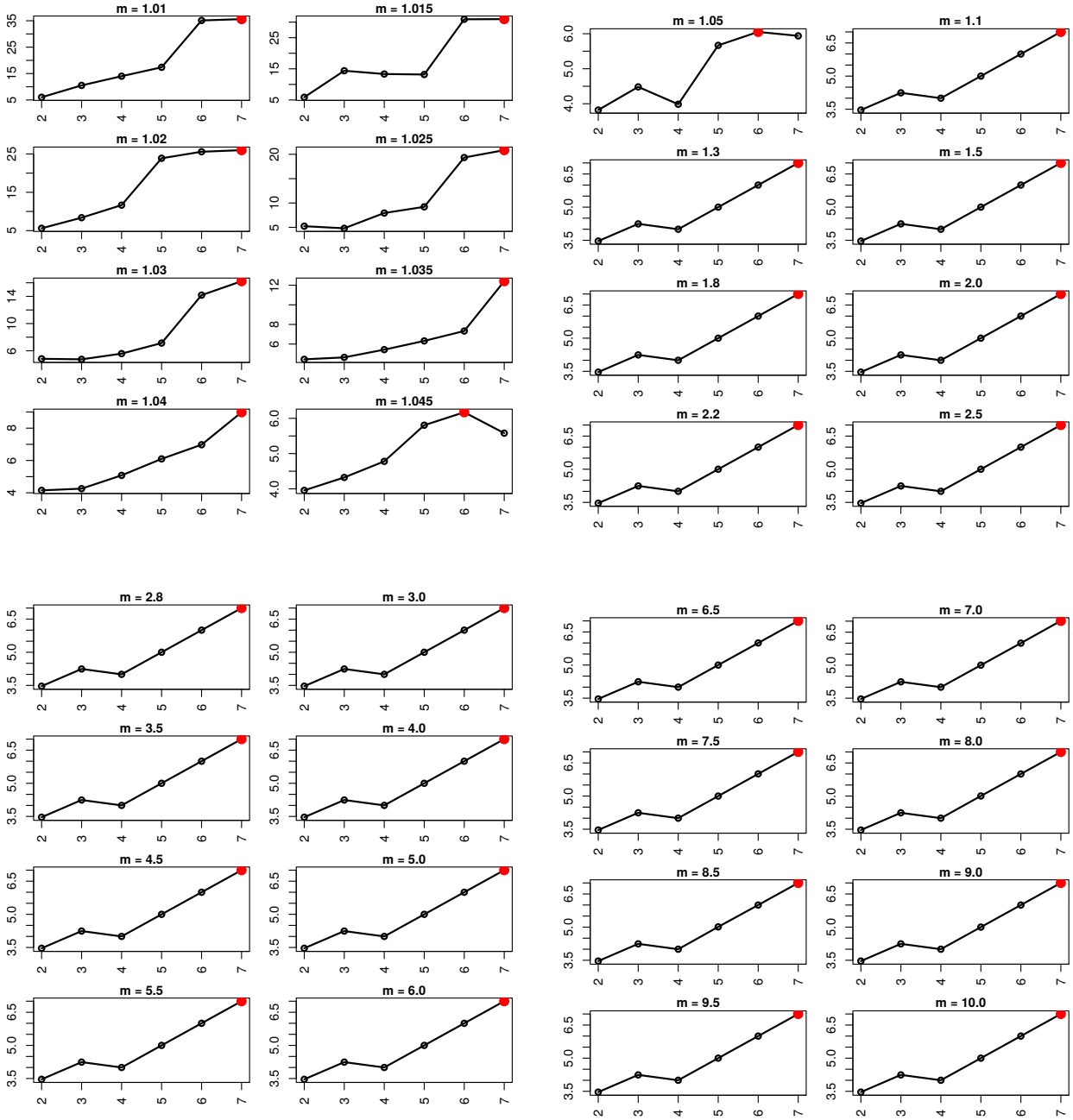


Tabela 76 – Reviews



ANEXO G – GD

Tabela 77 – NewYorkTimes

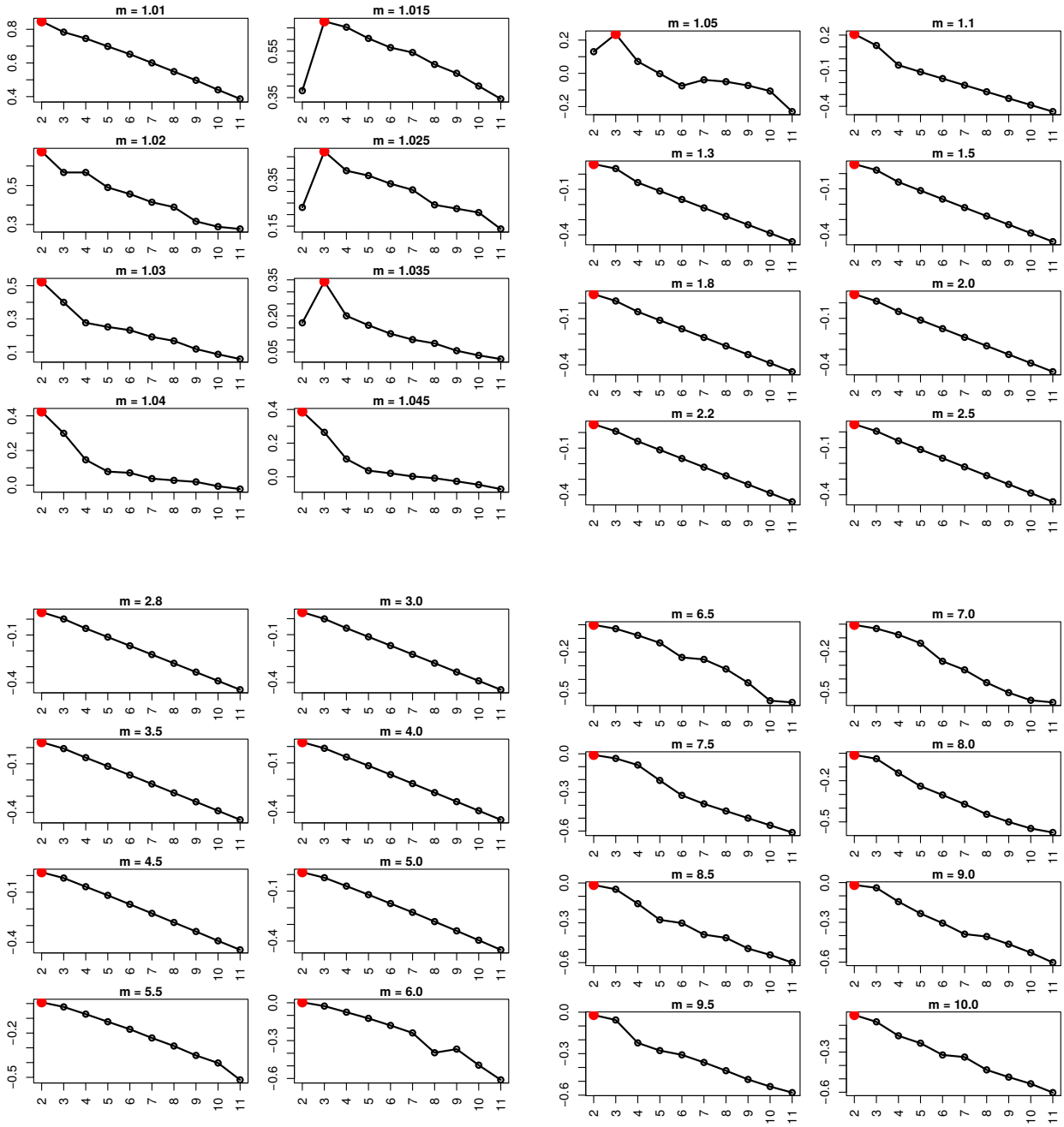


Tabela 78 – IAarticles

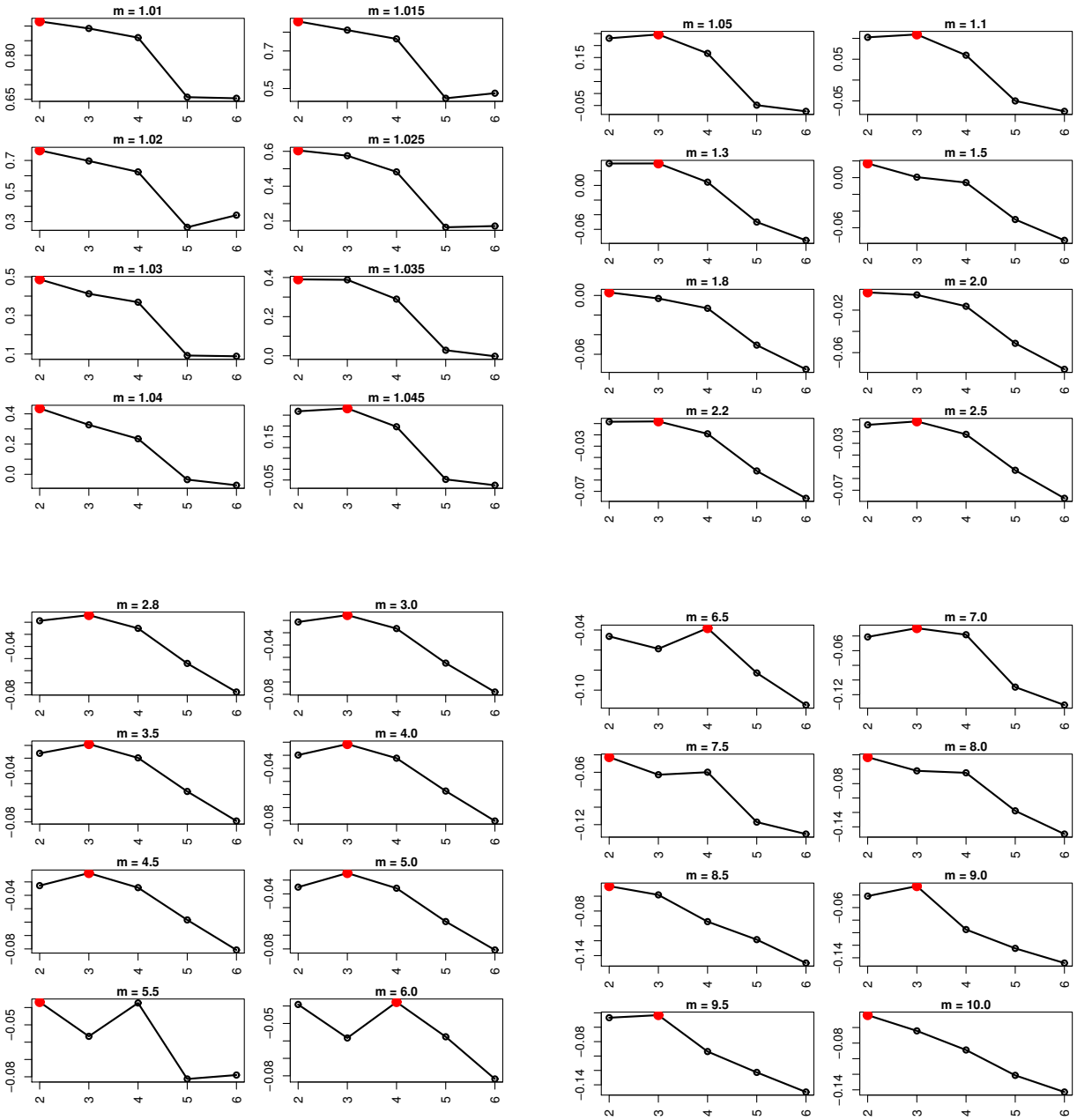


Tabela 79 – Opinosis

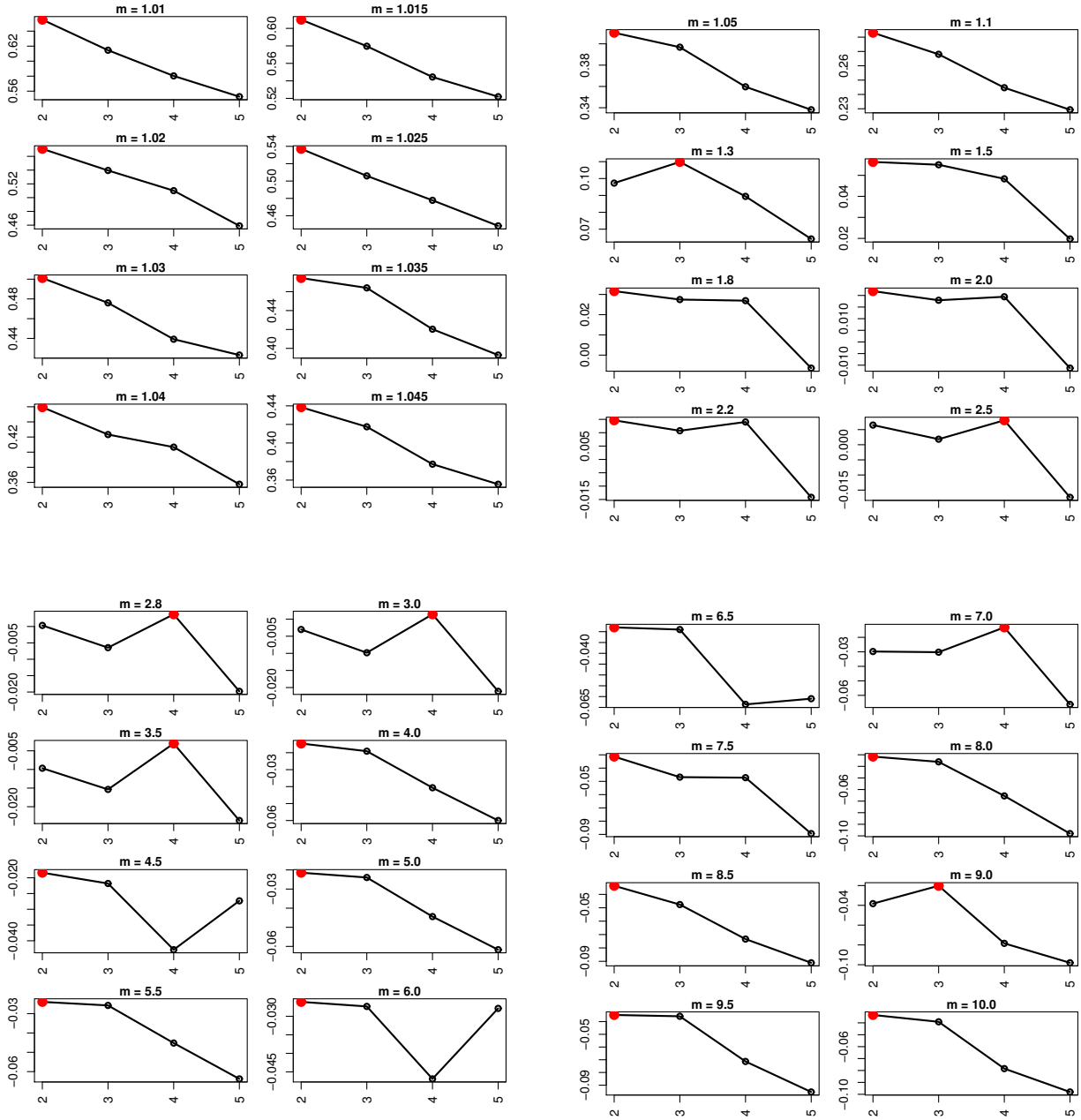
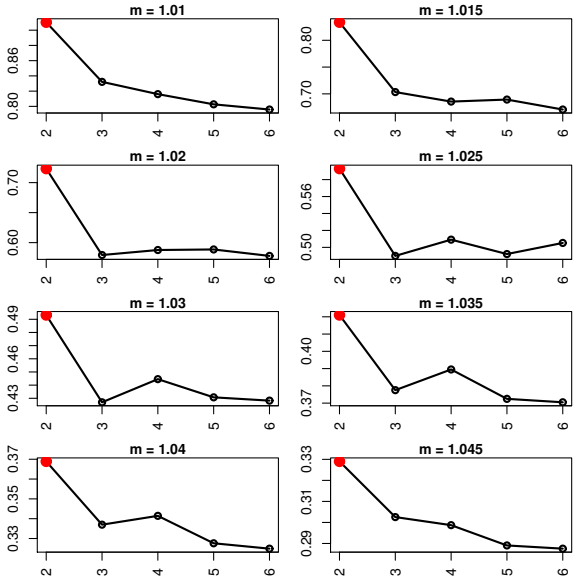
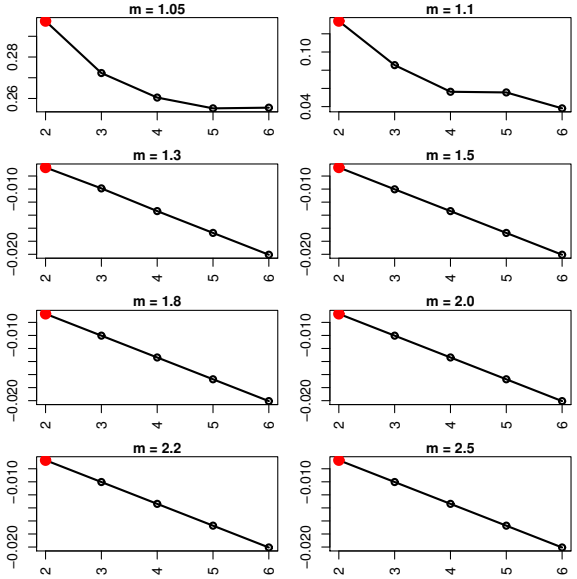
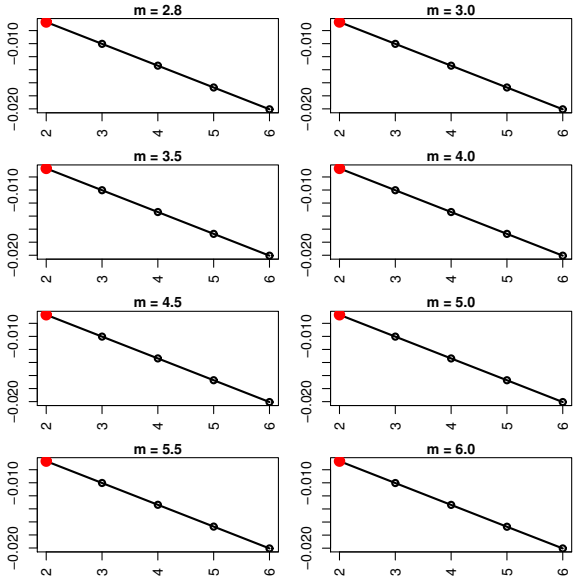


Tabela 80 – CSTR







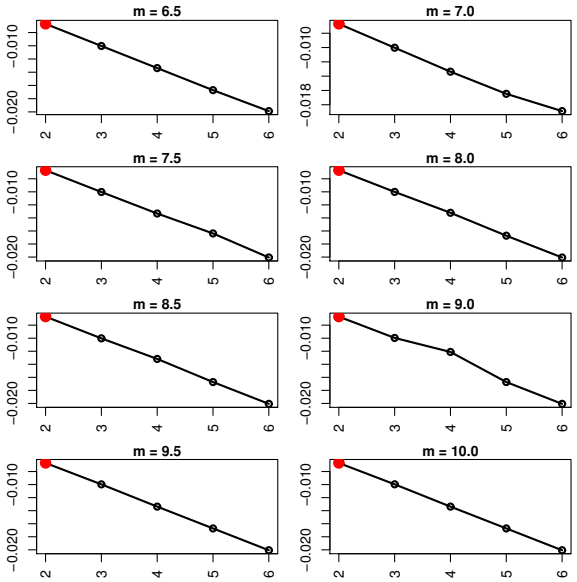


Tabela 81 – SyskillWebert

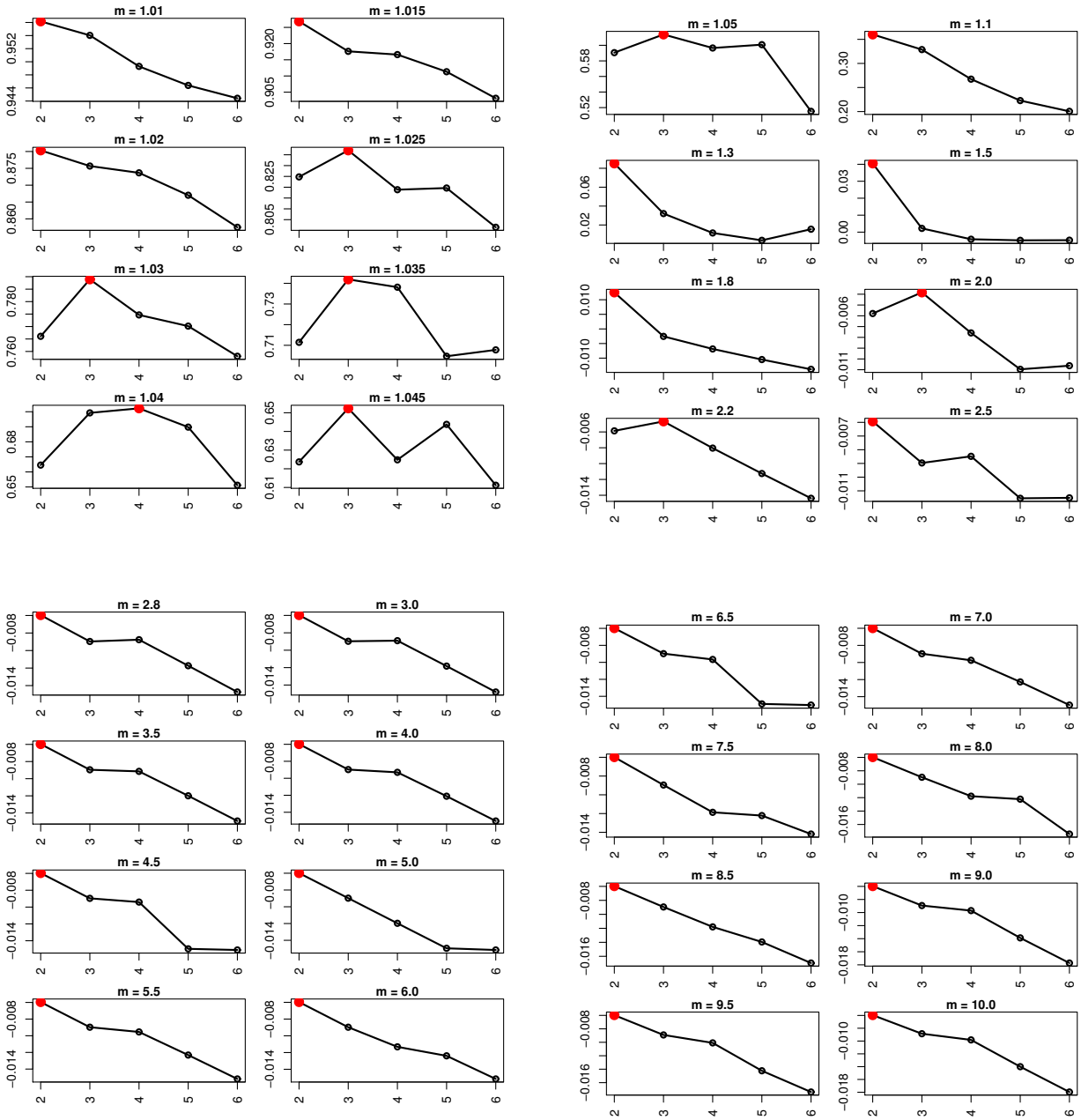


Tabela 82 – Hitech

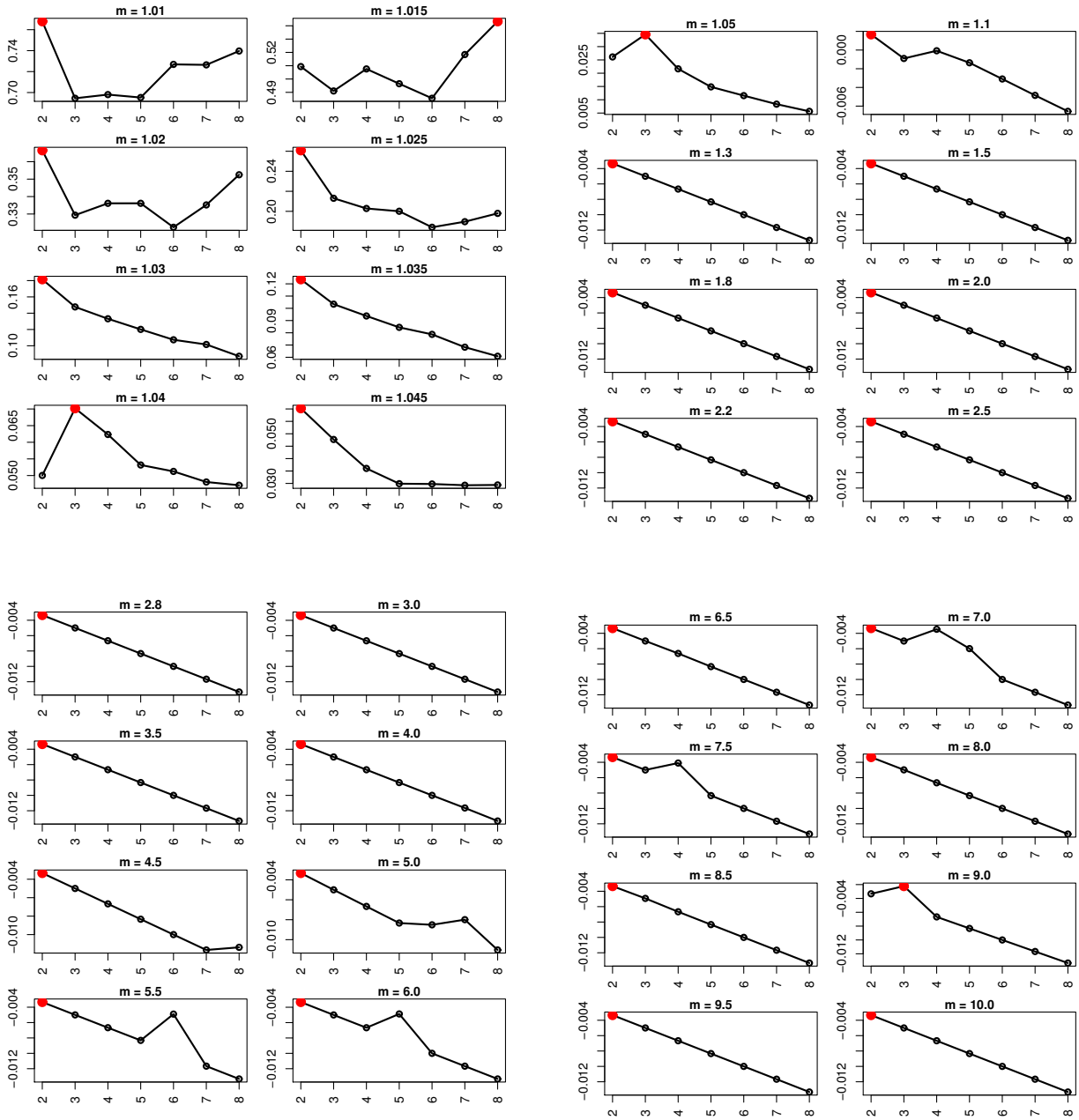


Tabela 83 – WAP

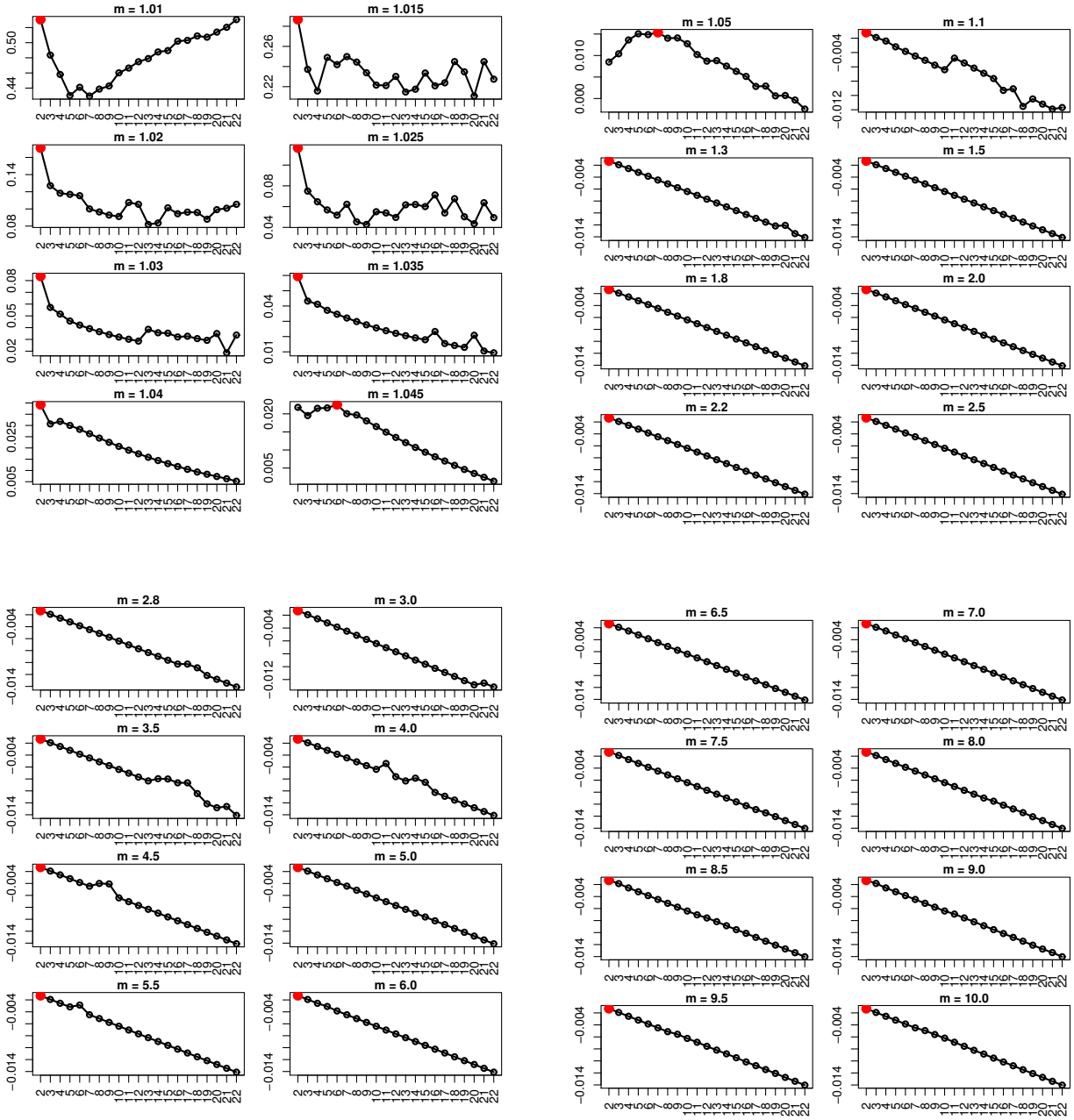


Tabela 84 – NSF

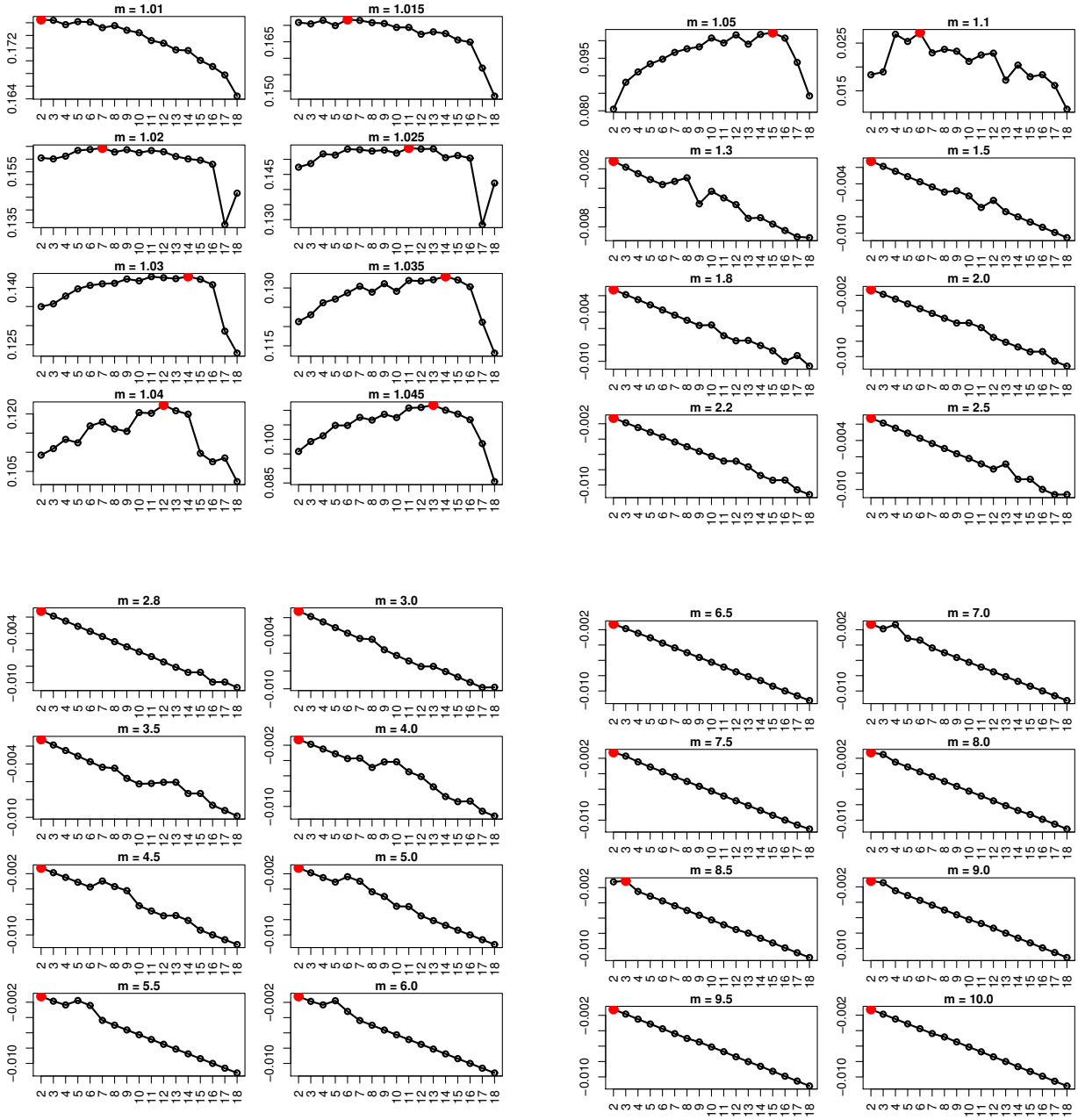


Tabela 85 – Irish-Sentiment

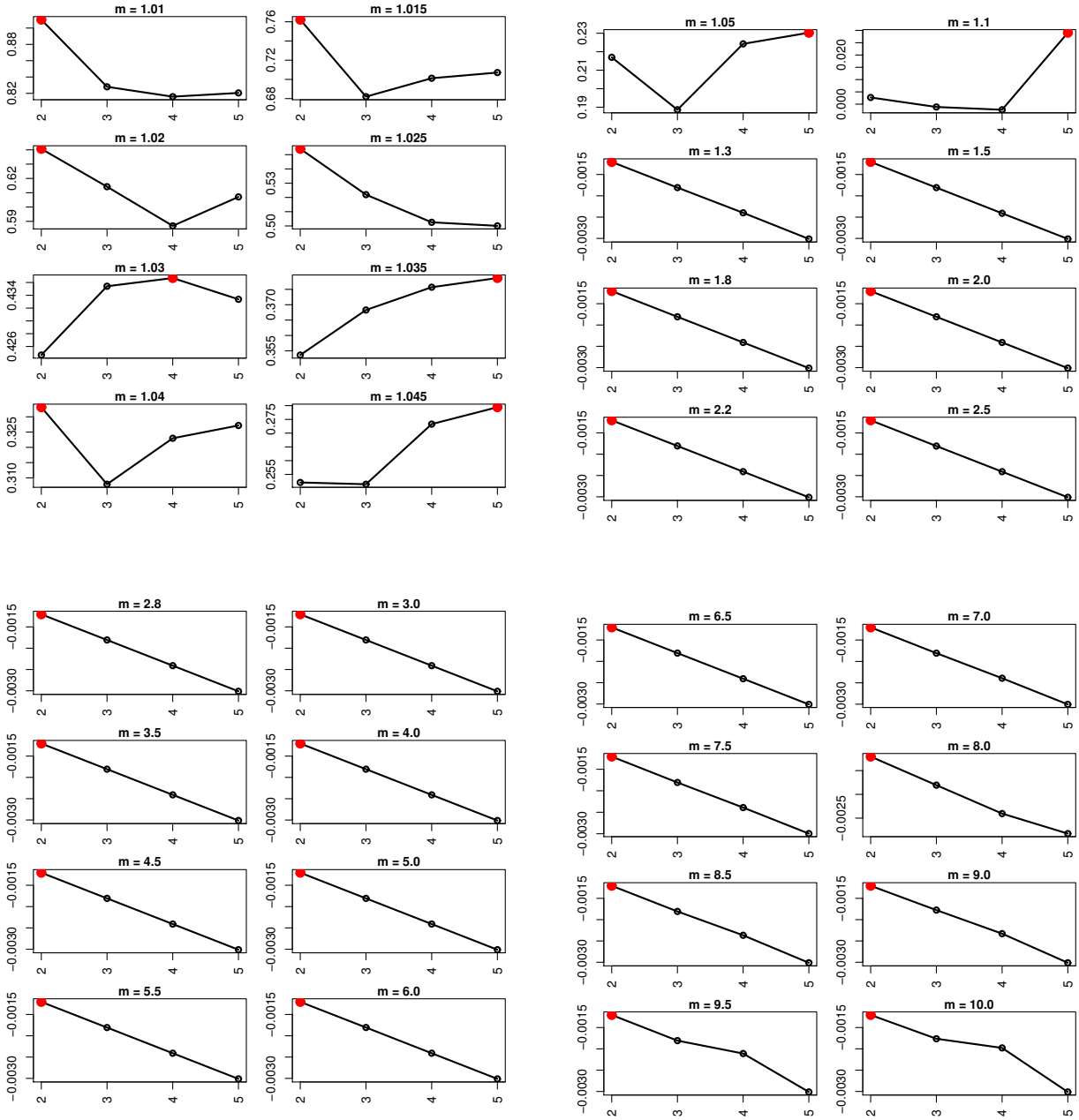


Tabela 86 – 20Newsgroups

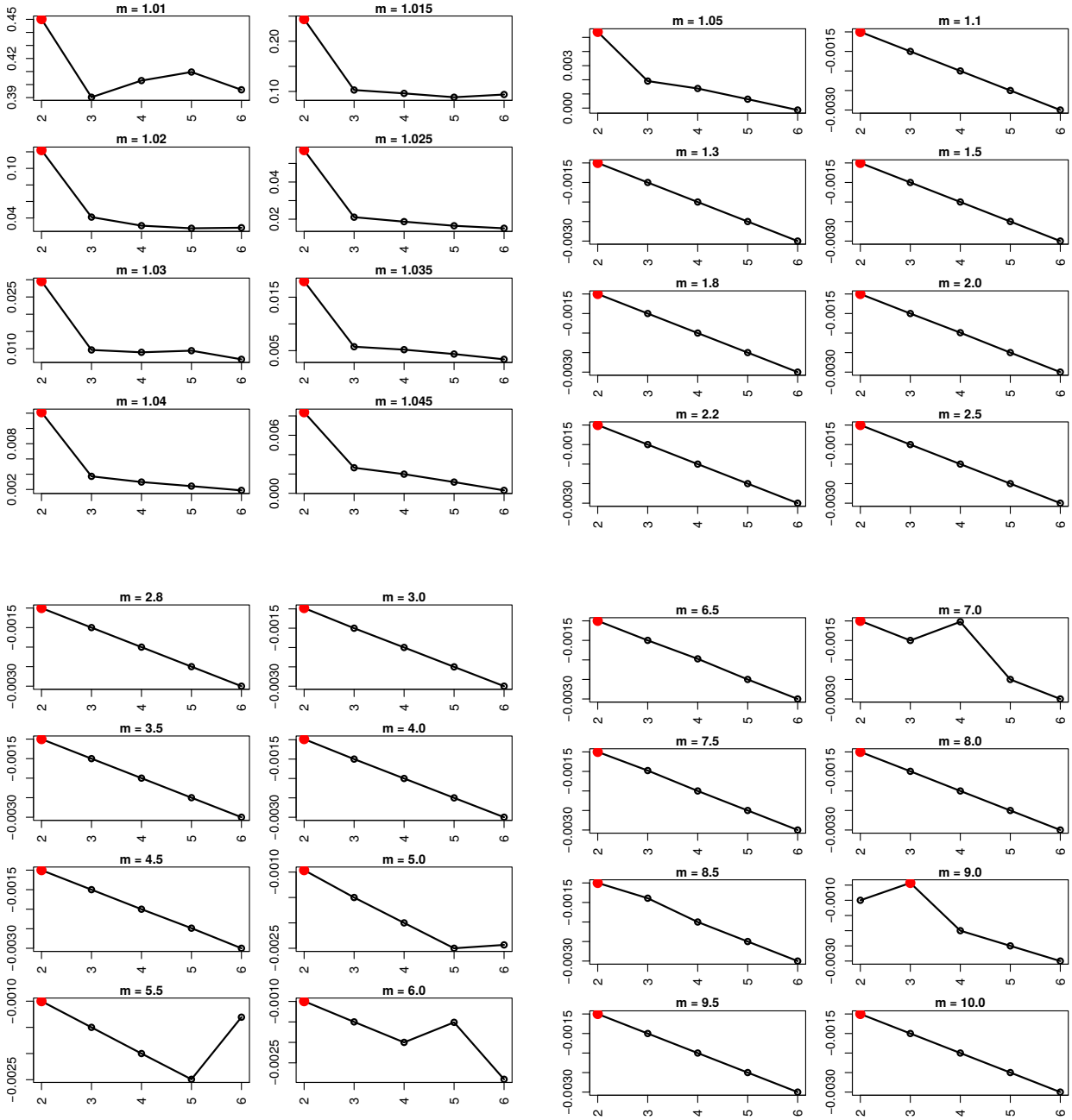


Tabela 87 – La1s

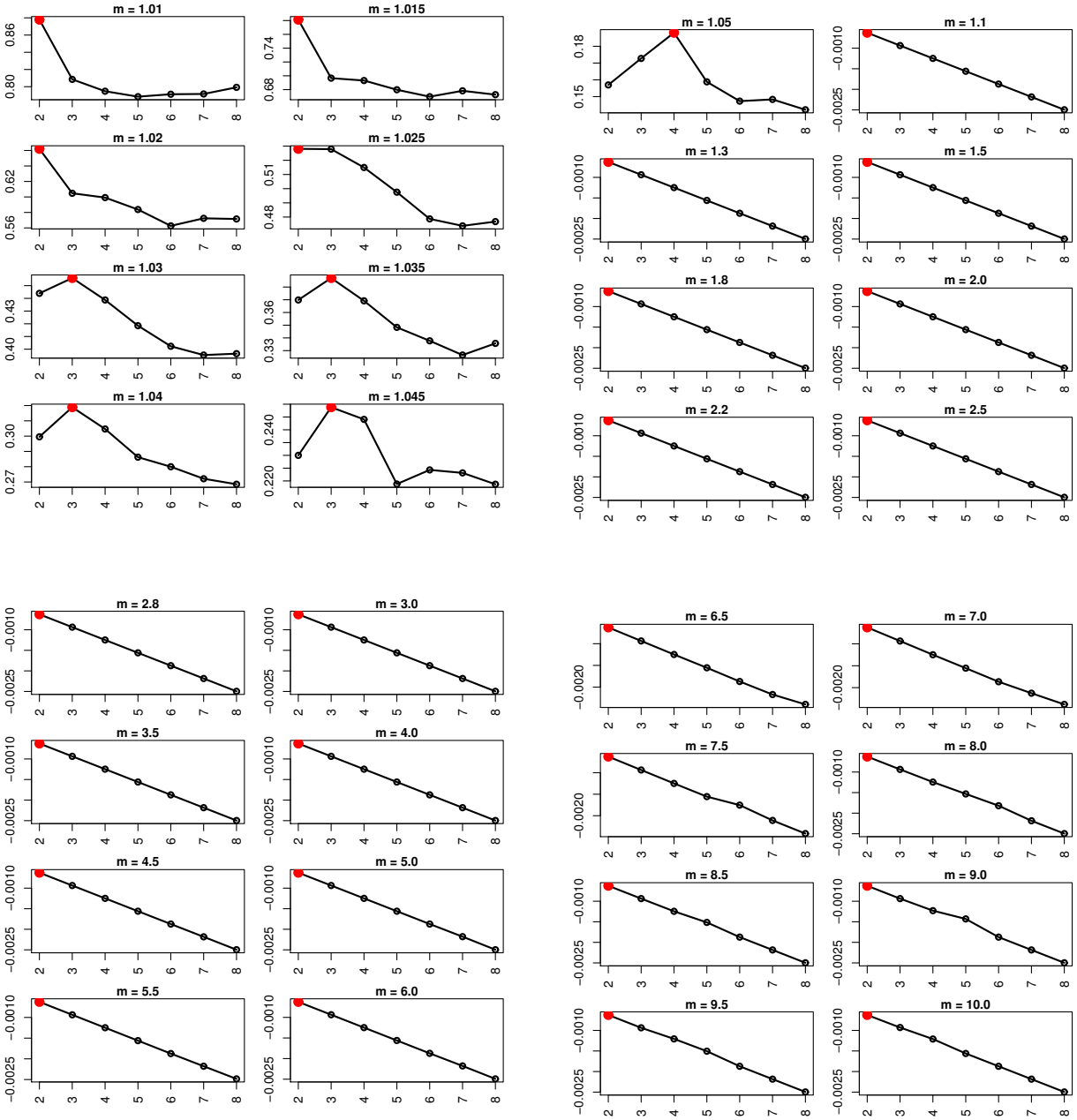
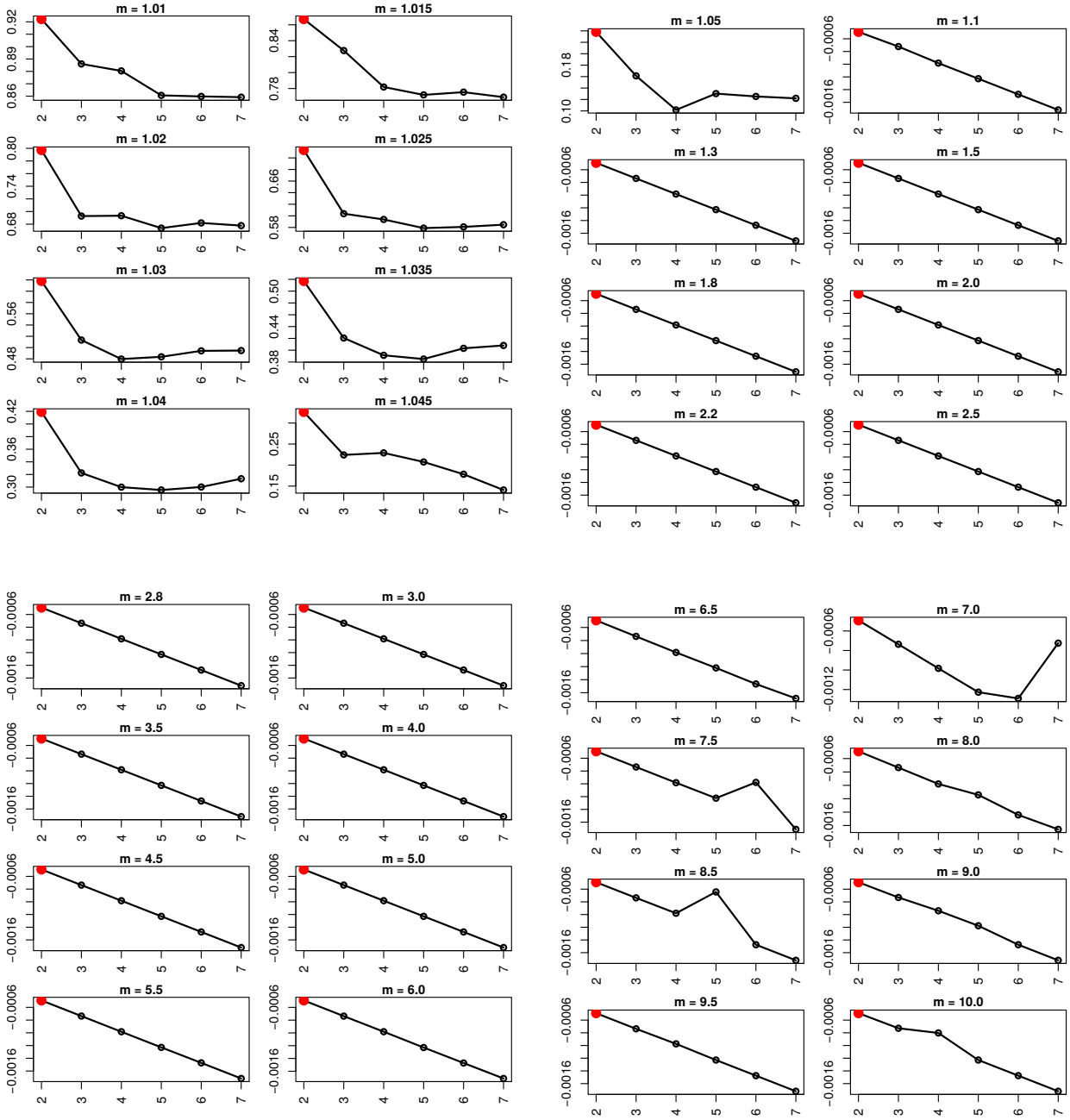


Tabela 88 – Reviews



ANEXO H – FS

Tabela 89 – NewYorkTimes

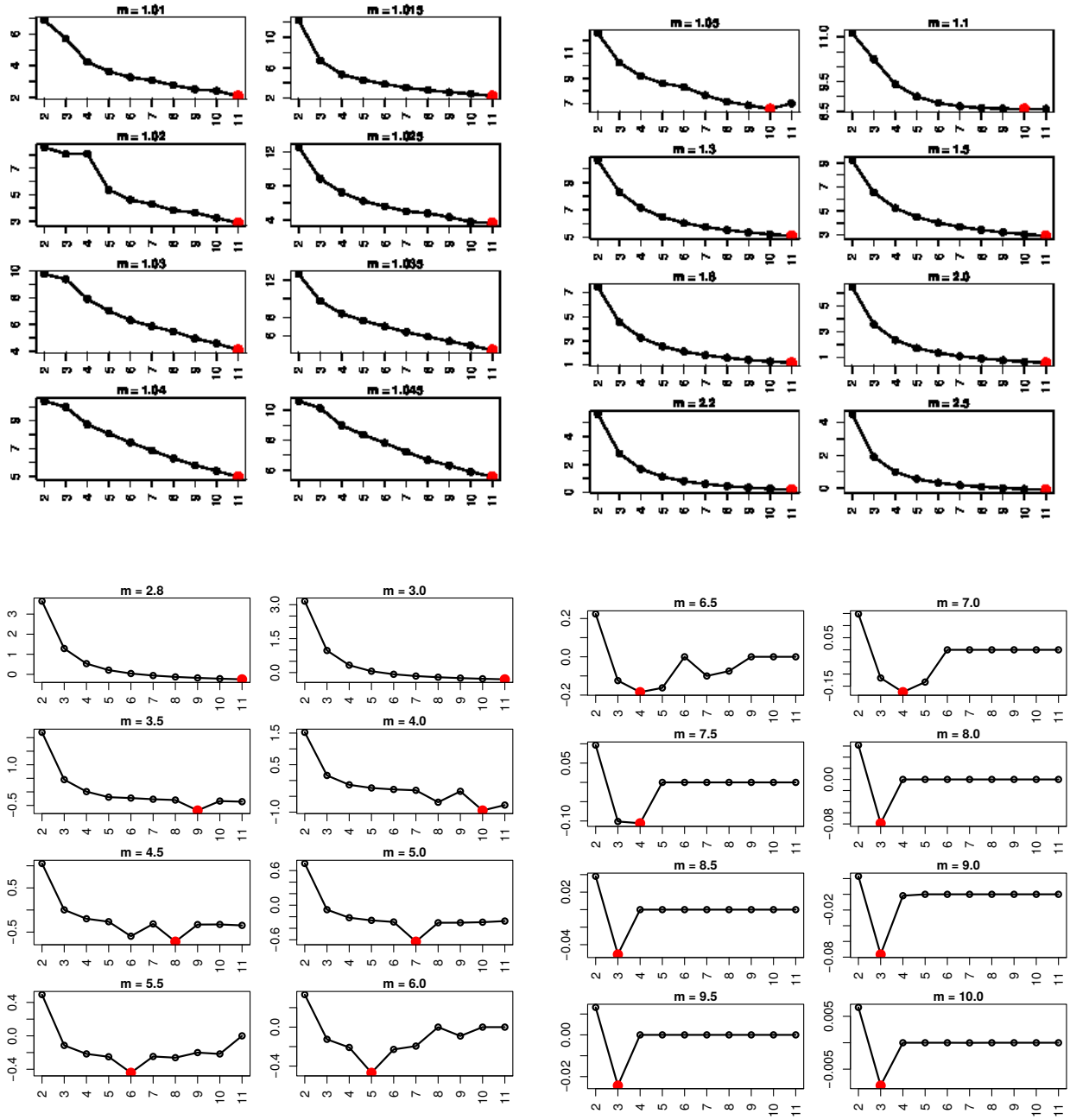


Tabela 90 – IAarticles

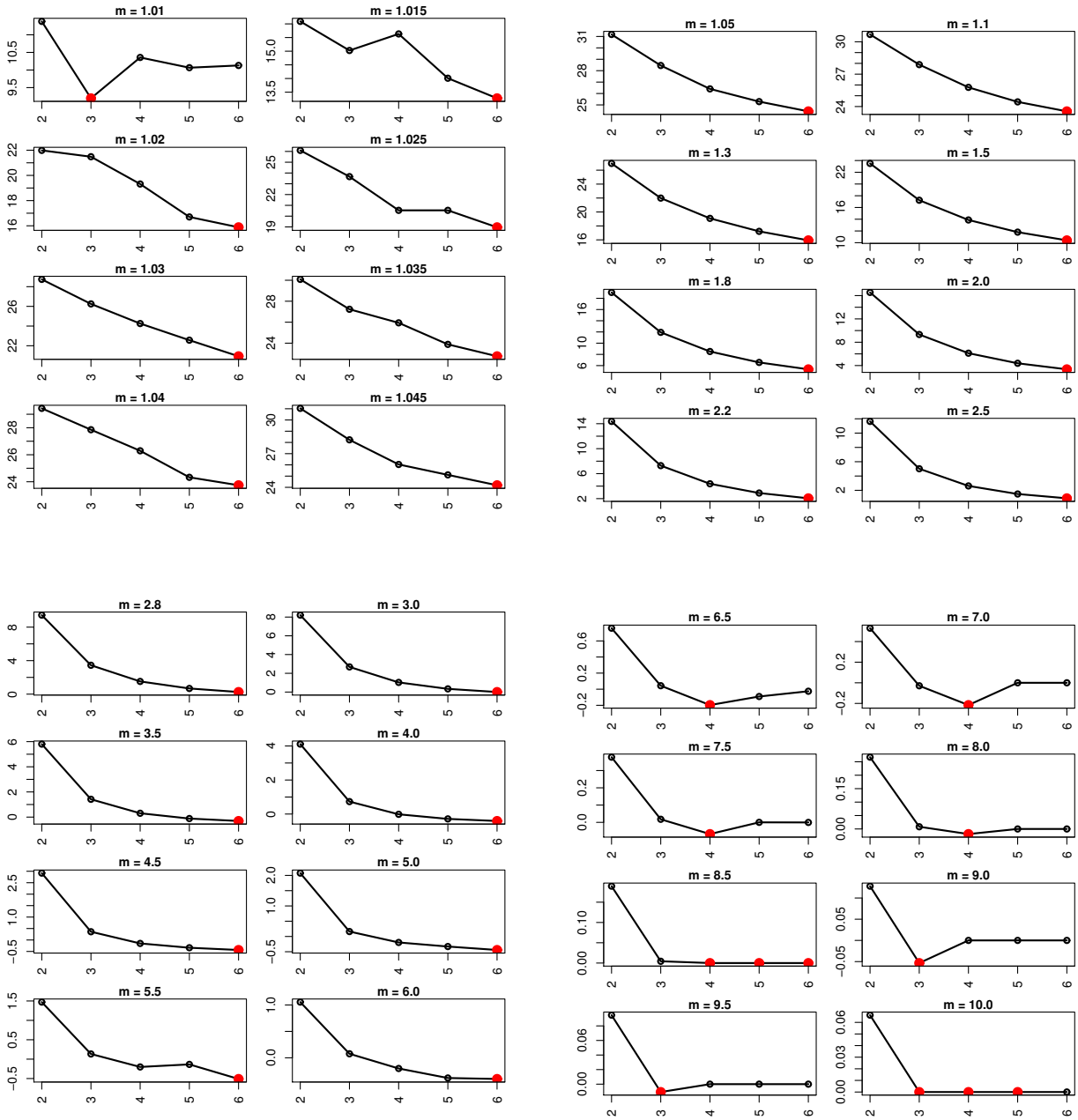


Tabela 91 – Opinosis

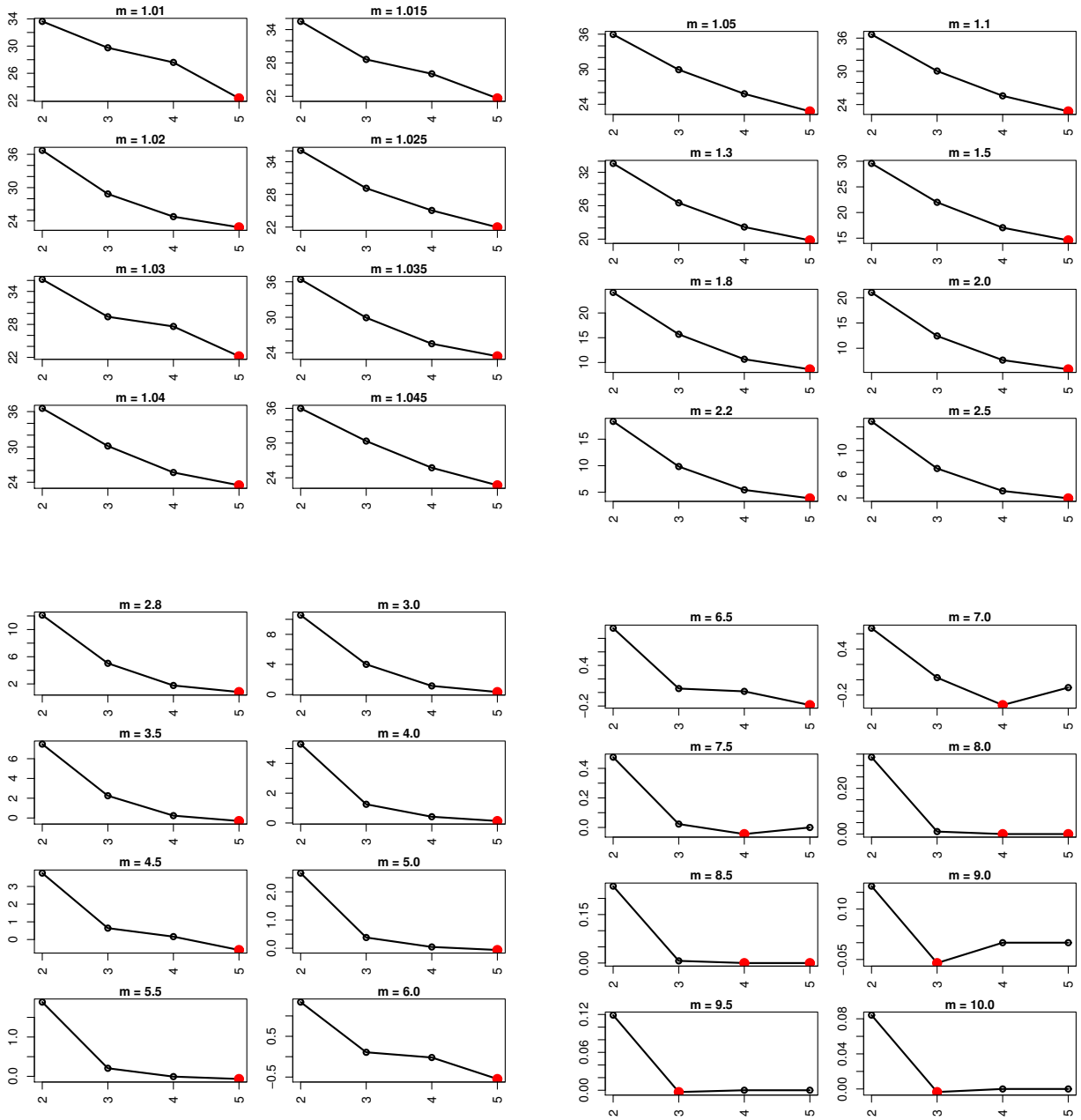


Tabela 92 – CSTR

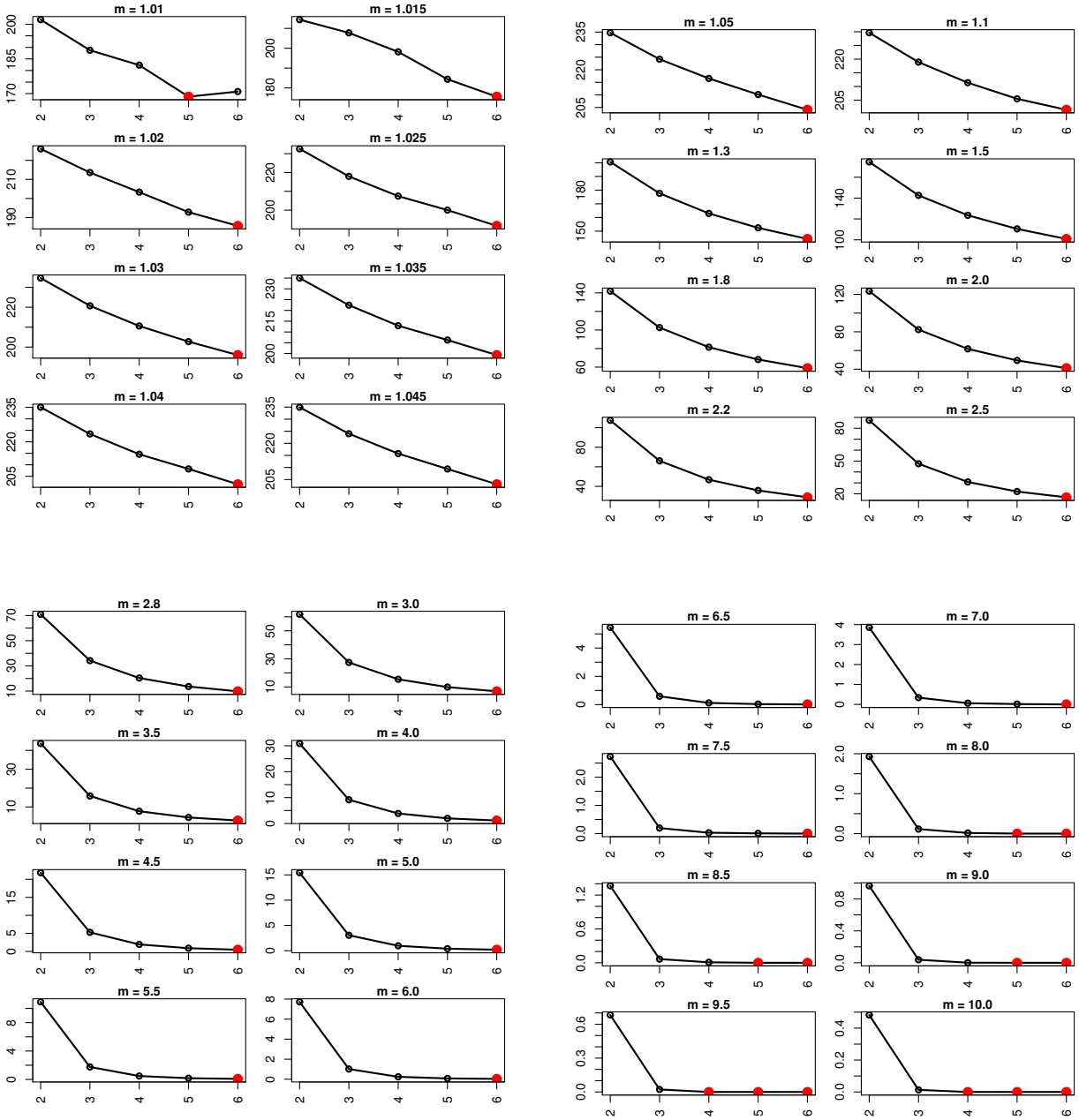


Tabela 93 – SyskillWebert

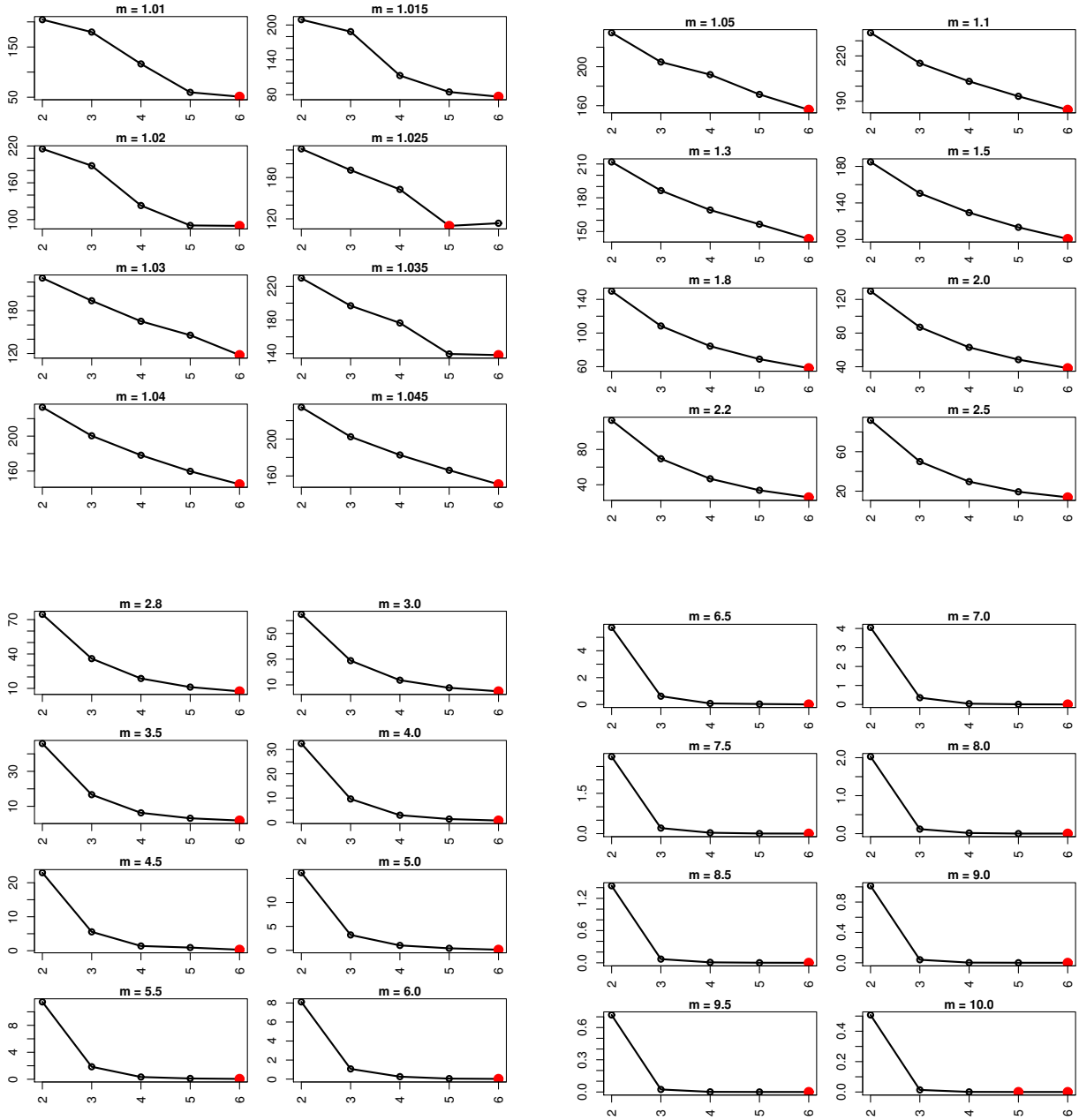


Tabela 94 – Hitech

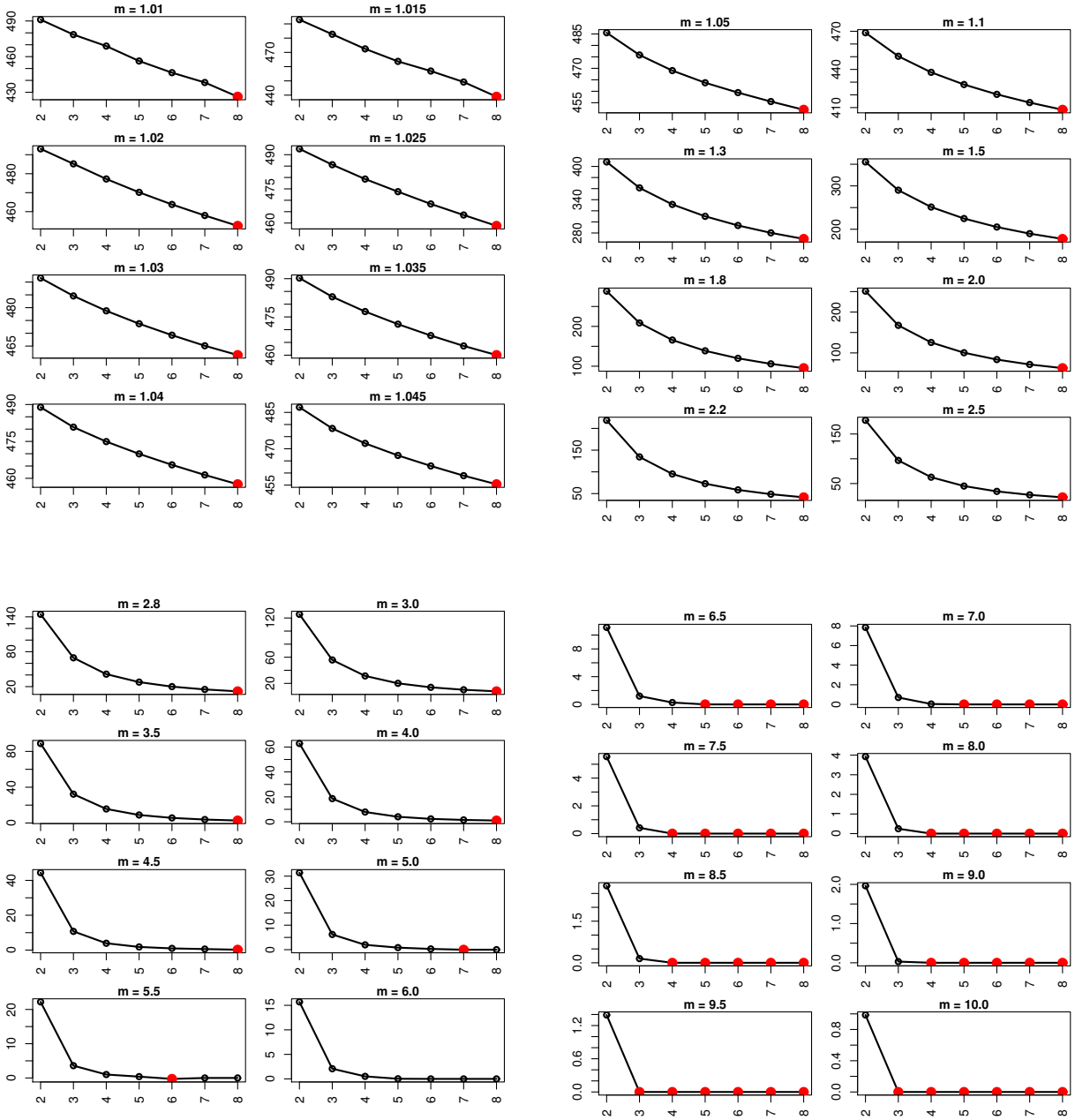


Tabela 95 – WAP

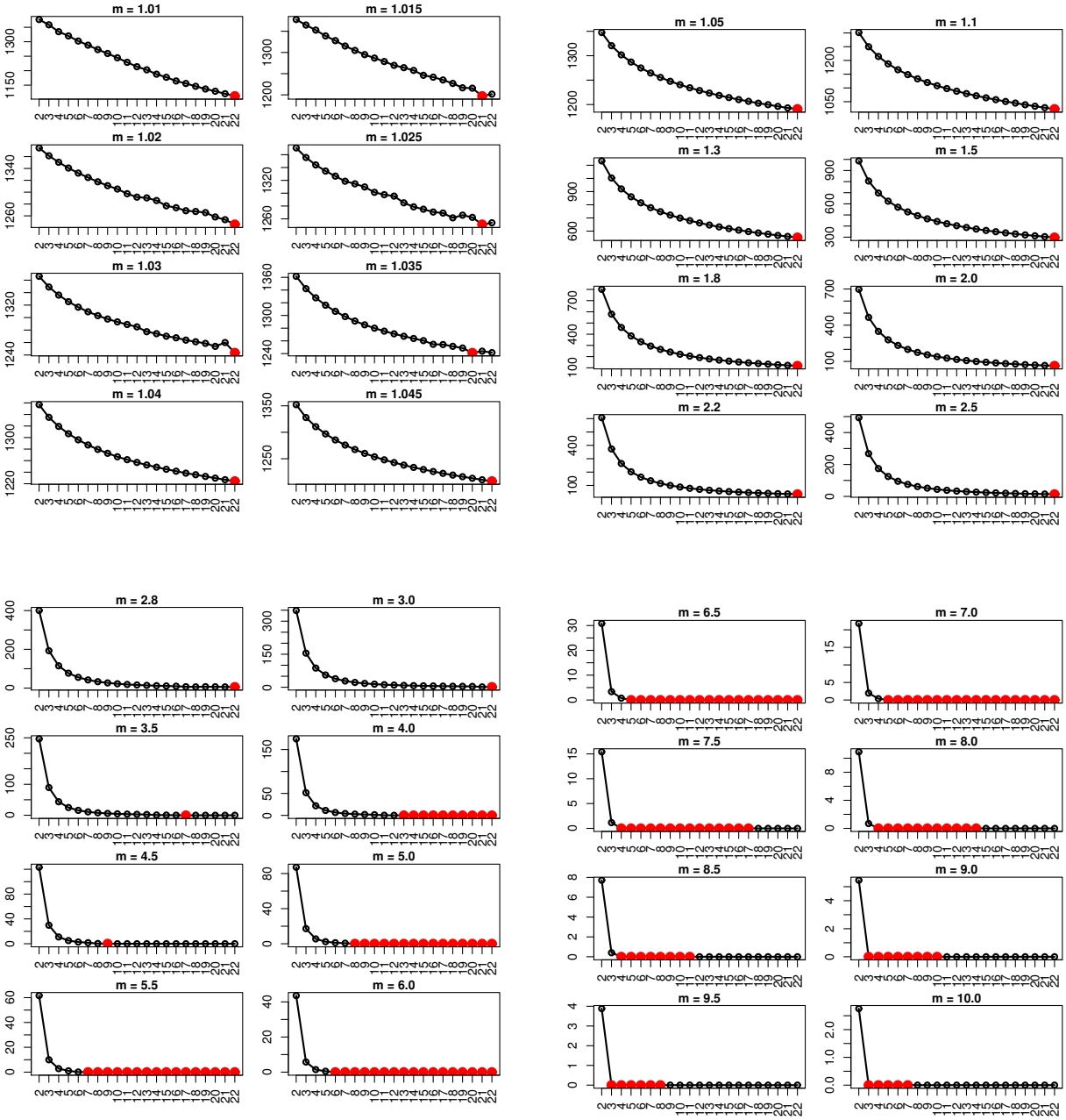


Tabela 96 – NSF

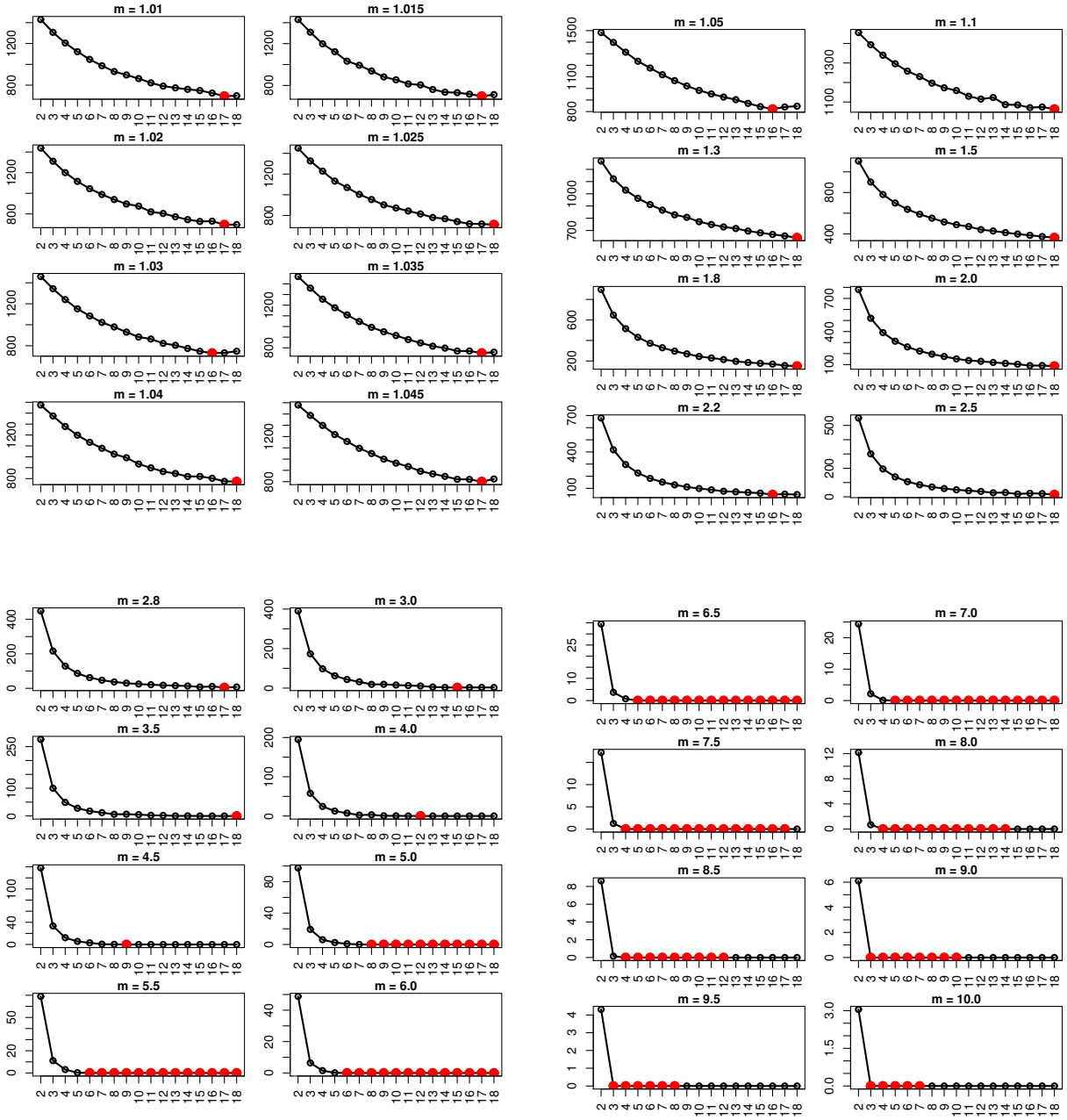


Tabela 97 – Irish-Sentiment

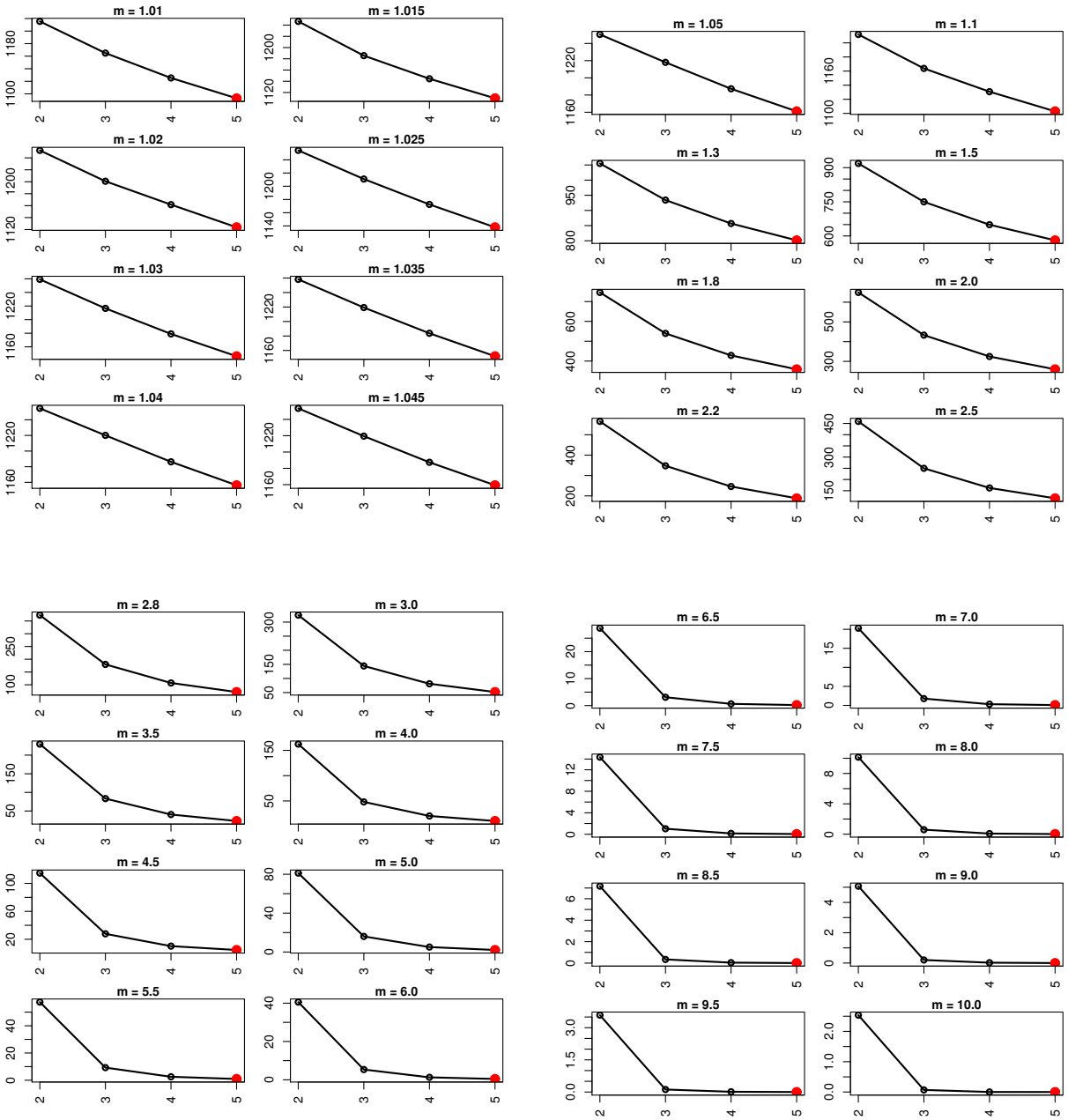


Tabela 98 – 20Newsgroups

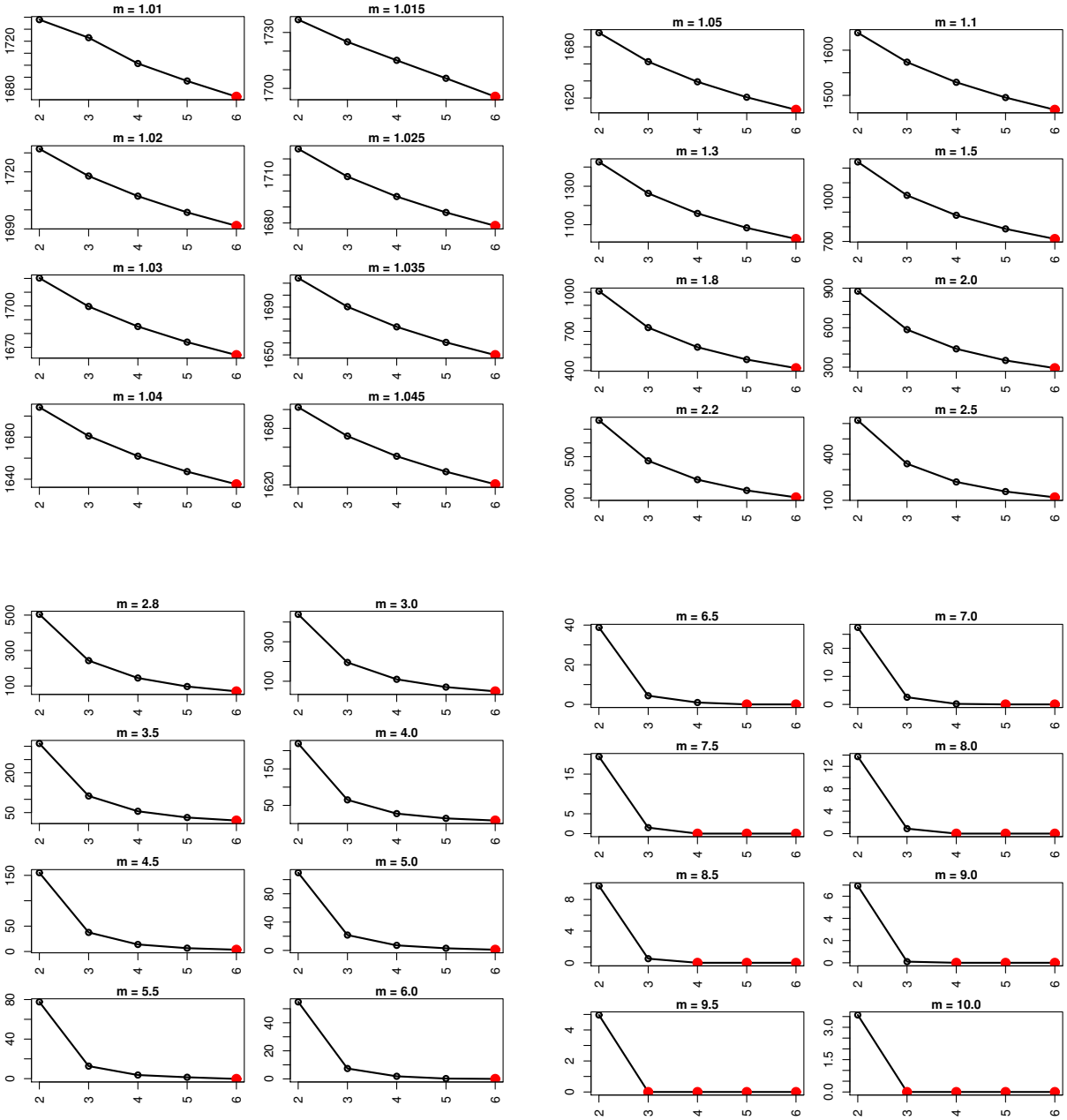


Tabela 99 – La1s

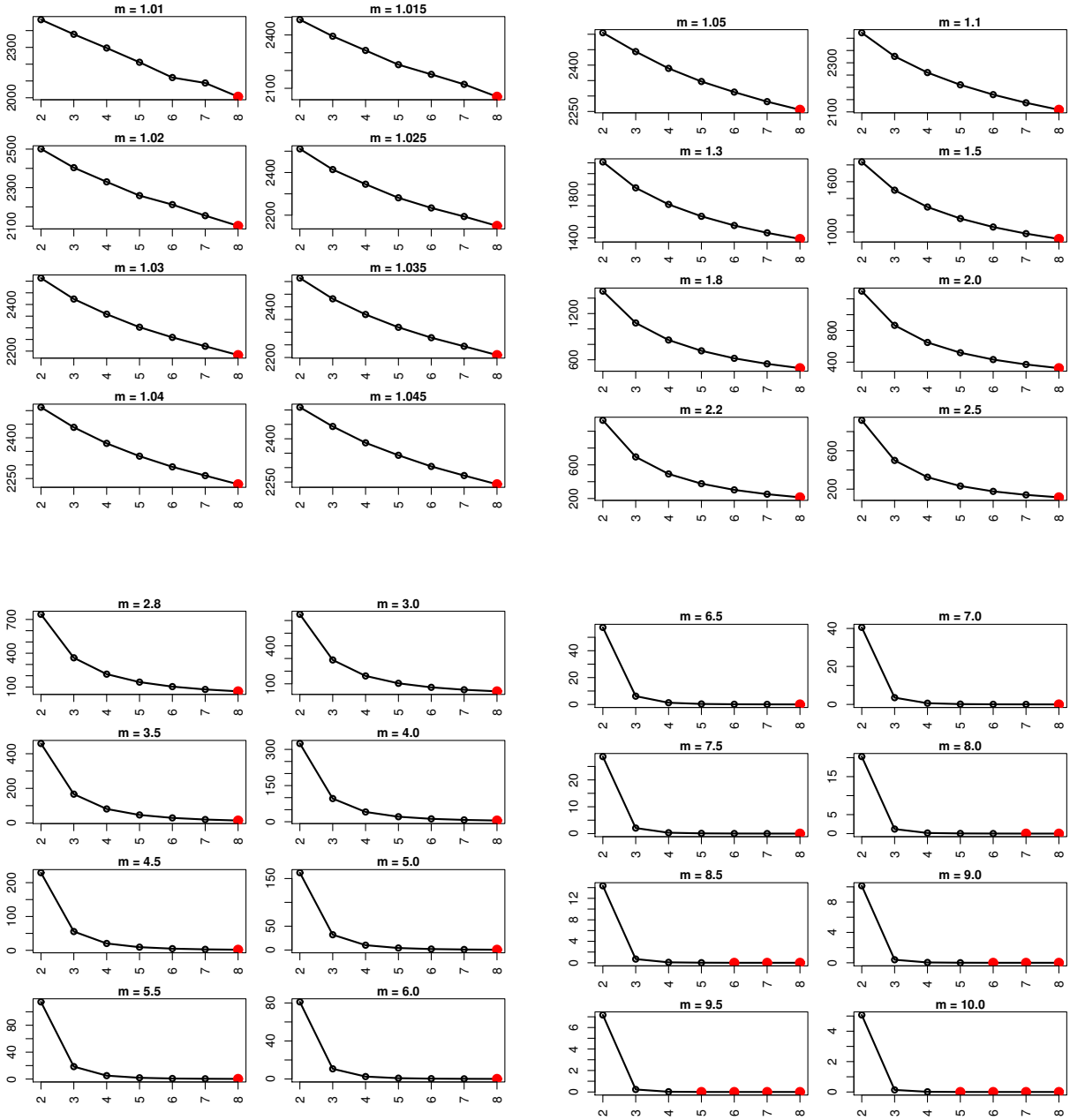
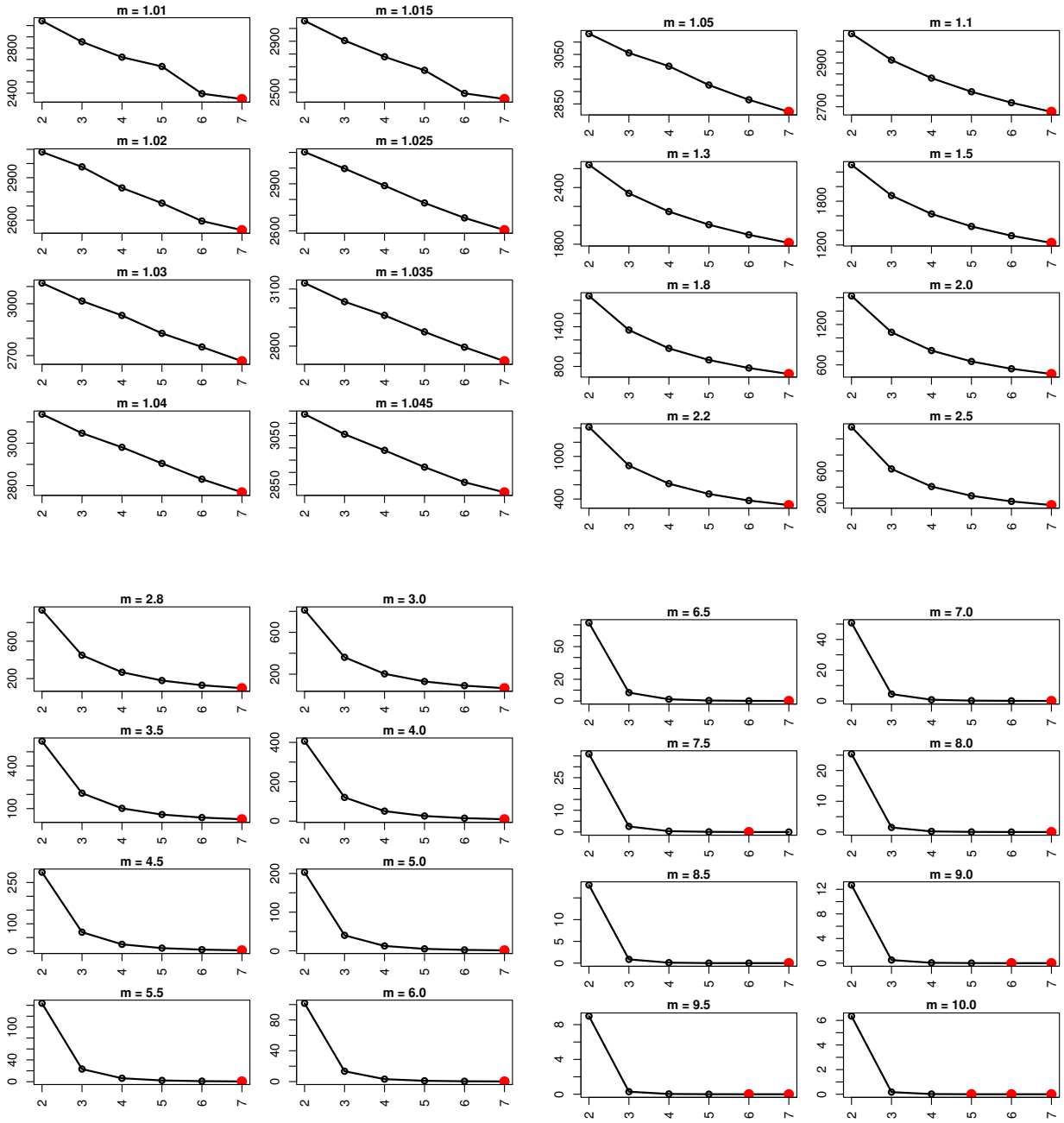


Tabela 100 – Reviews



ANEXO I – FHV

Tabela 101 – NewYorkTimes

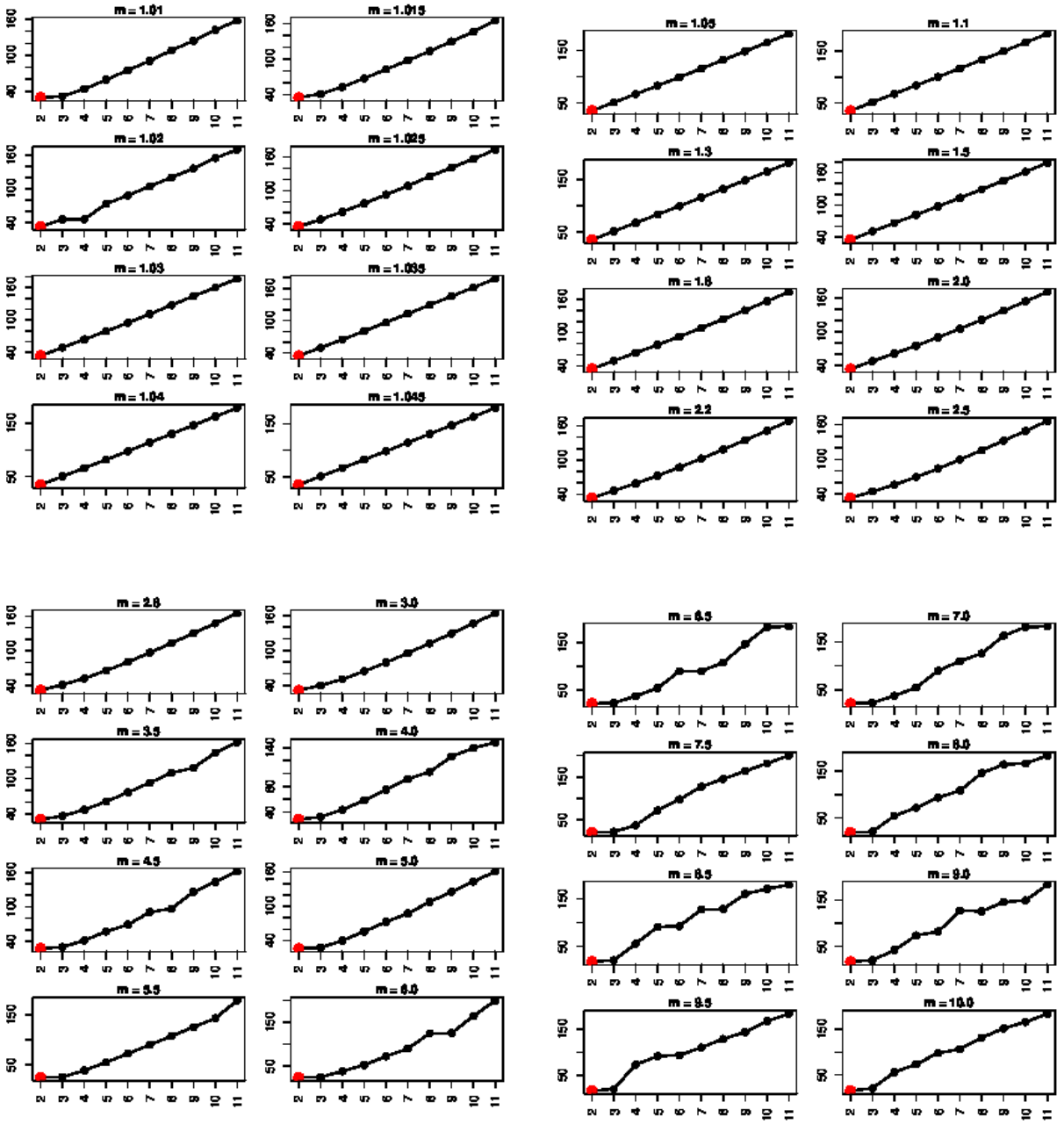


Tabela 102 – IAarticles

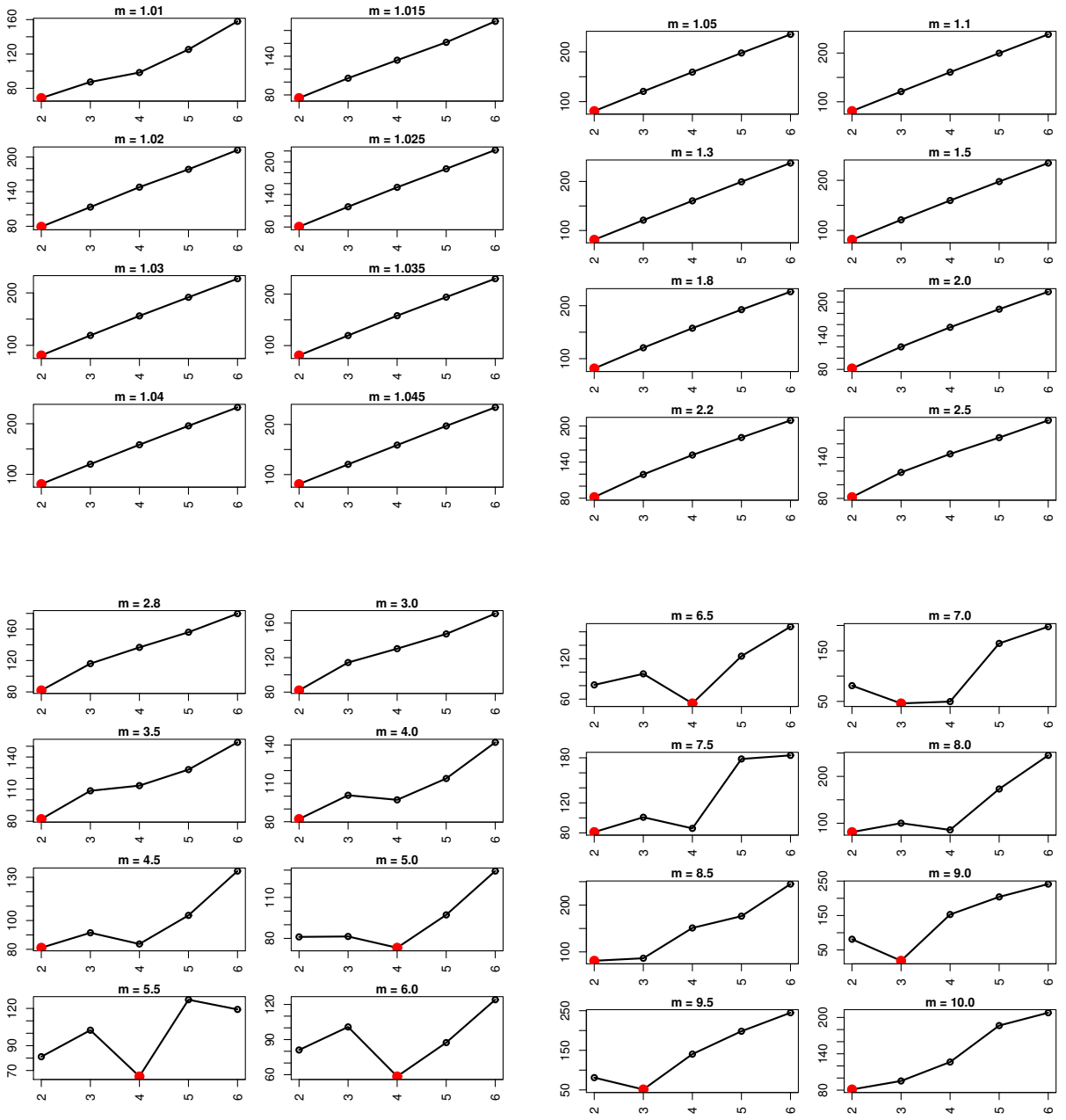


Tabela 103 – Opinosis

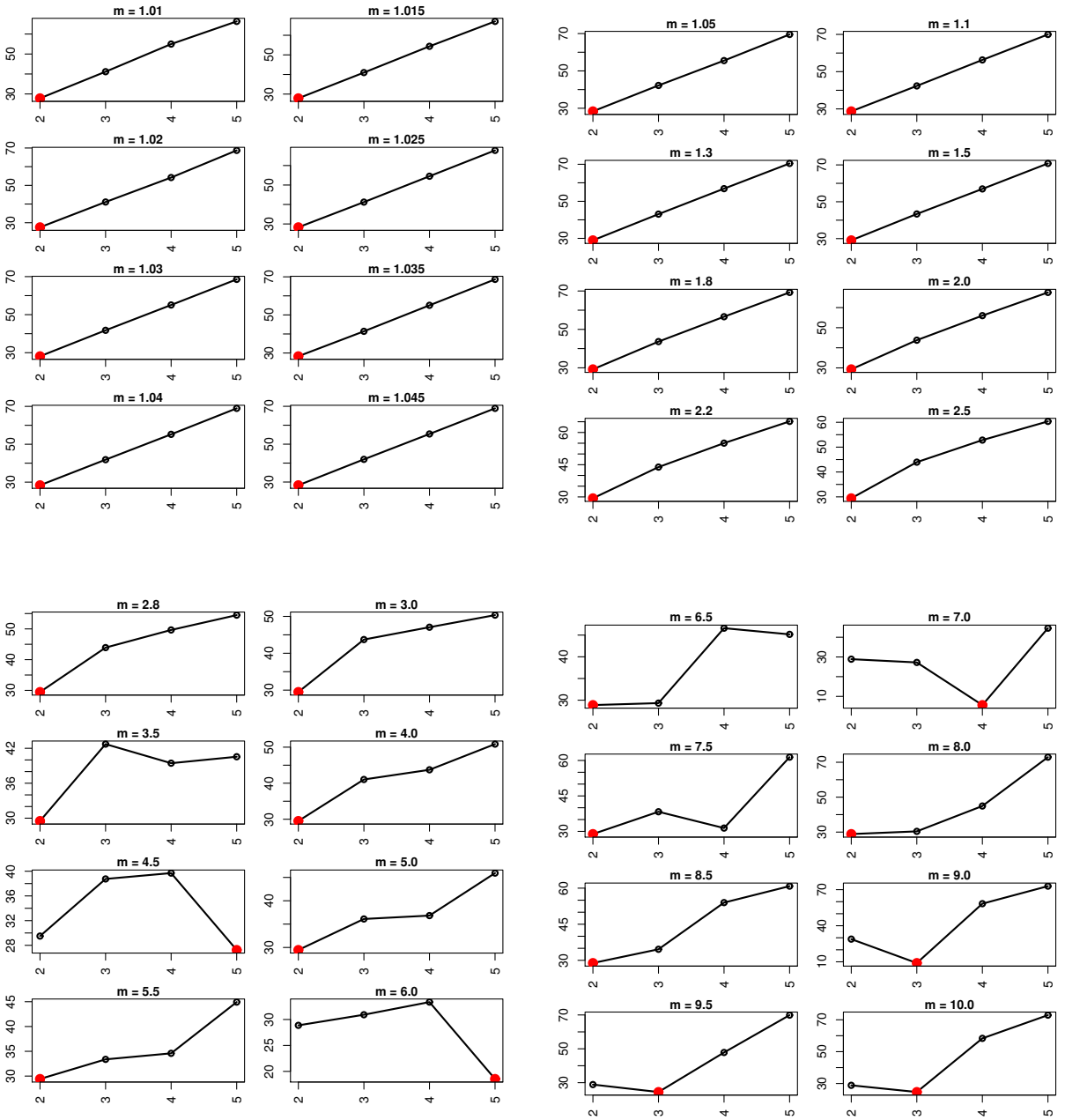
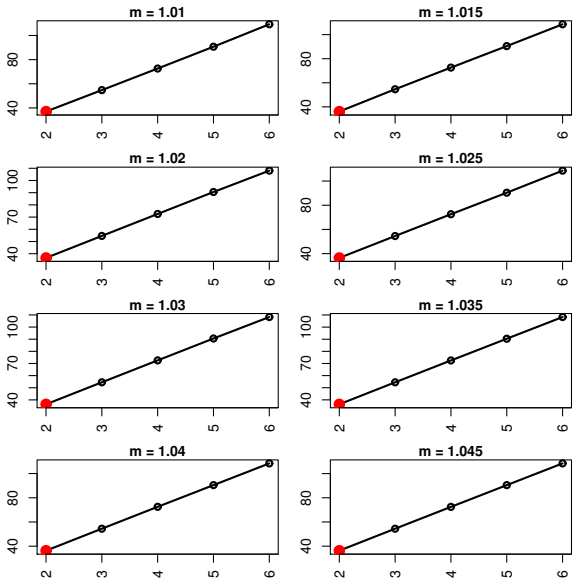
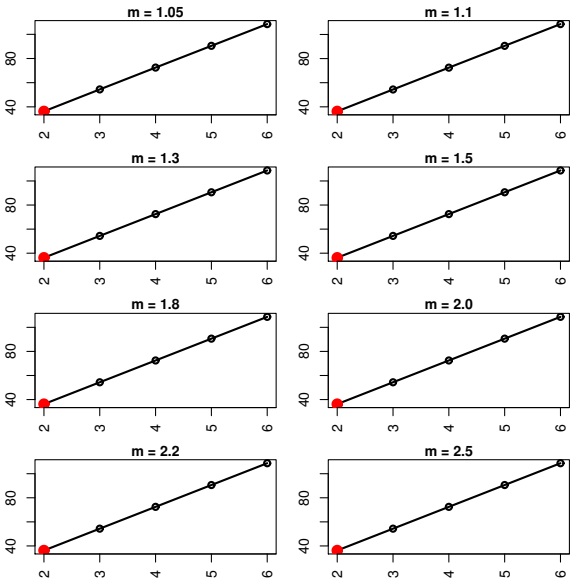
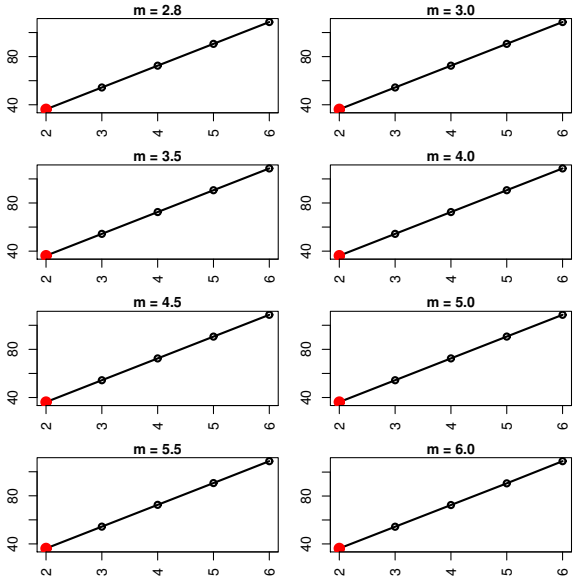


Tabela 104 – CSTR







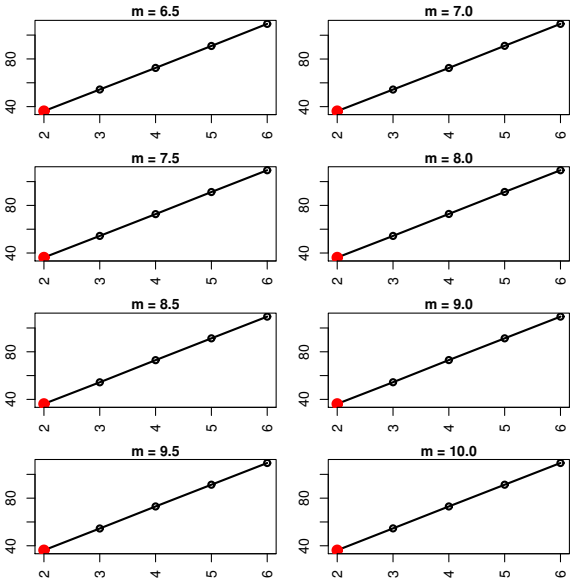


Tabela 105 – SyskillWebert

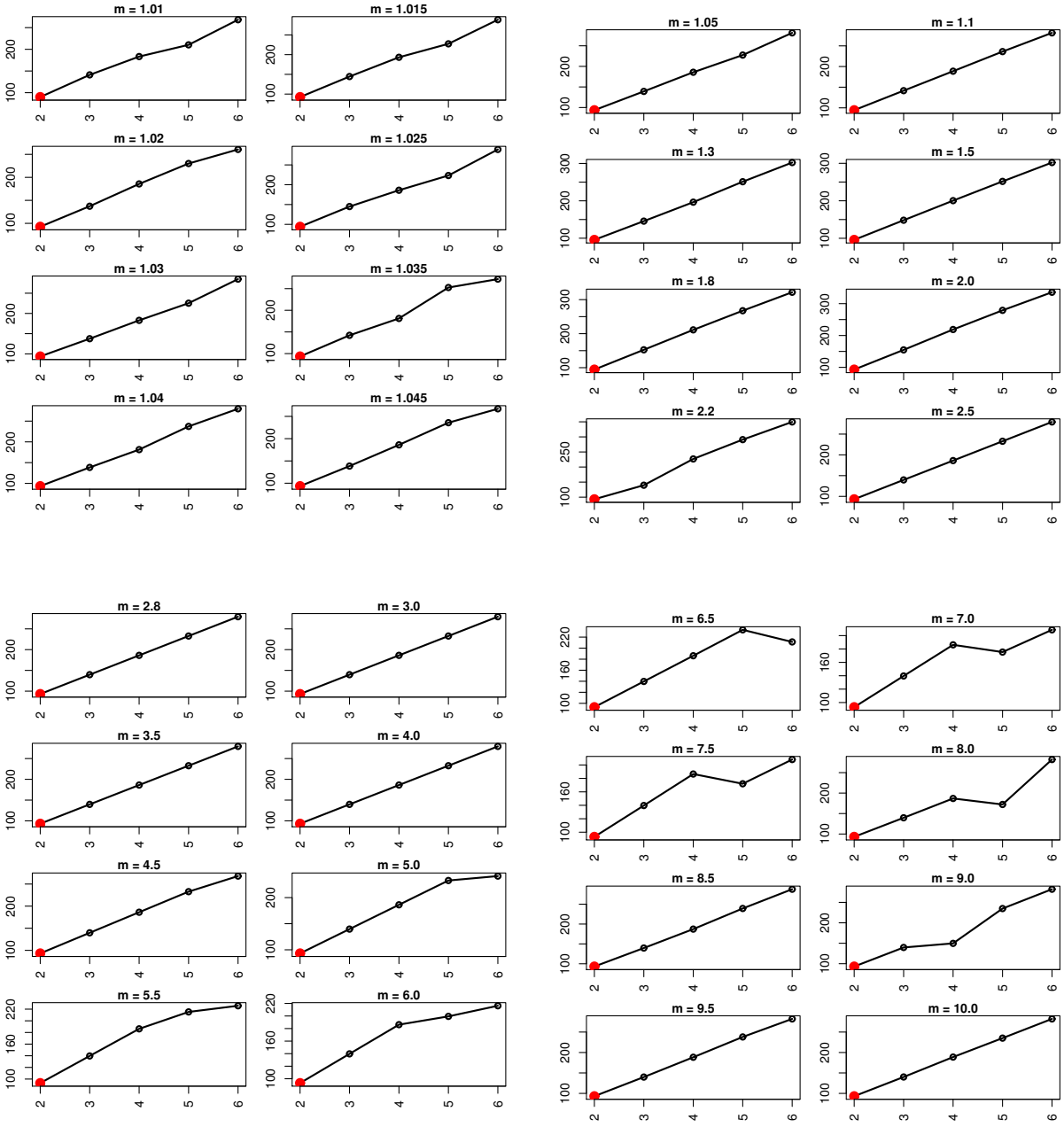


Tabela 106 – Hitech

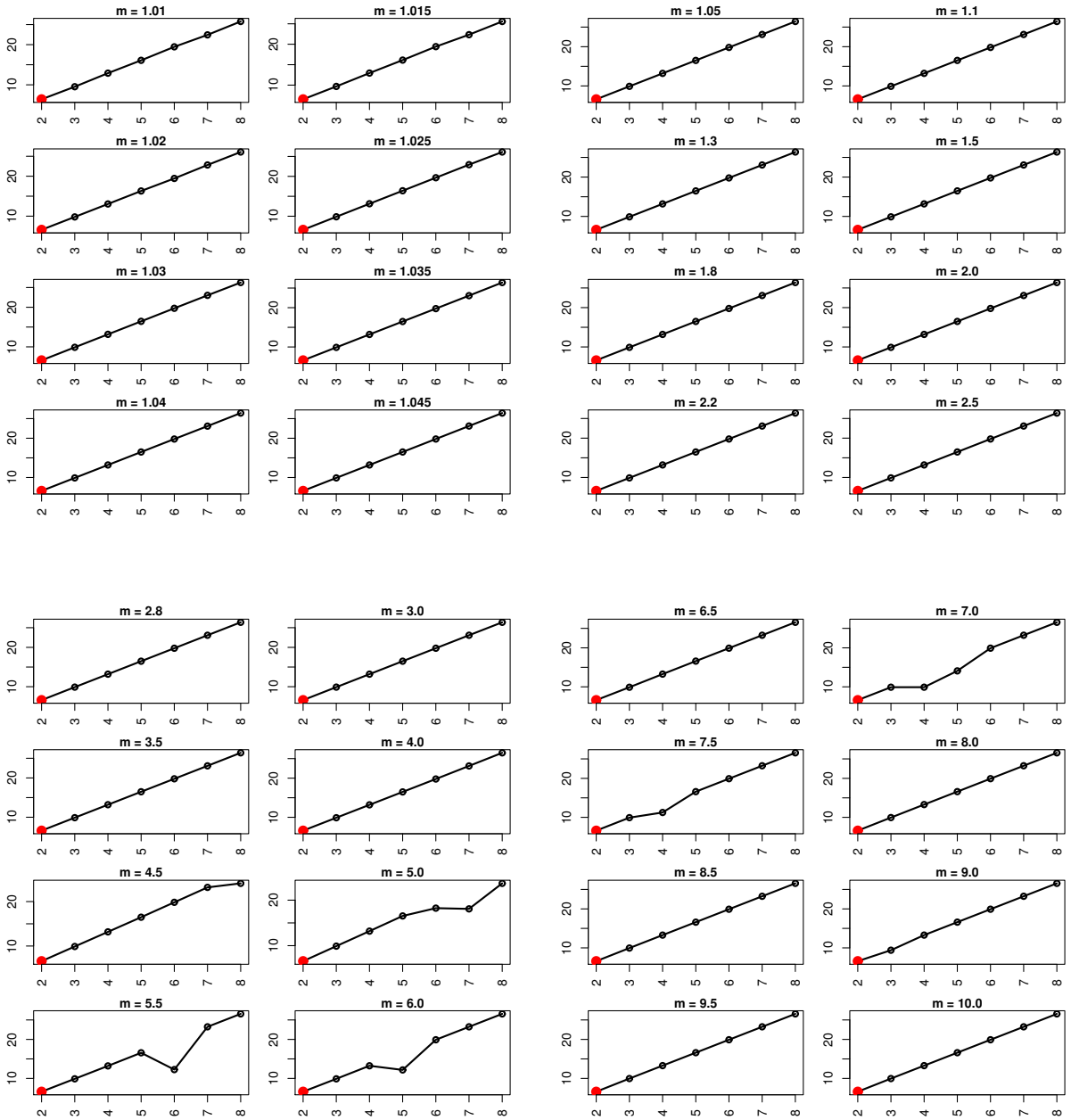


Tabela 107 – WAP

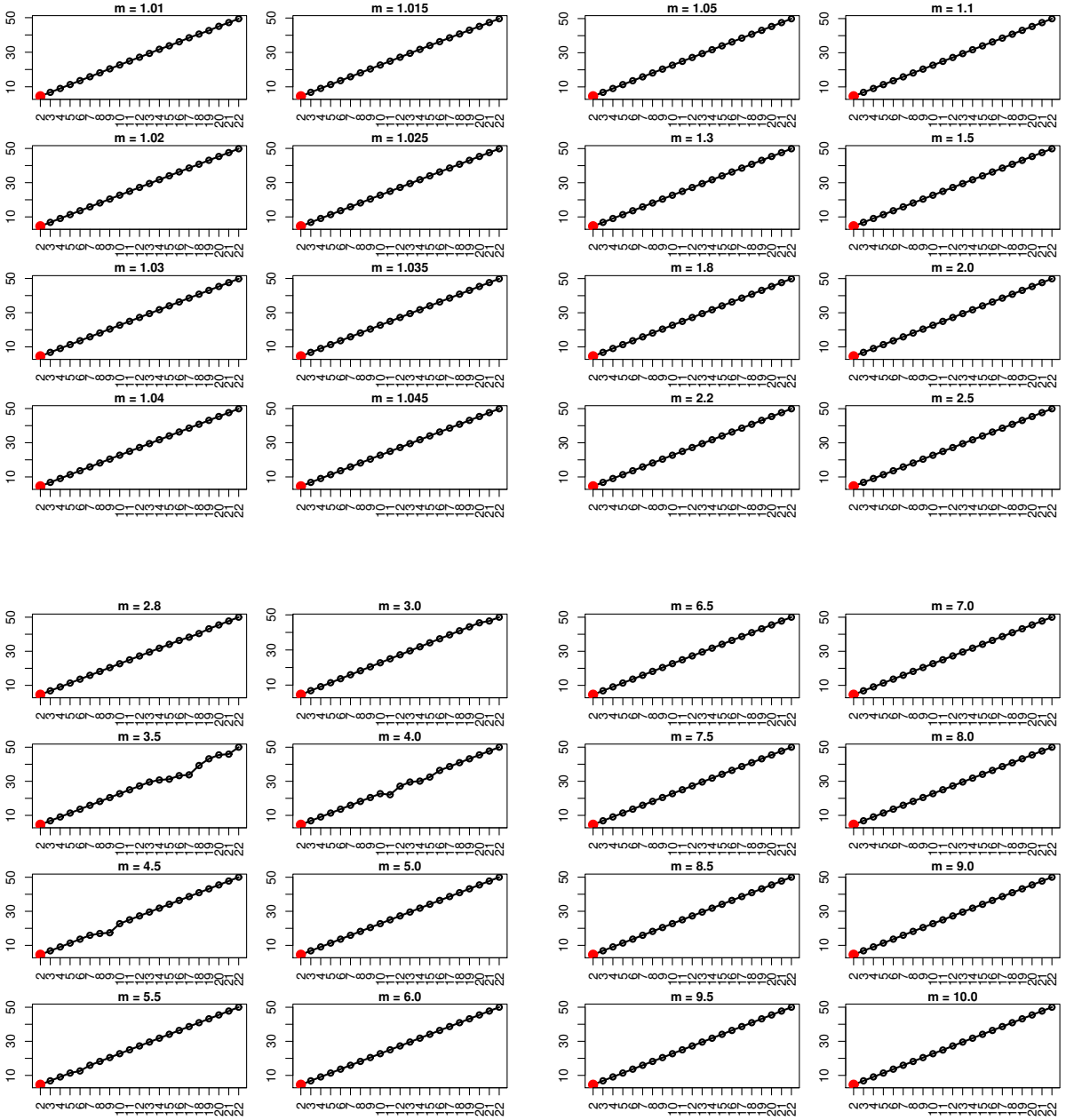


Tabela 108 – NSF

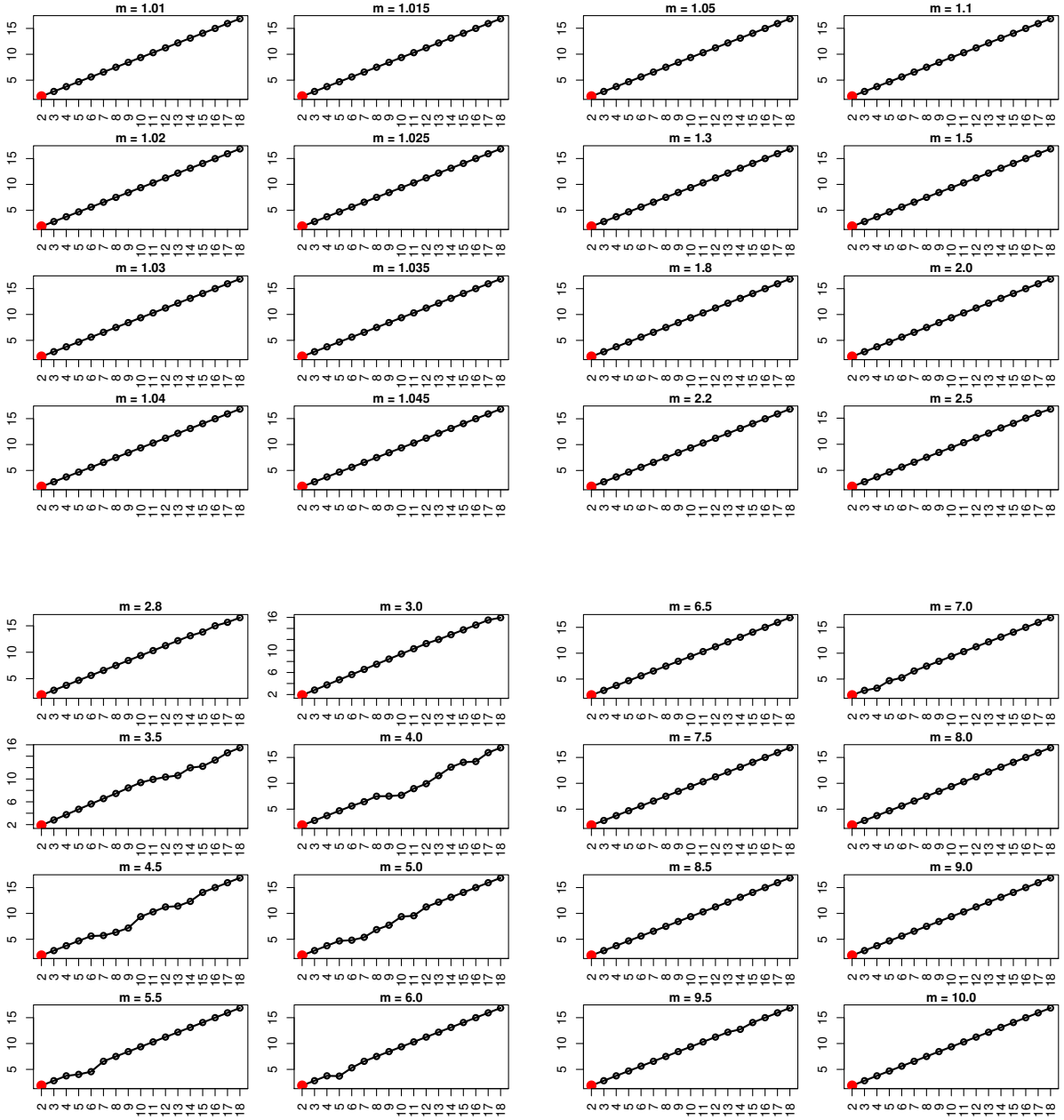


Tabela 109 – Irish-Sentiment

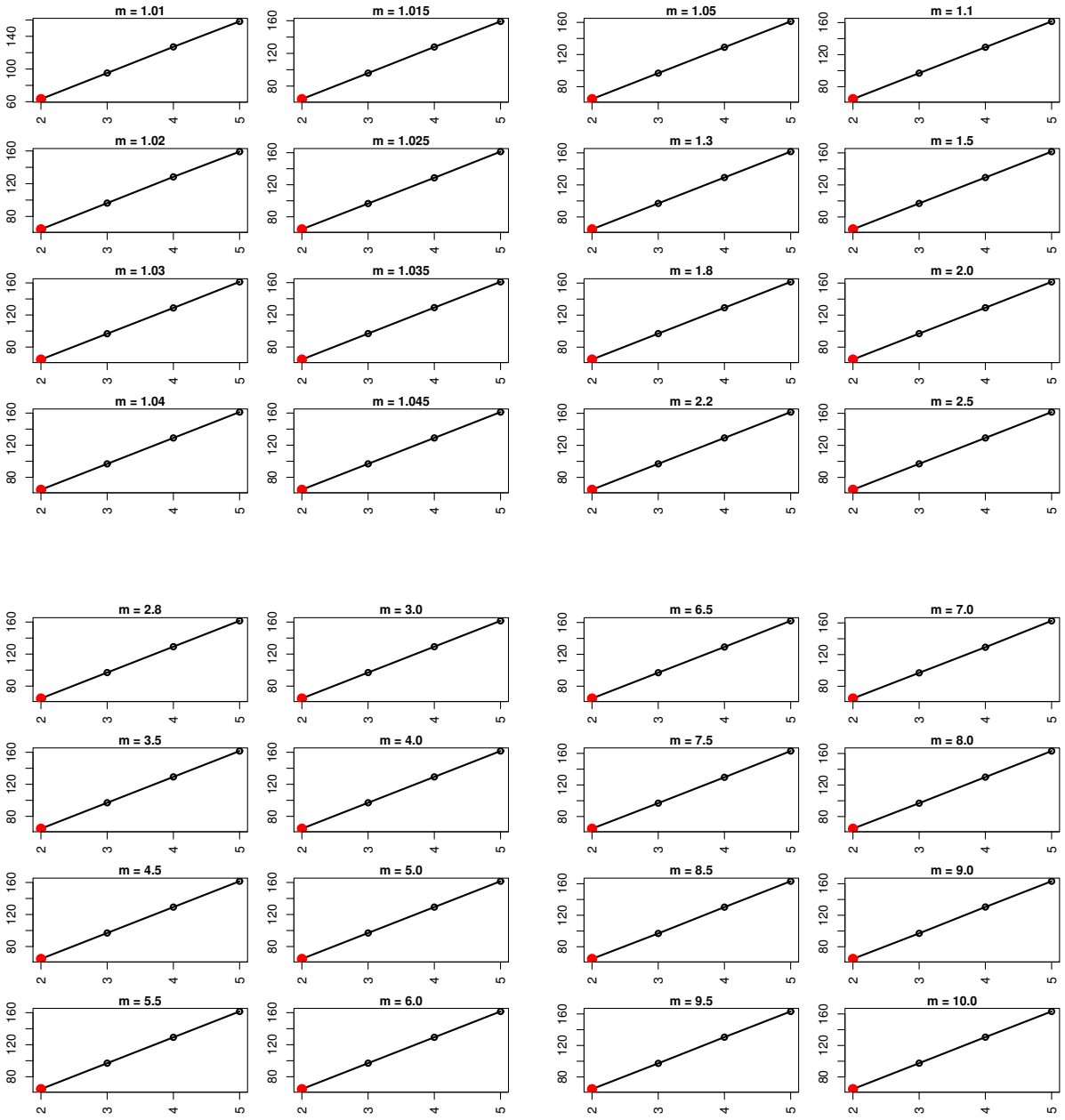


Tabela 110 – 20Newsgroups

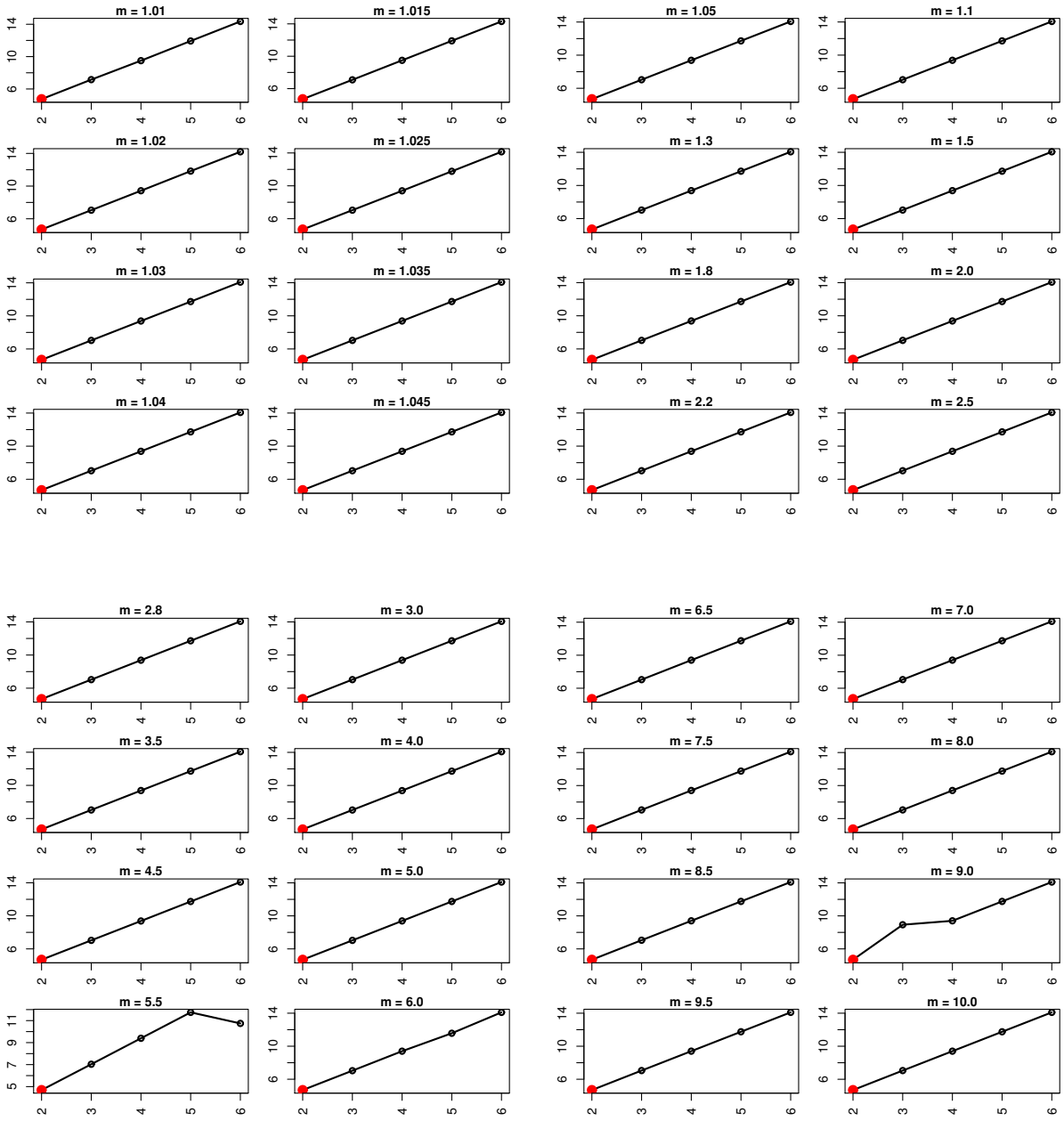


Tabela 111 – La1s

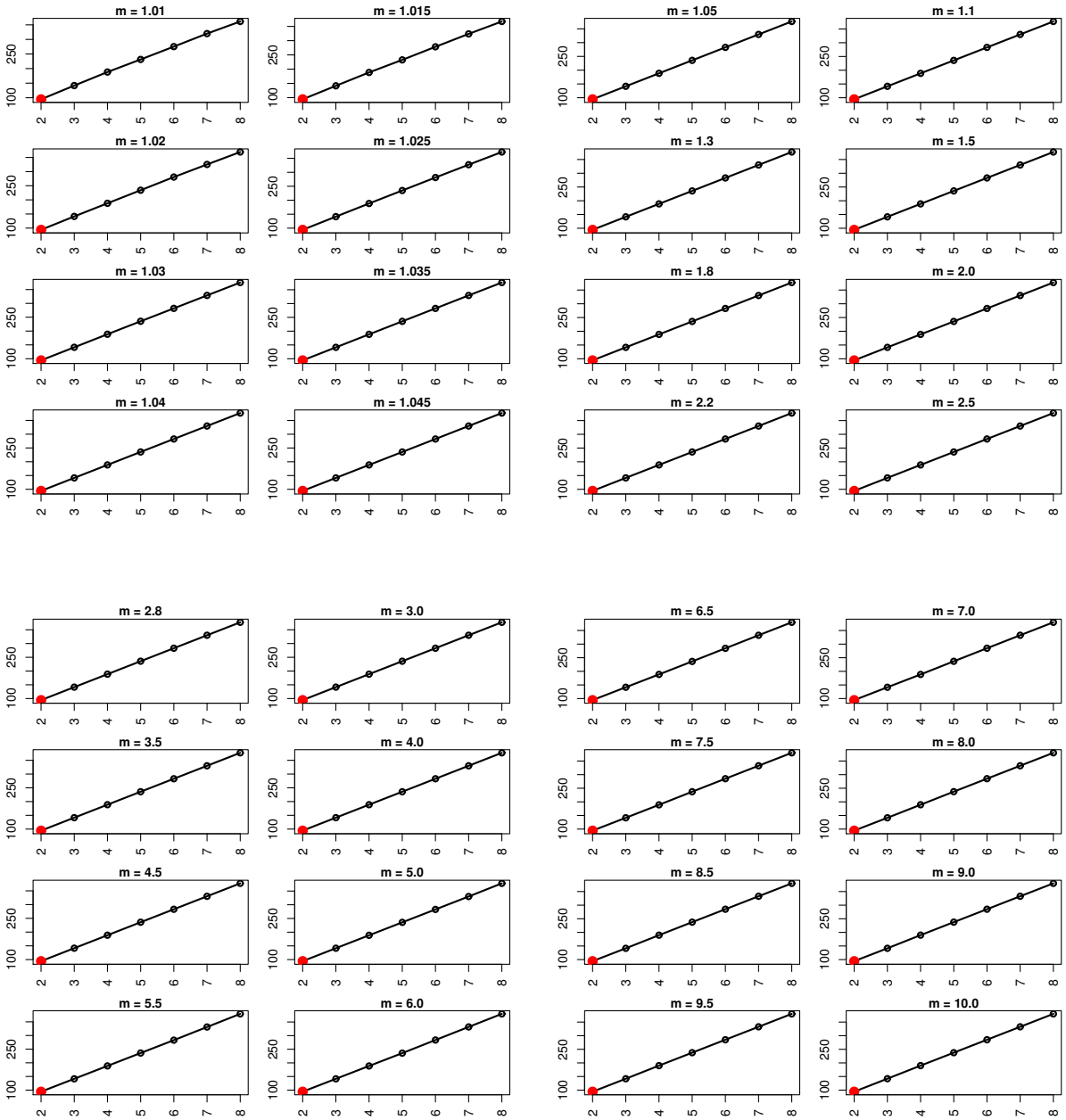
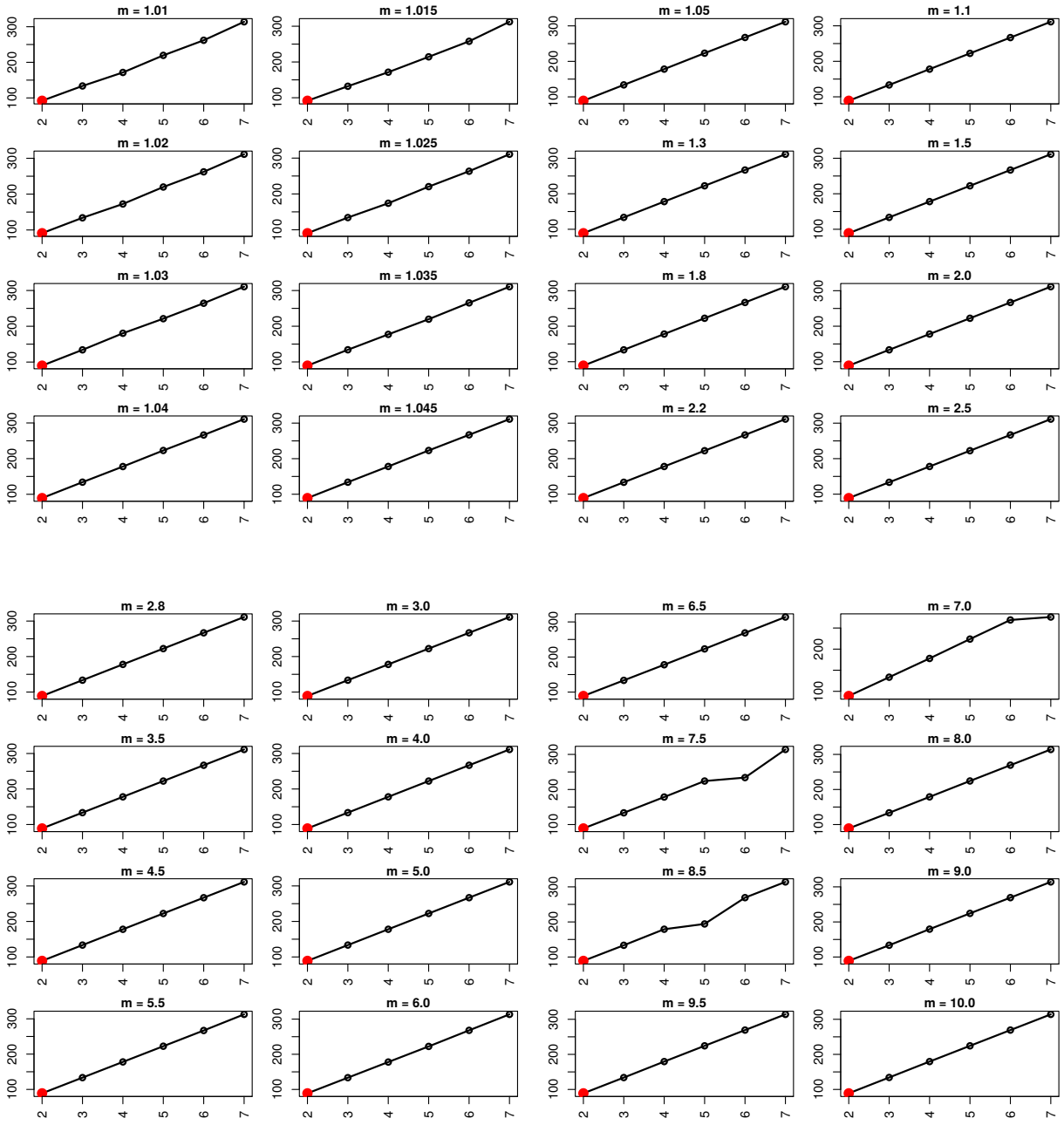


Tabela 112 – Reviews



ANEXO J – XB

Tabela 113 – NewYorkTimes

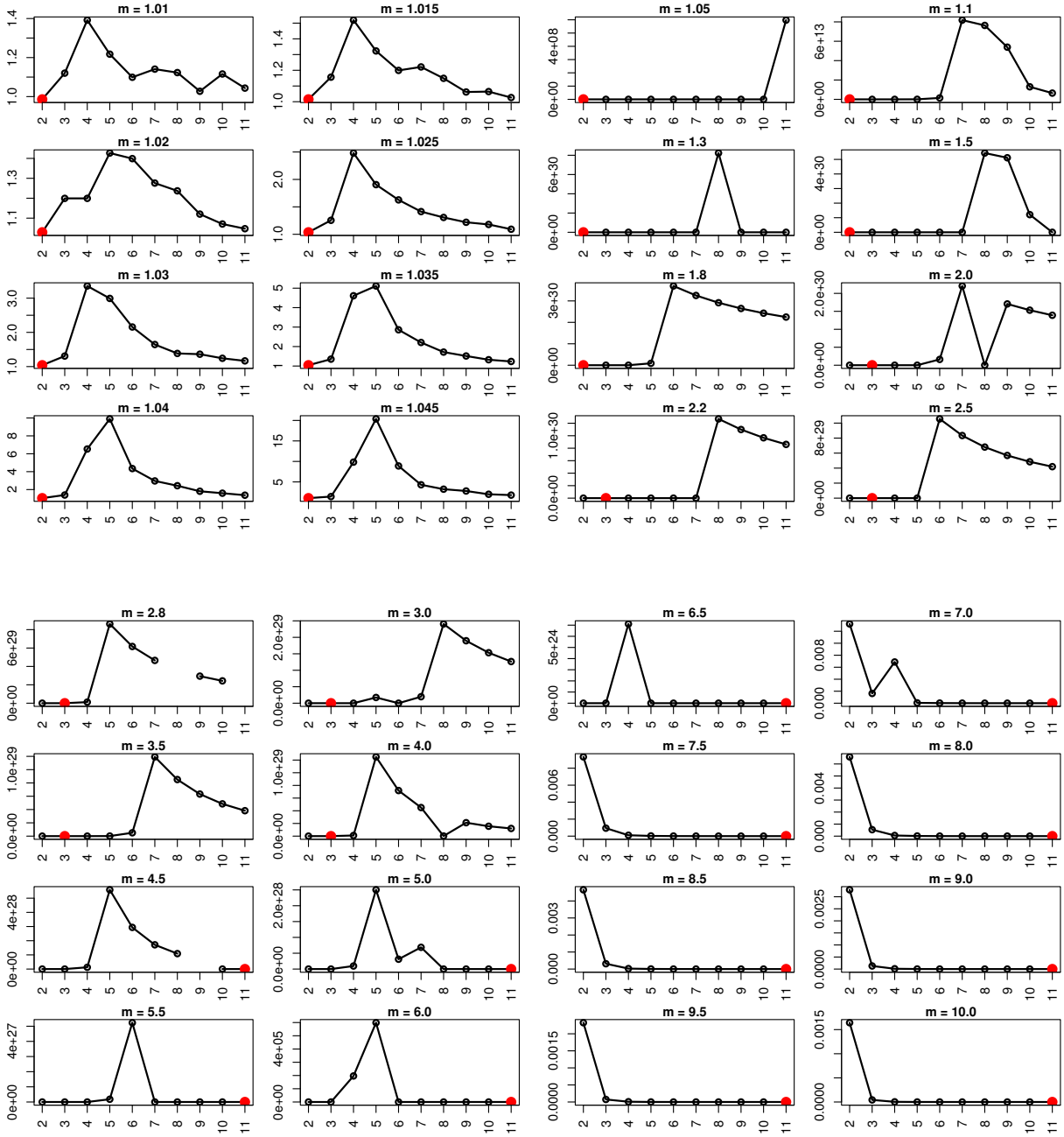


Tabela 114 – IAarticles

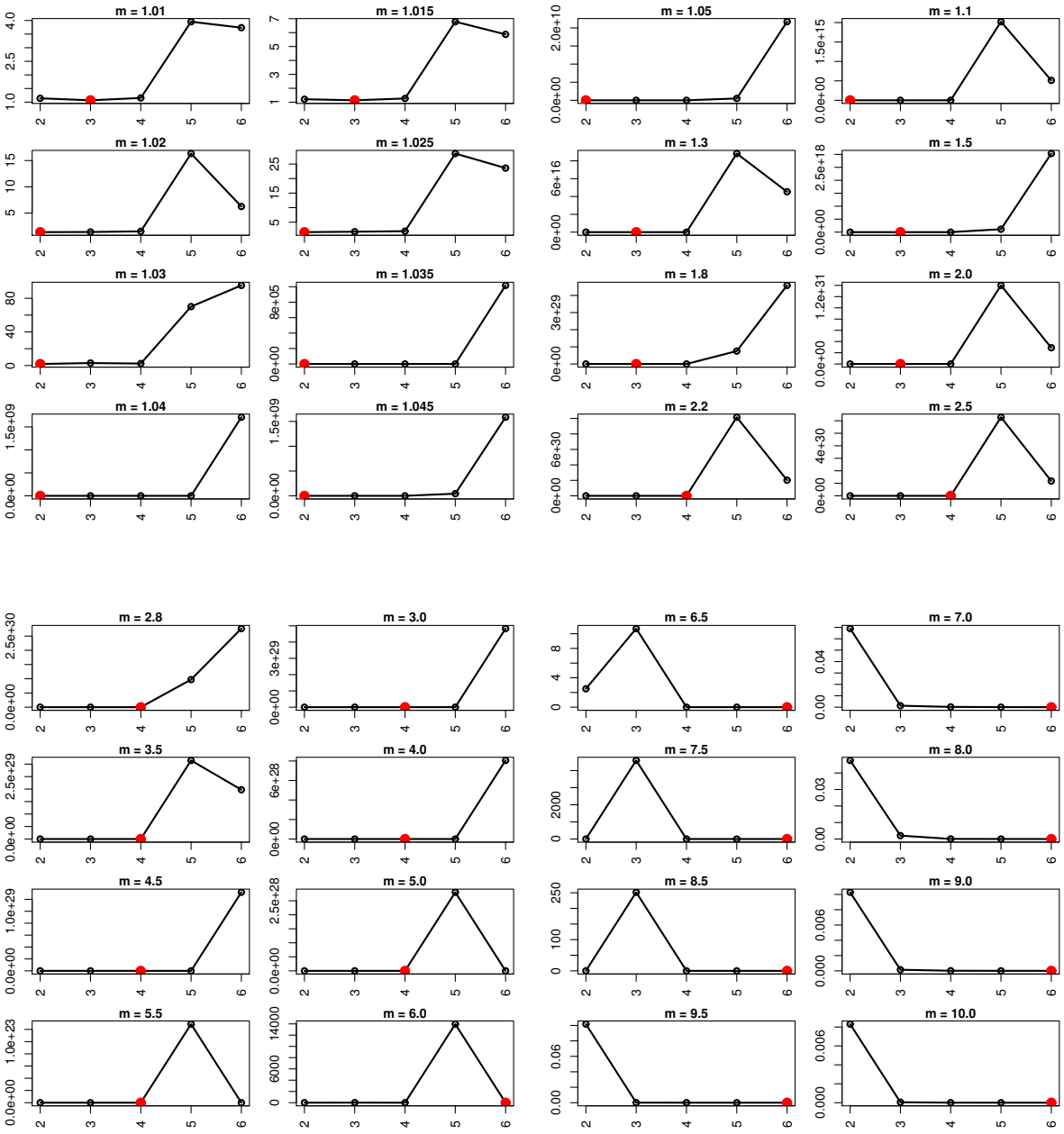


Tabela 115 – Opinosis

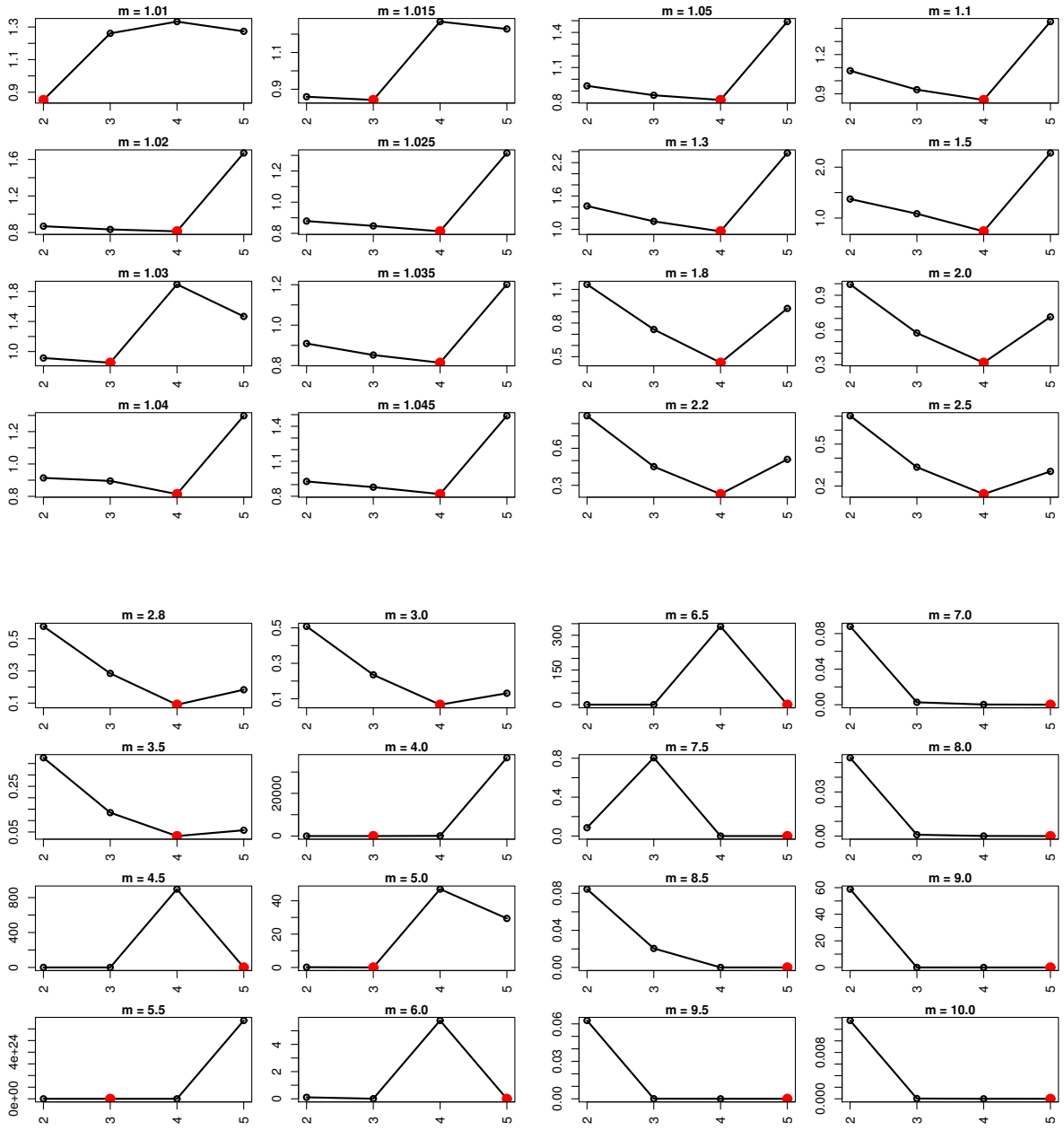


Tabela 116 – CSTR

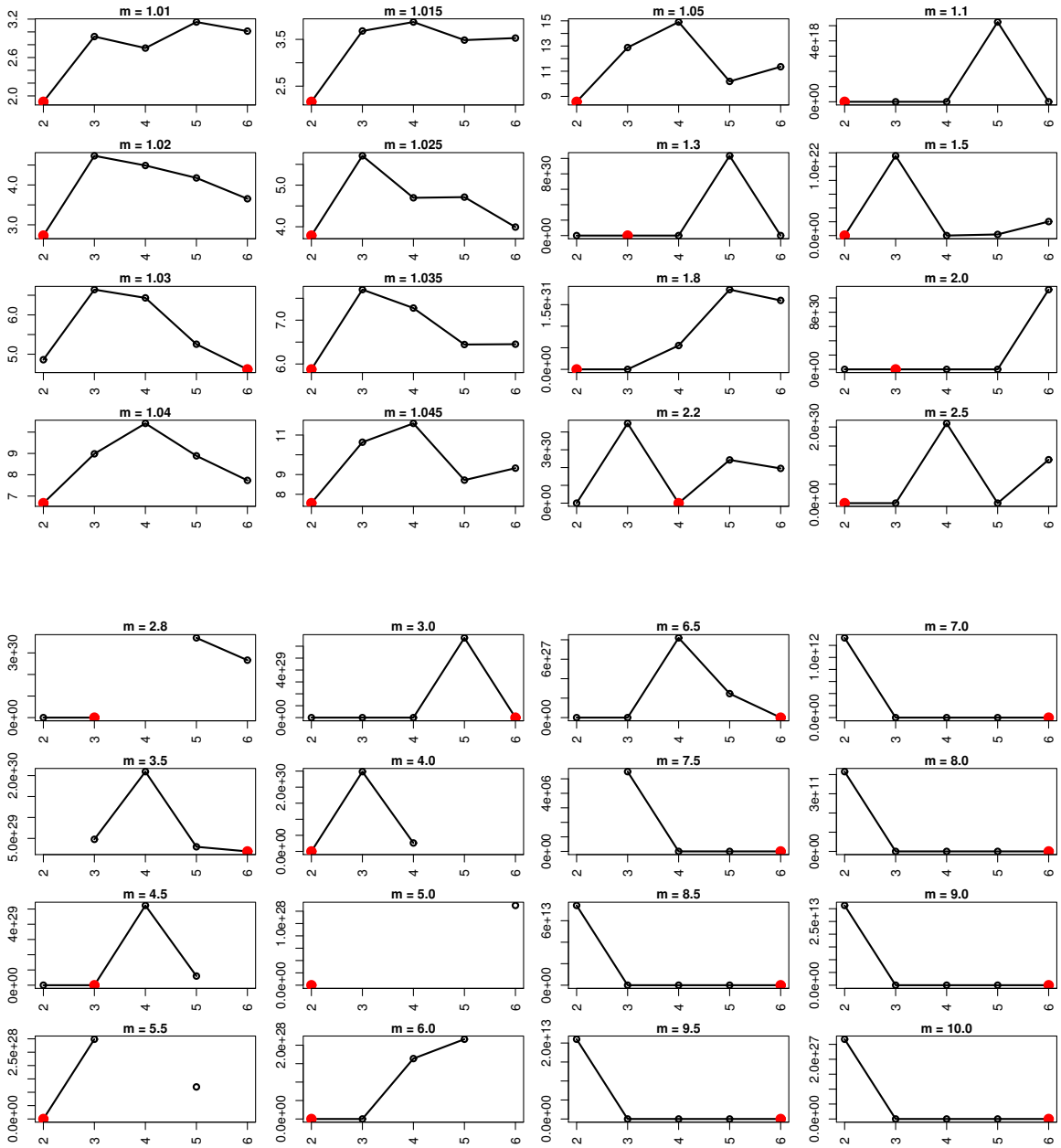


Tabela 117 – SyskillWebert

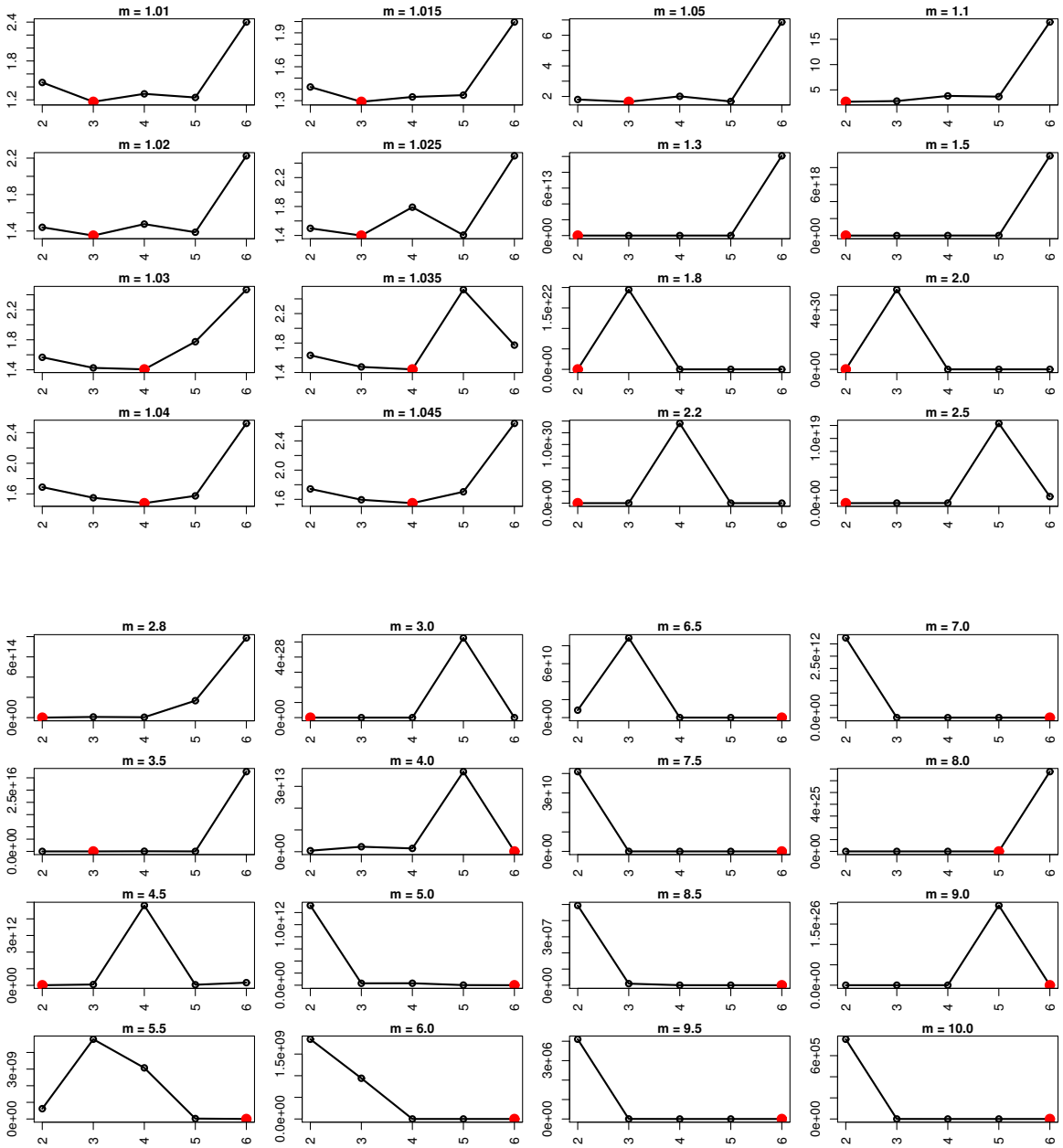


Tabela 118 – Hitech

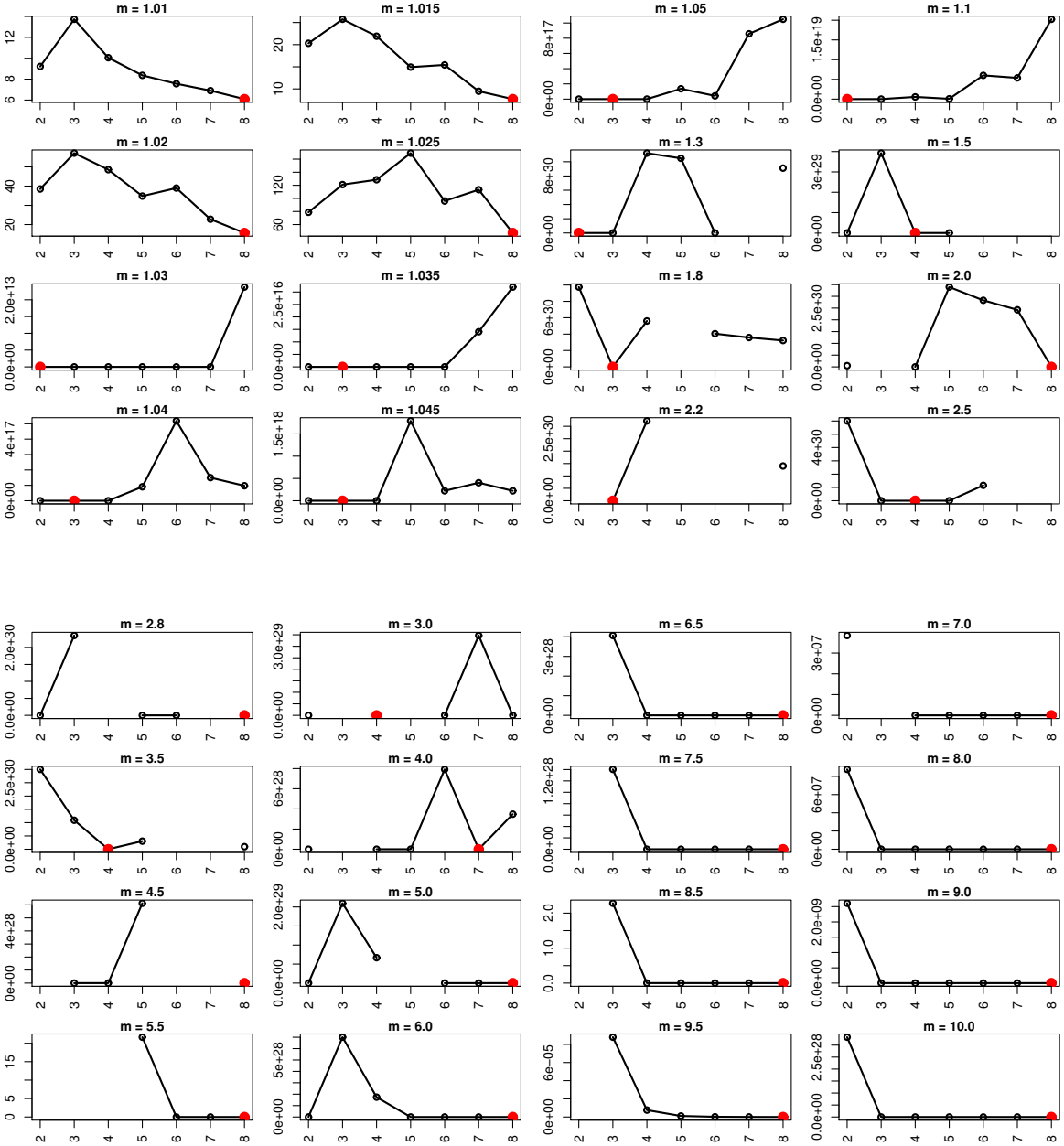


Tabela 119 – WAP

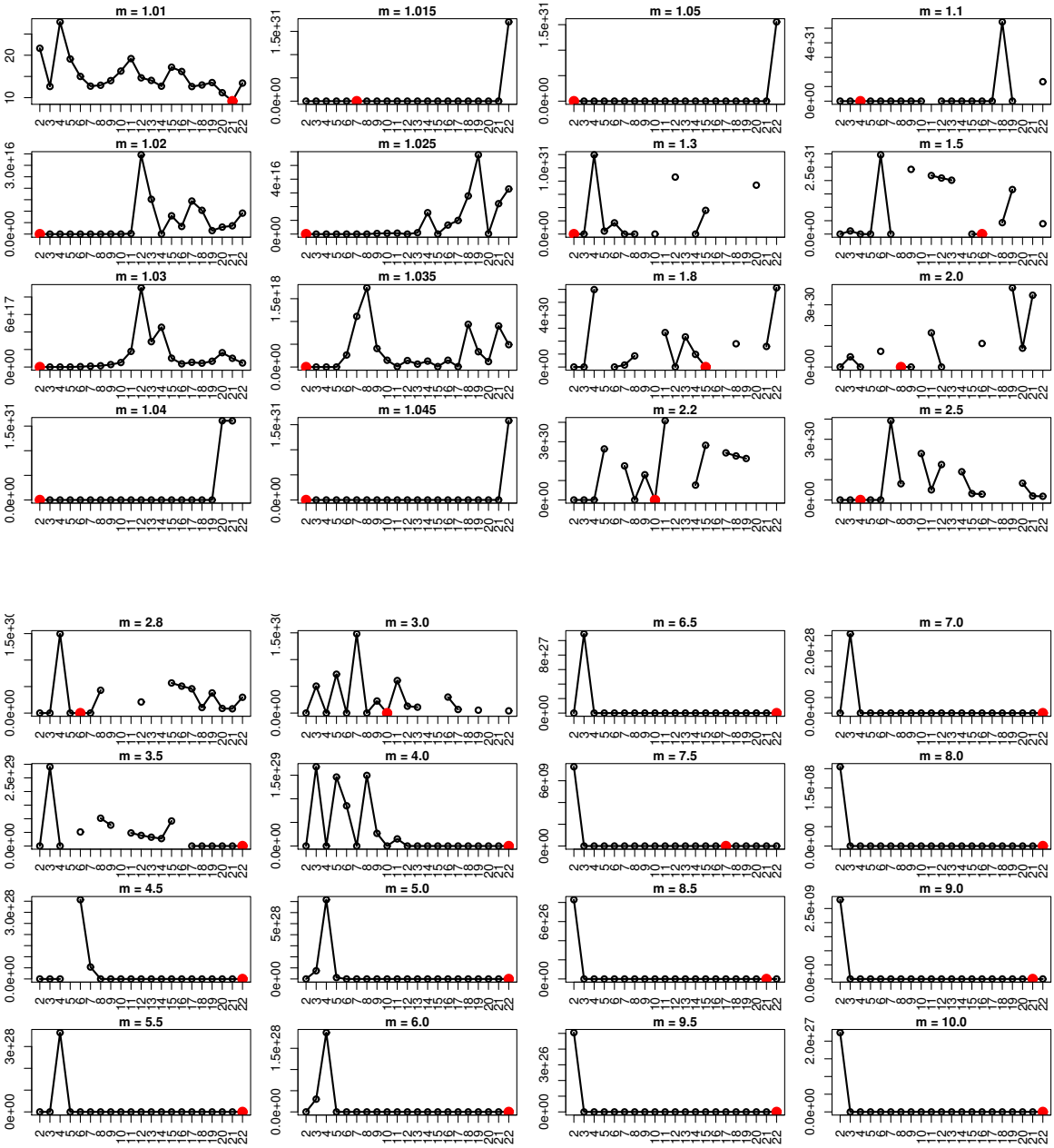


Tabela 120 – NSF

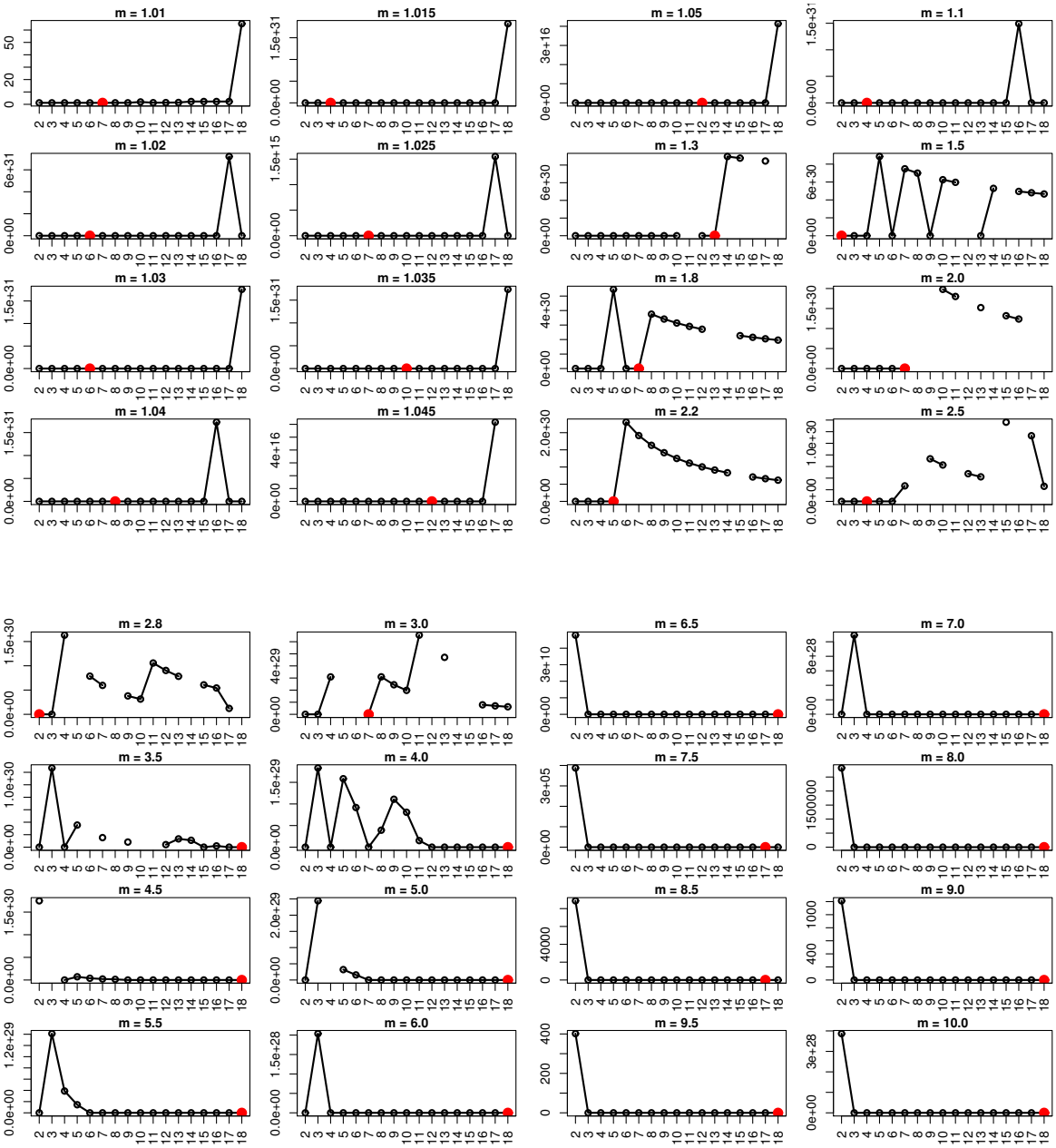


Tabela 121 – Irish-Sentiment

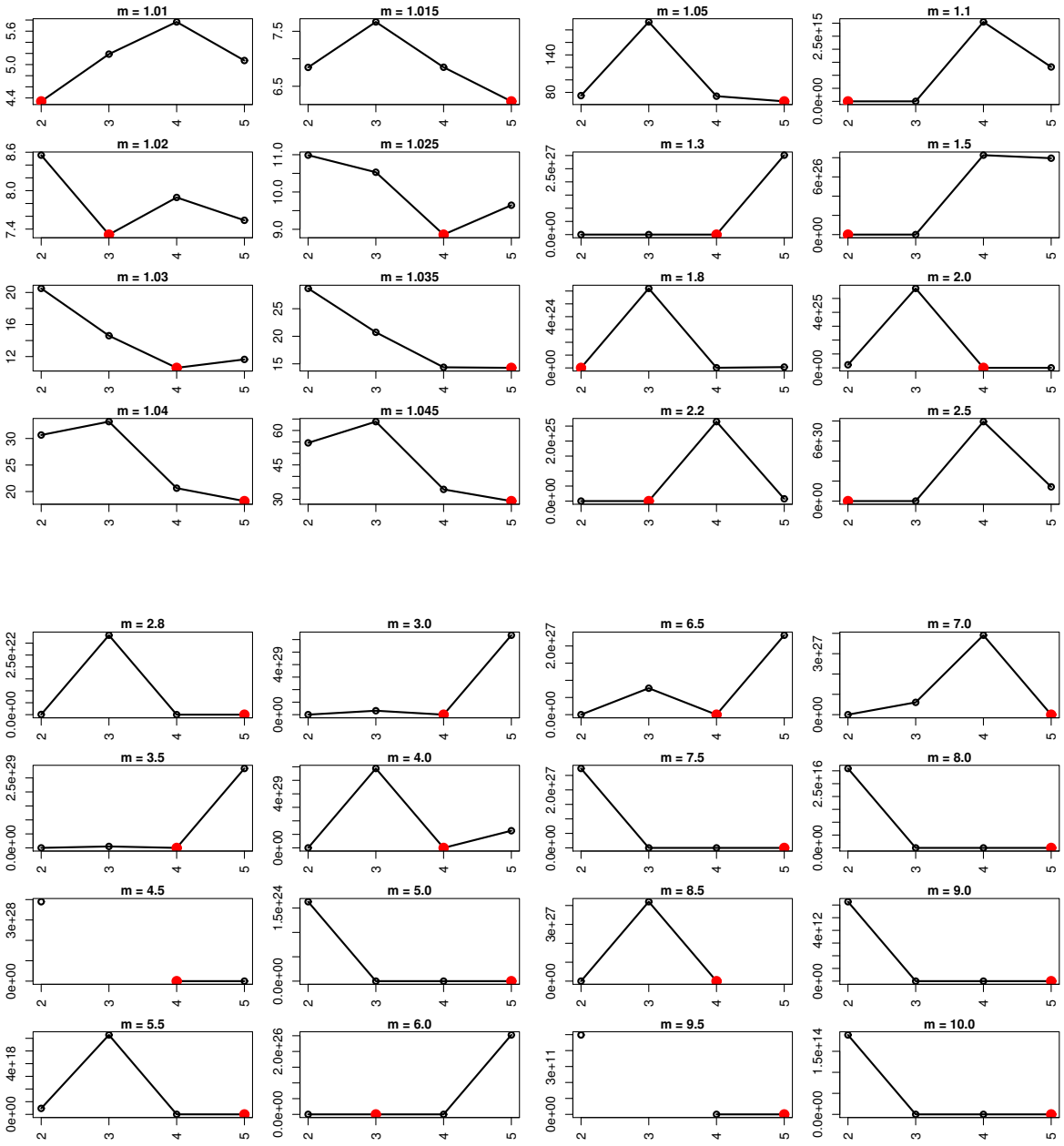


Tabela 122 – 20Newsgroups

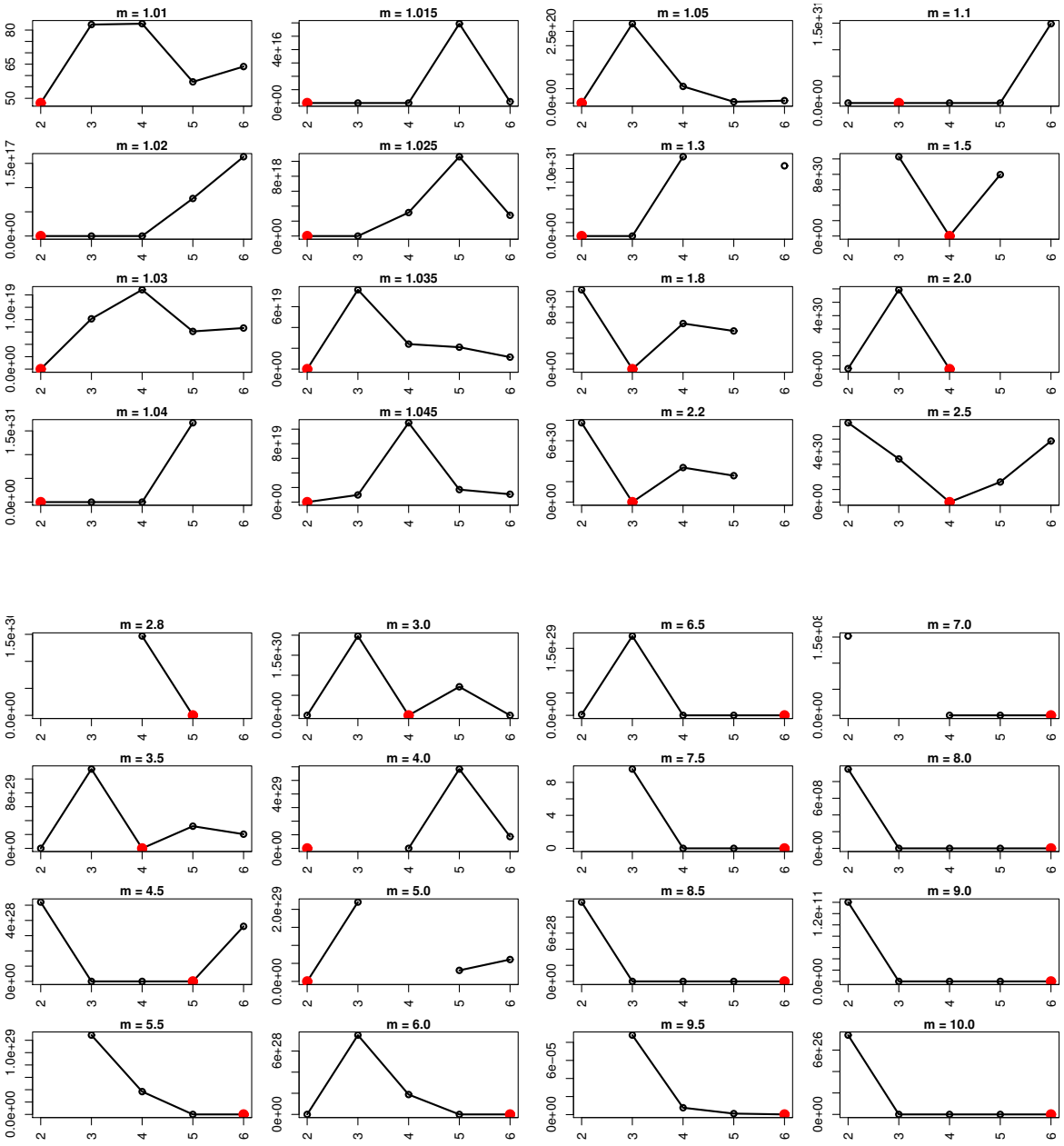


Tabela 123 – La1s

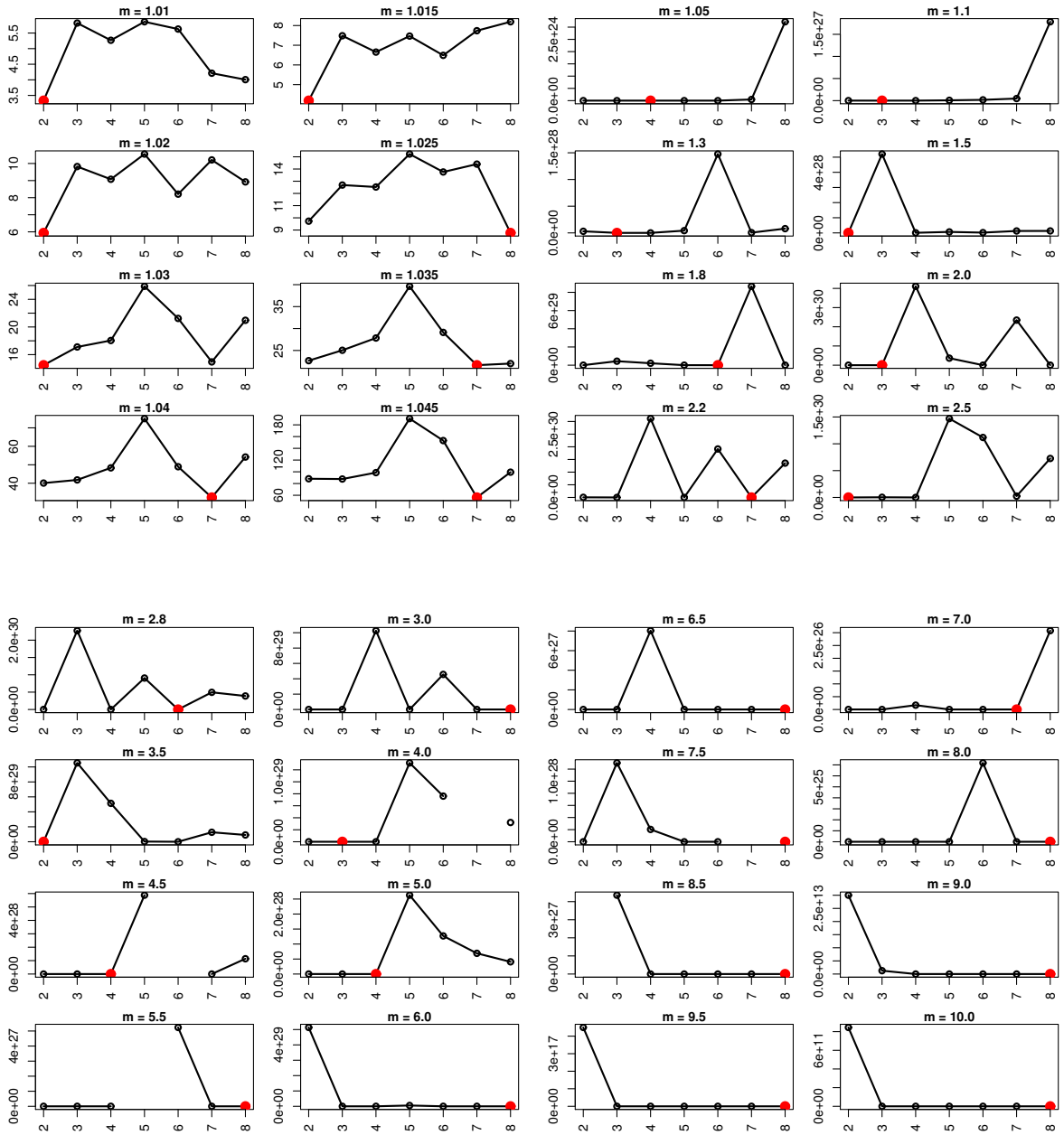
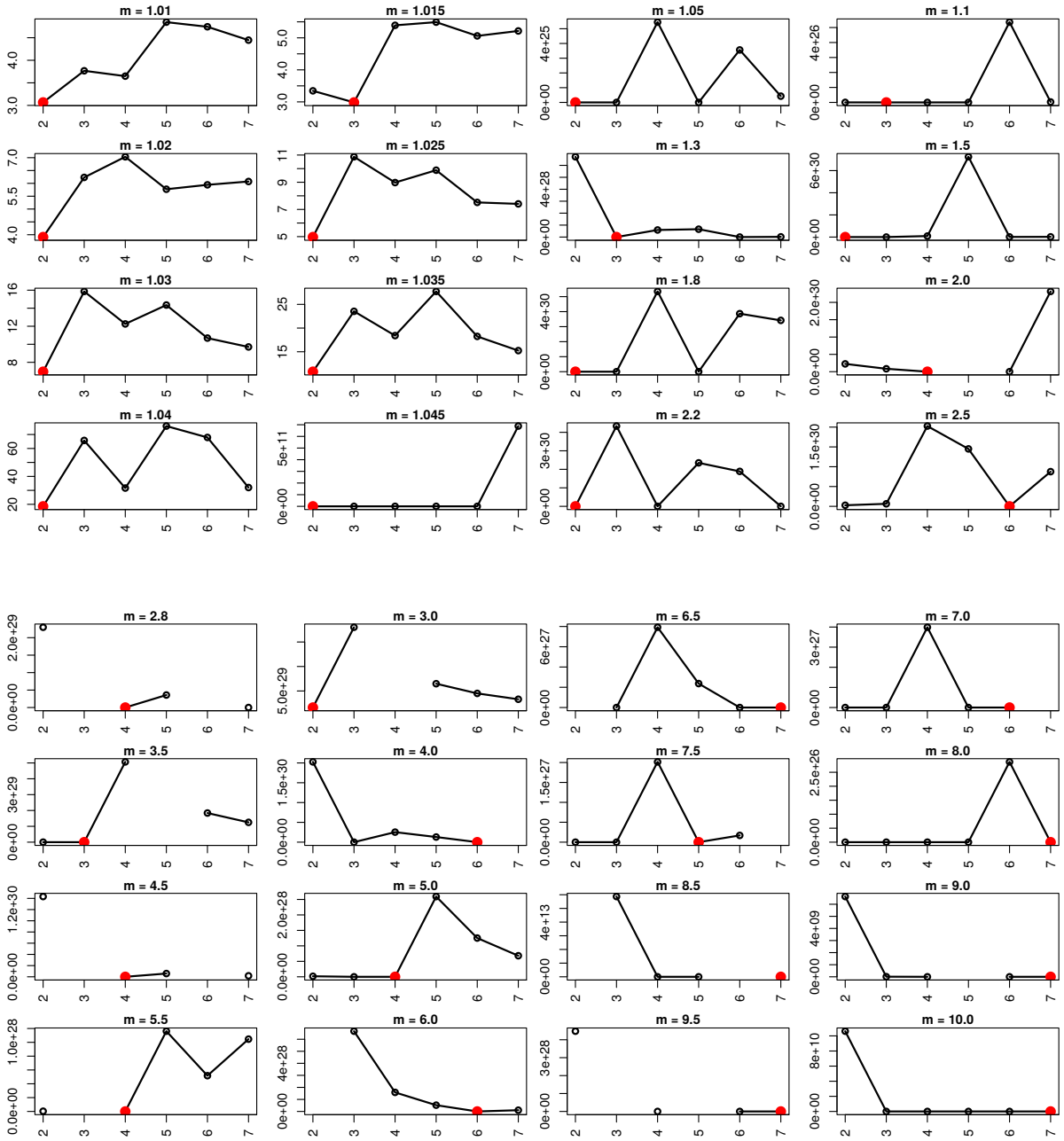


Tabela 124 – Reviews



ANEXO K – K

Tabela 125 – NewYorkTimes

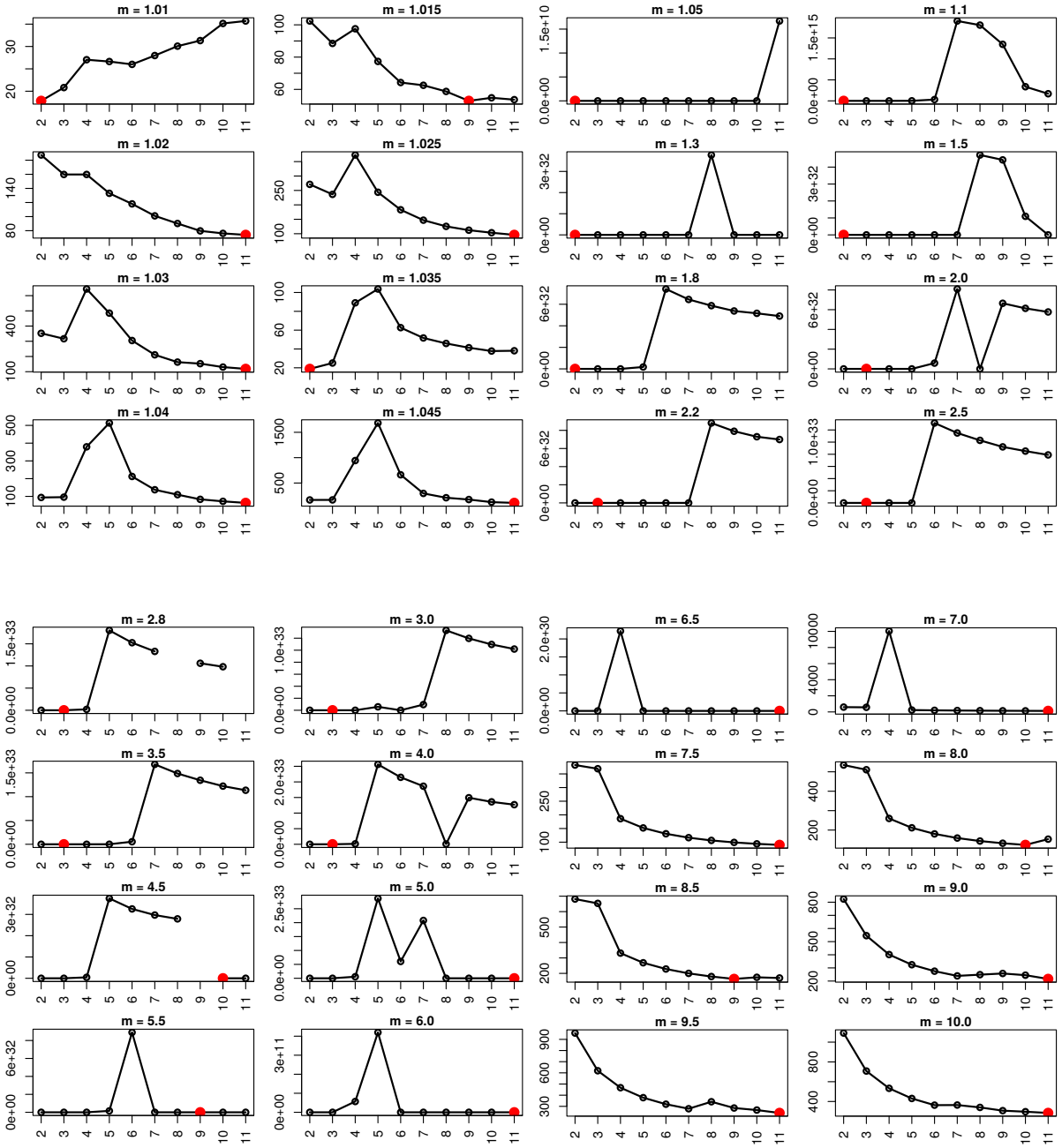


Tabela 126 – IAarticles

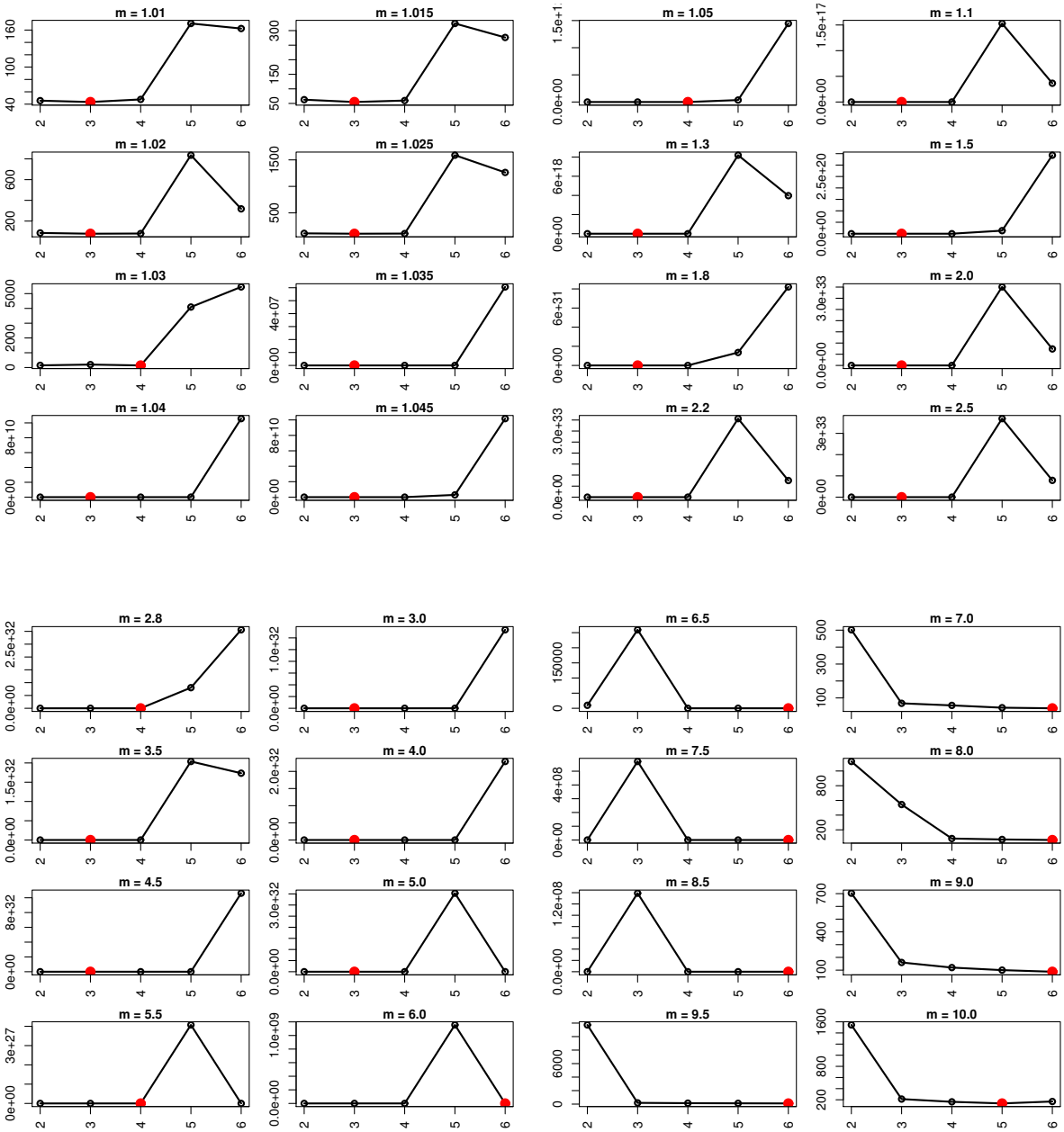


Tabela 127 – Opinosis

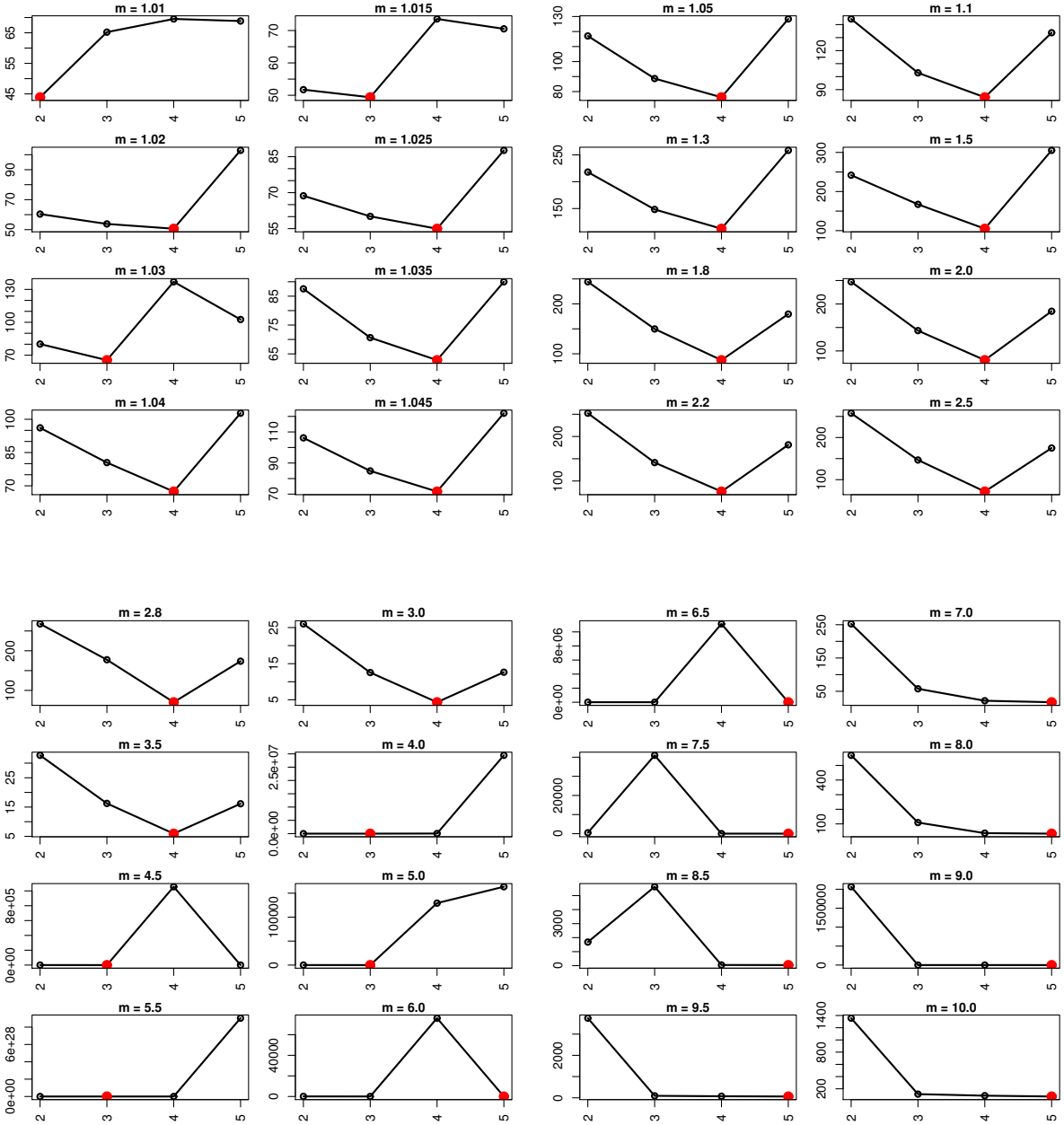


Tabela 128 – CSTR

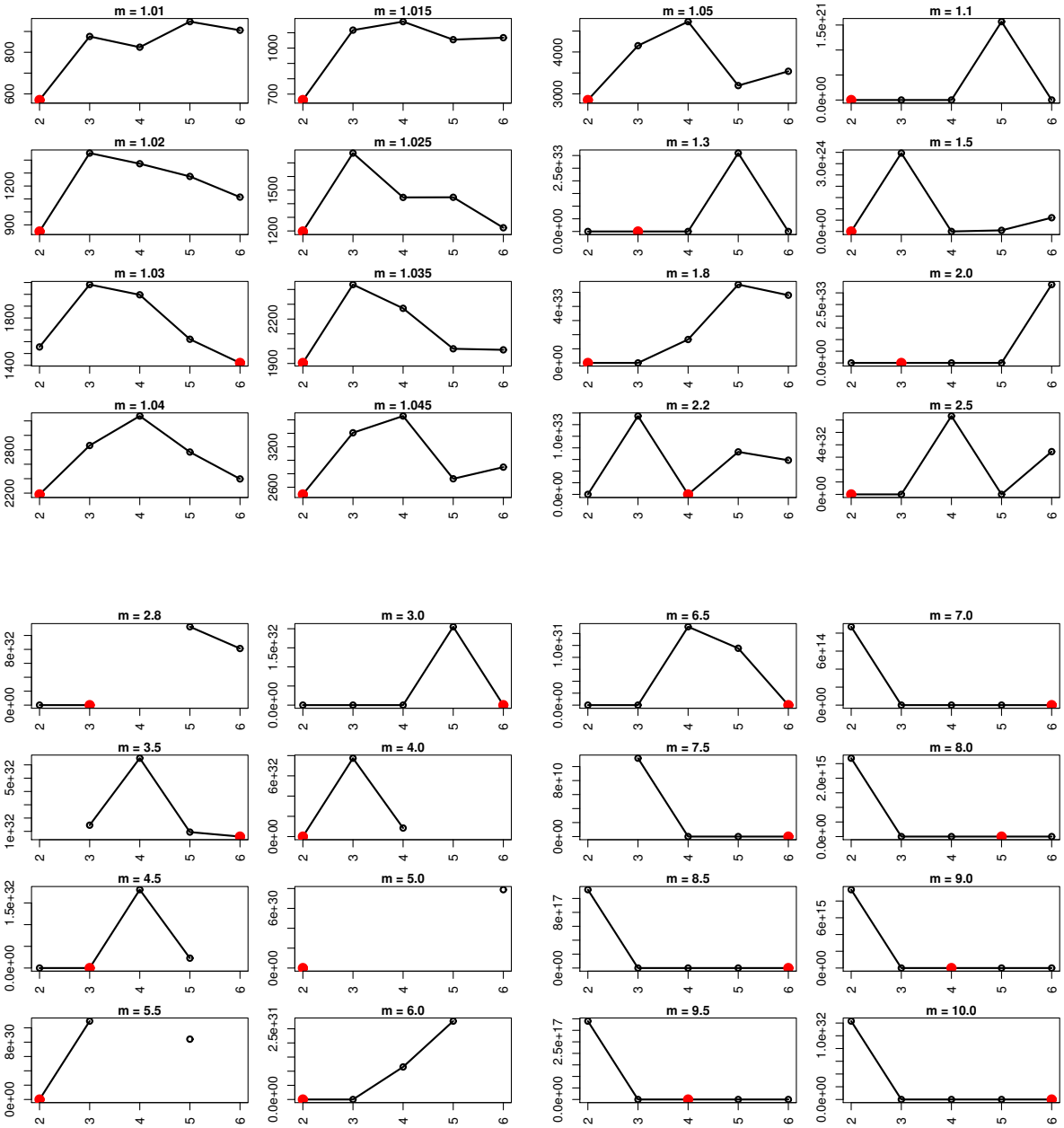


Tabela 129 – SyskillWebert

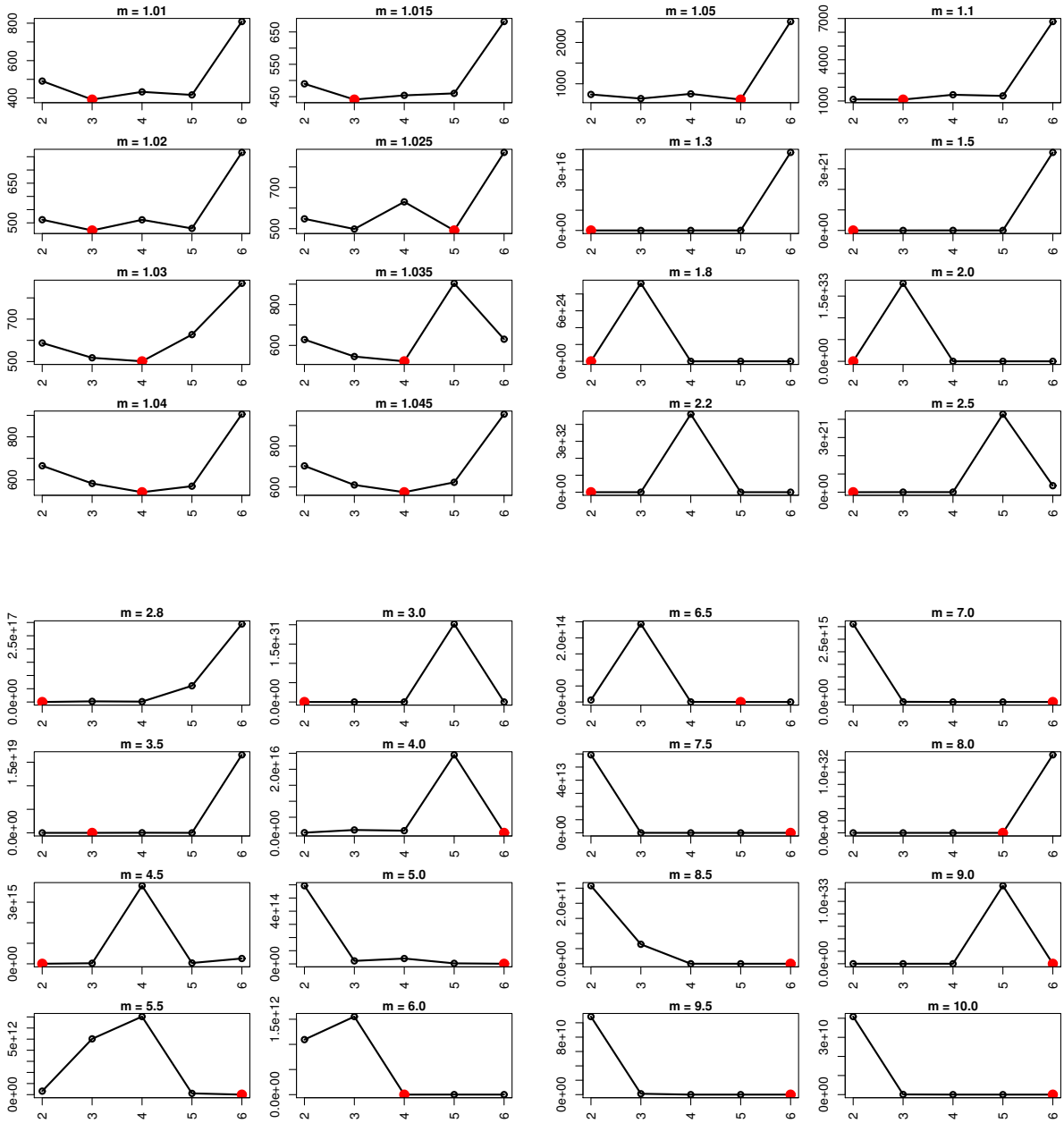


Tabela 130 – Hitech

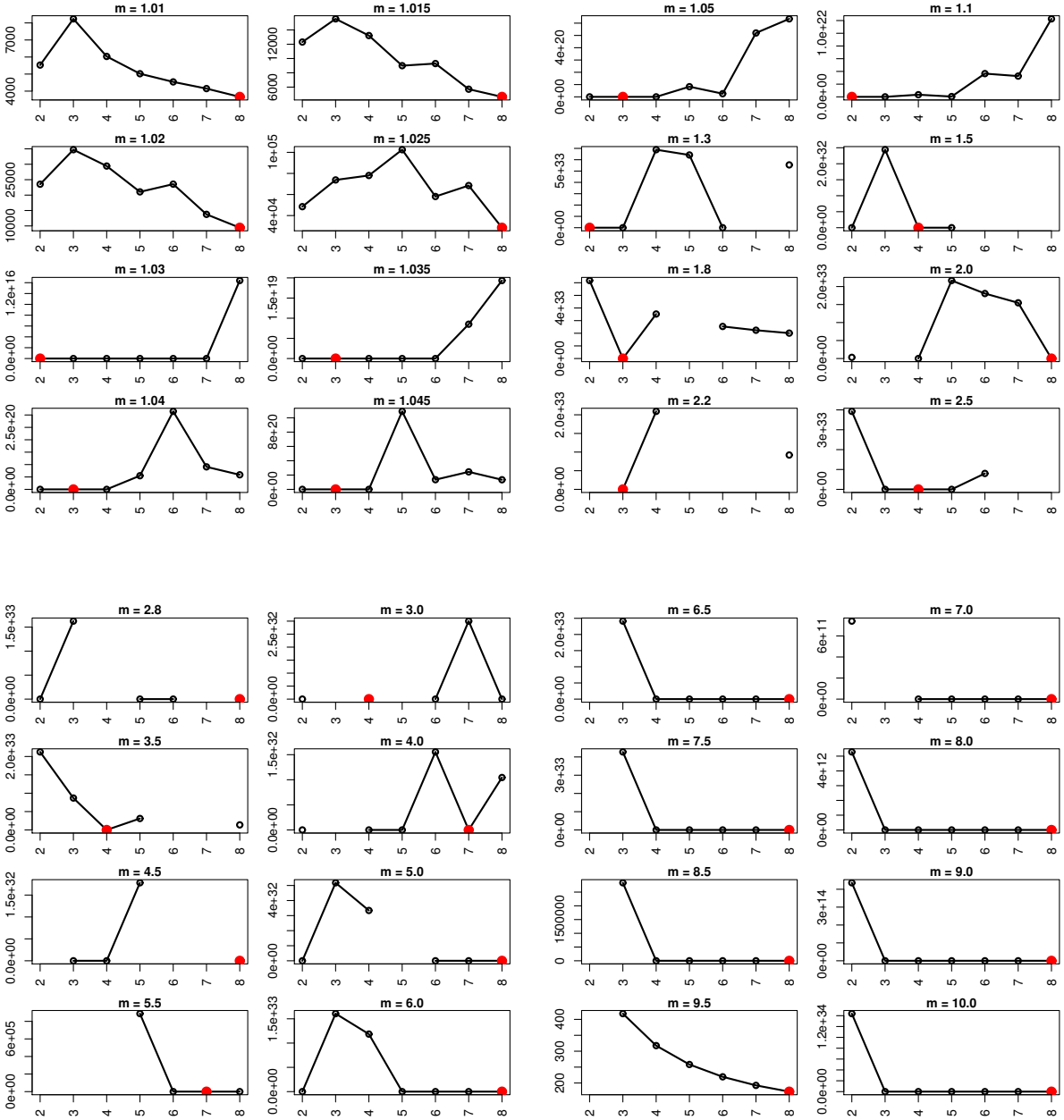


Tabela 131 – WAP

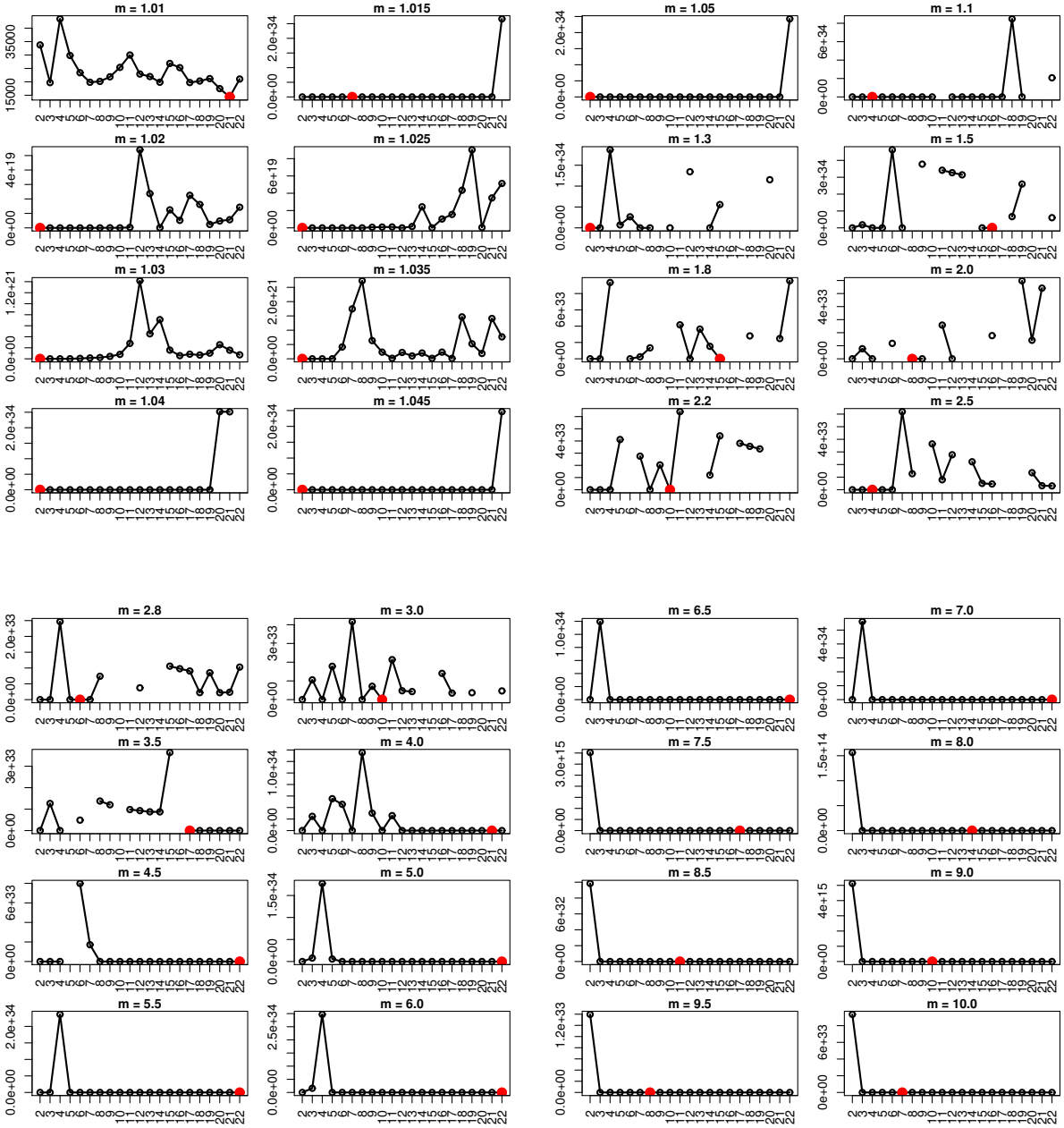


Tabela 132 – NSF

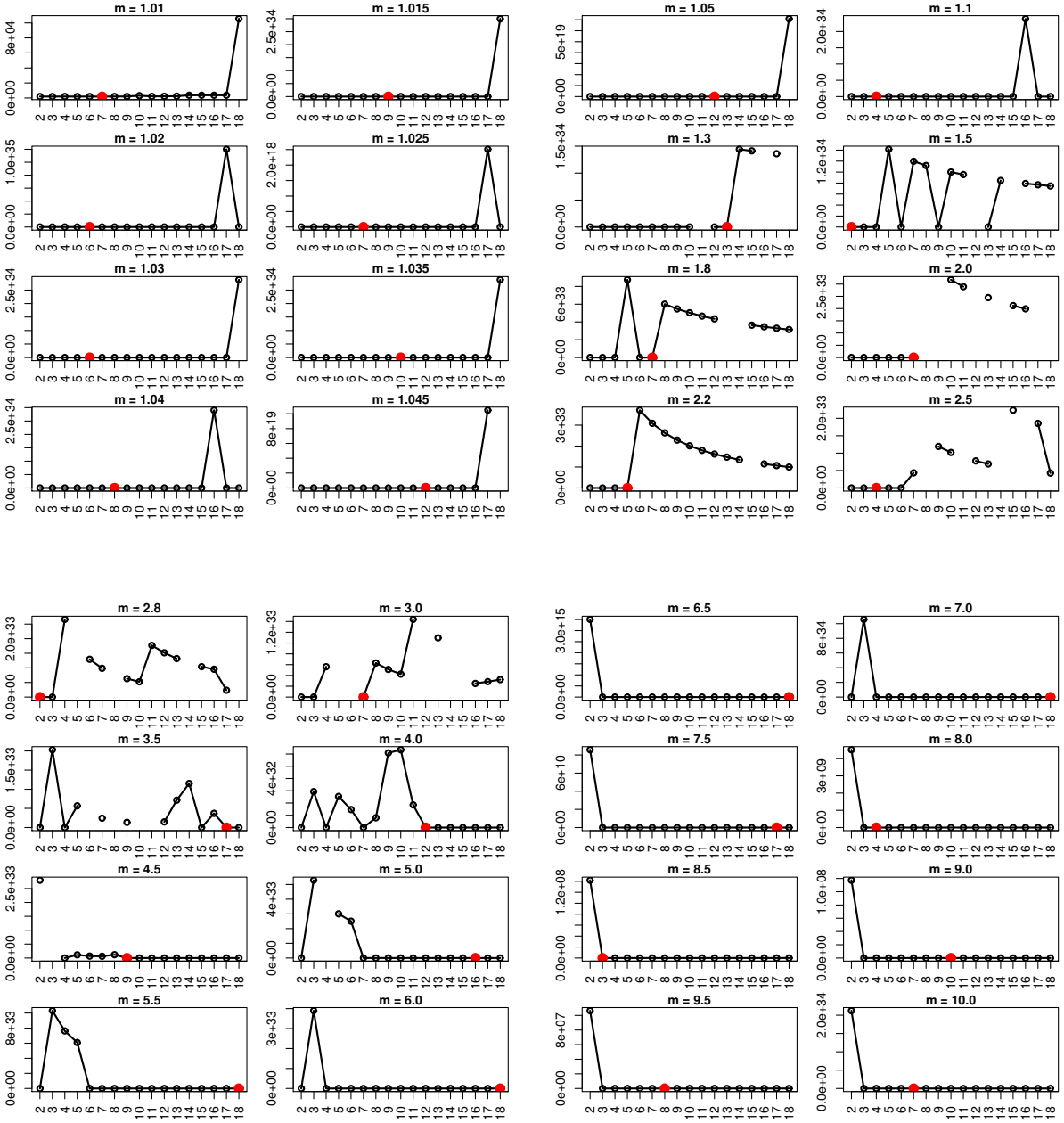


Tabela 133 – Irish-Sentiment

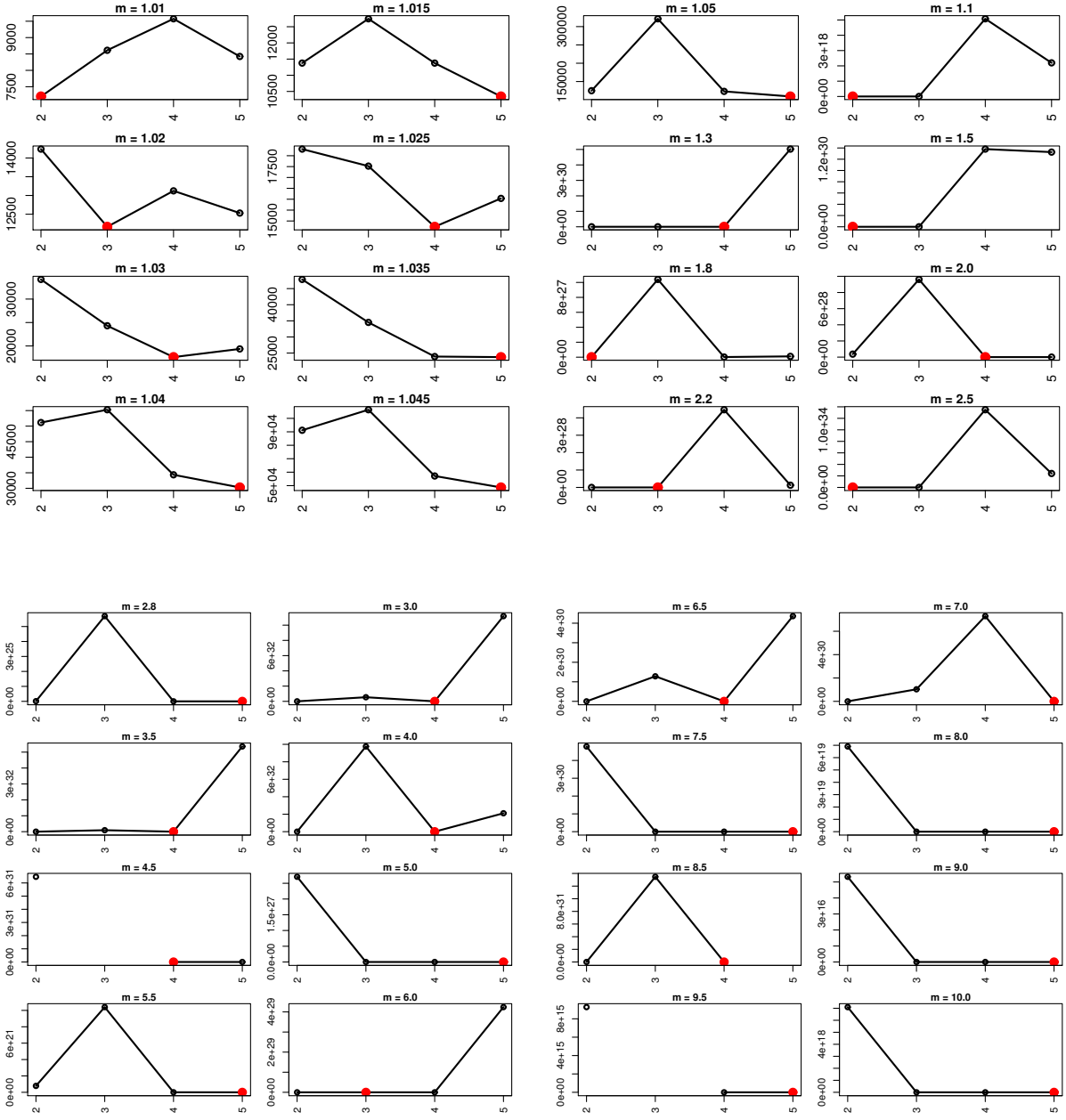


Tabela 134 – 20Newsgroups

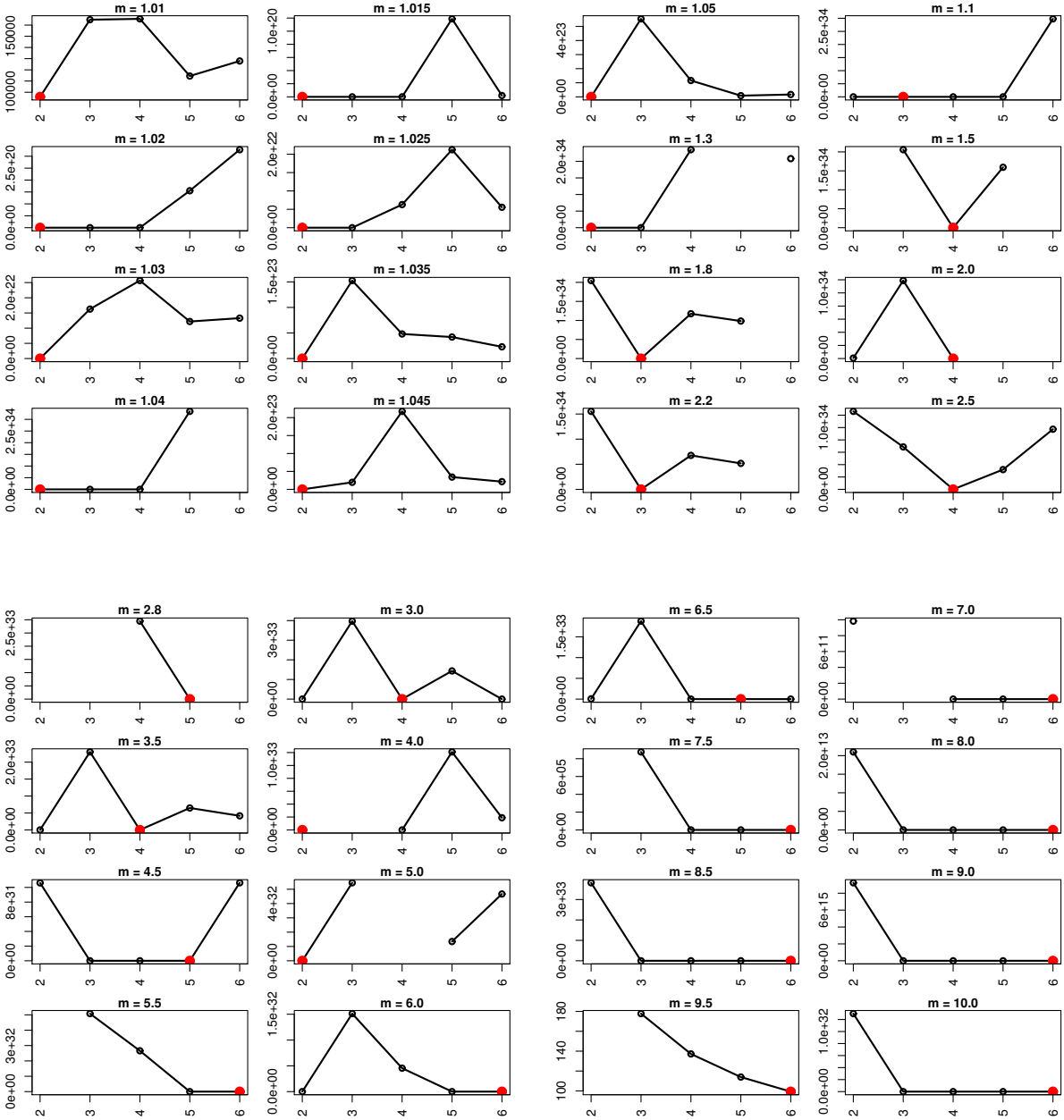


Tabela 135 – La1s

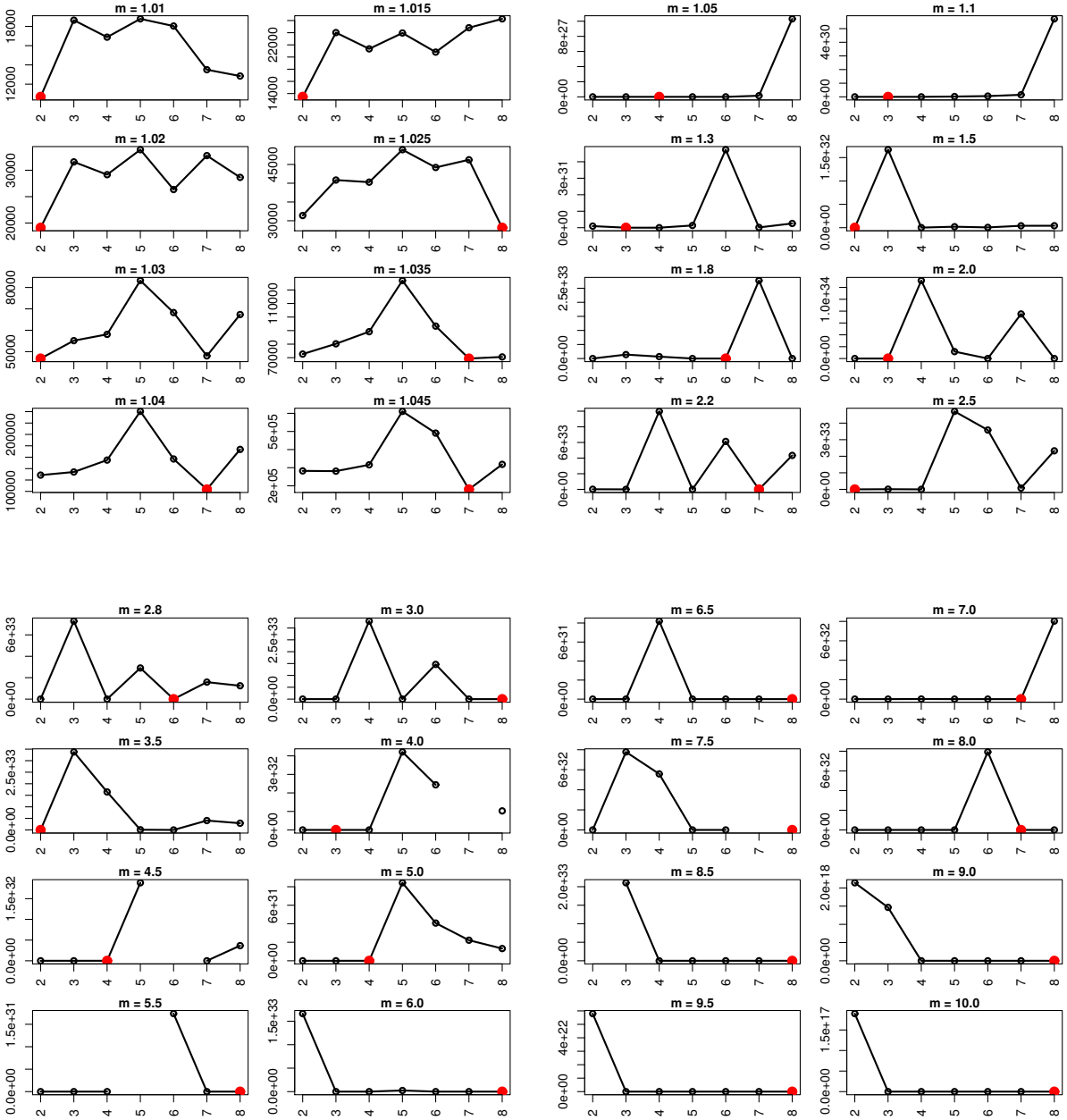
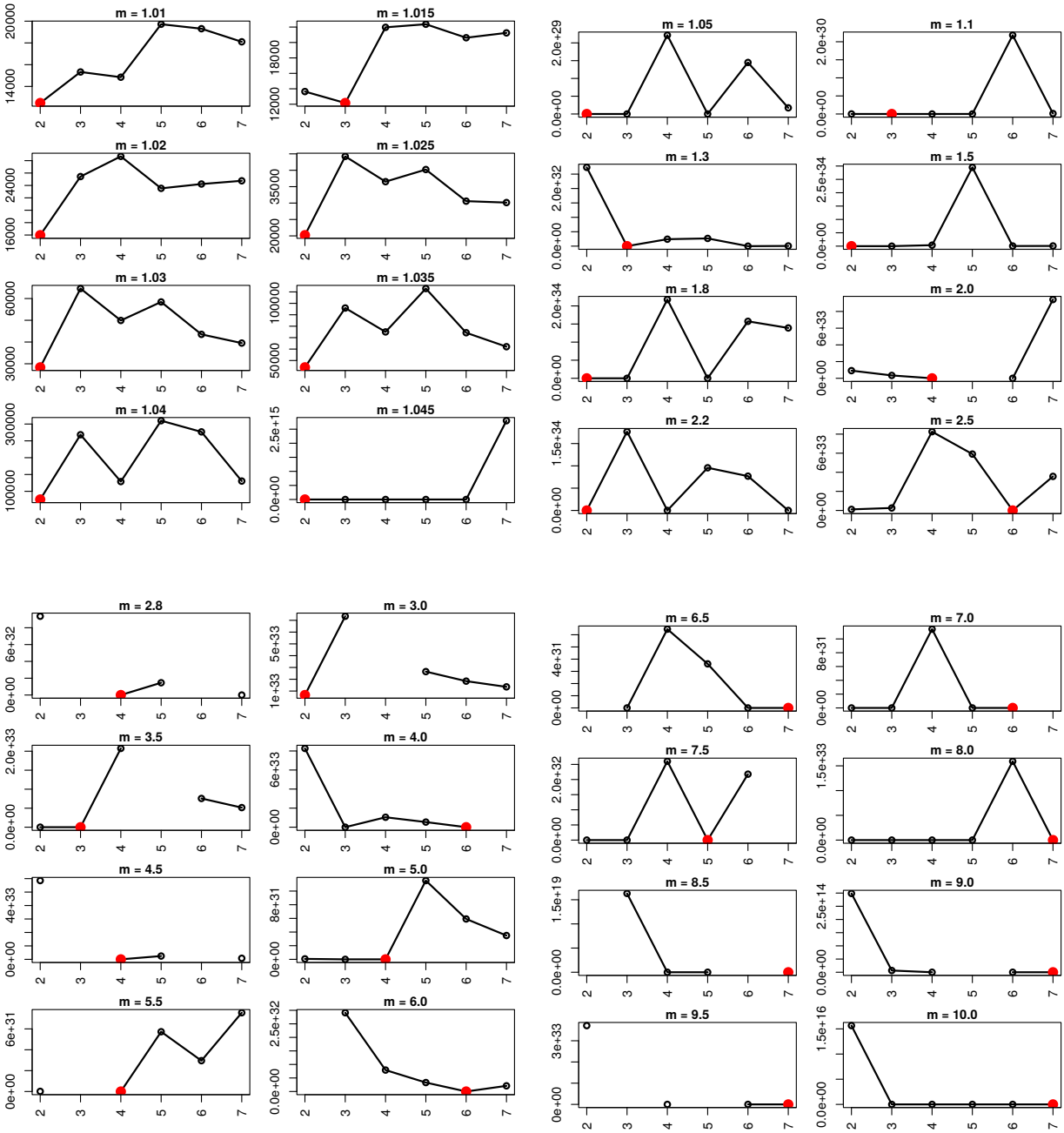


Tabela 136 – Reviews



ANEXO L – SC

Tabela 137 – NewYorkTimes

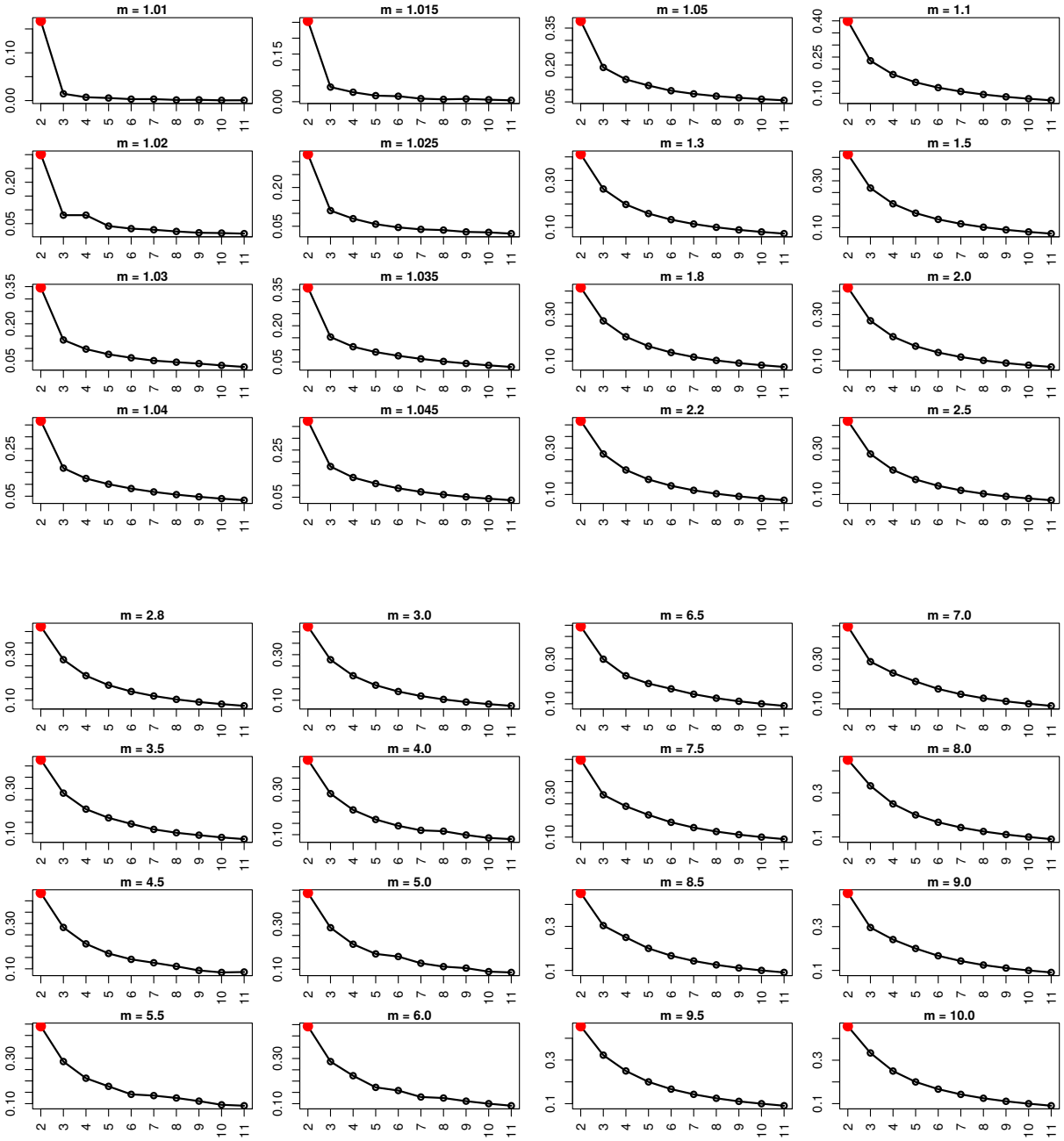


Tabela 138 – IAarticles

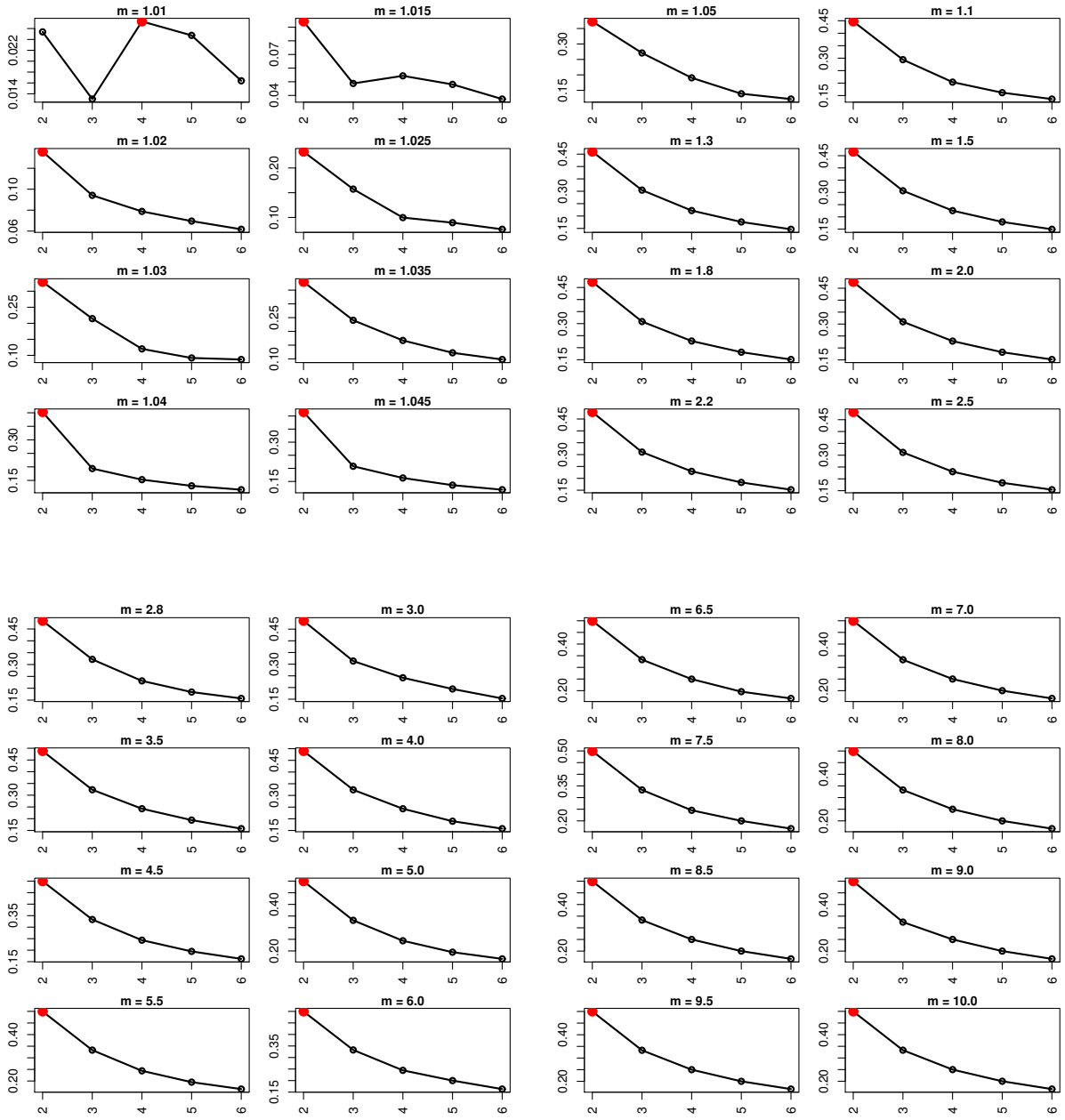


Tabela 139 – Opinosis

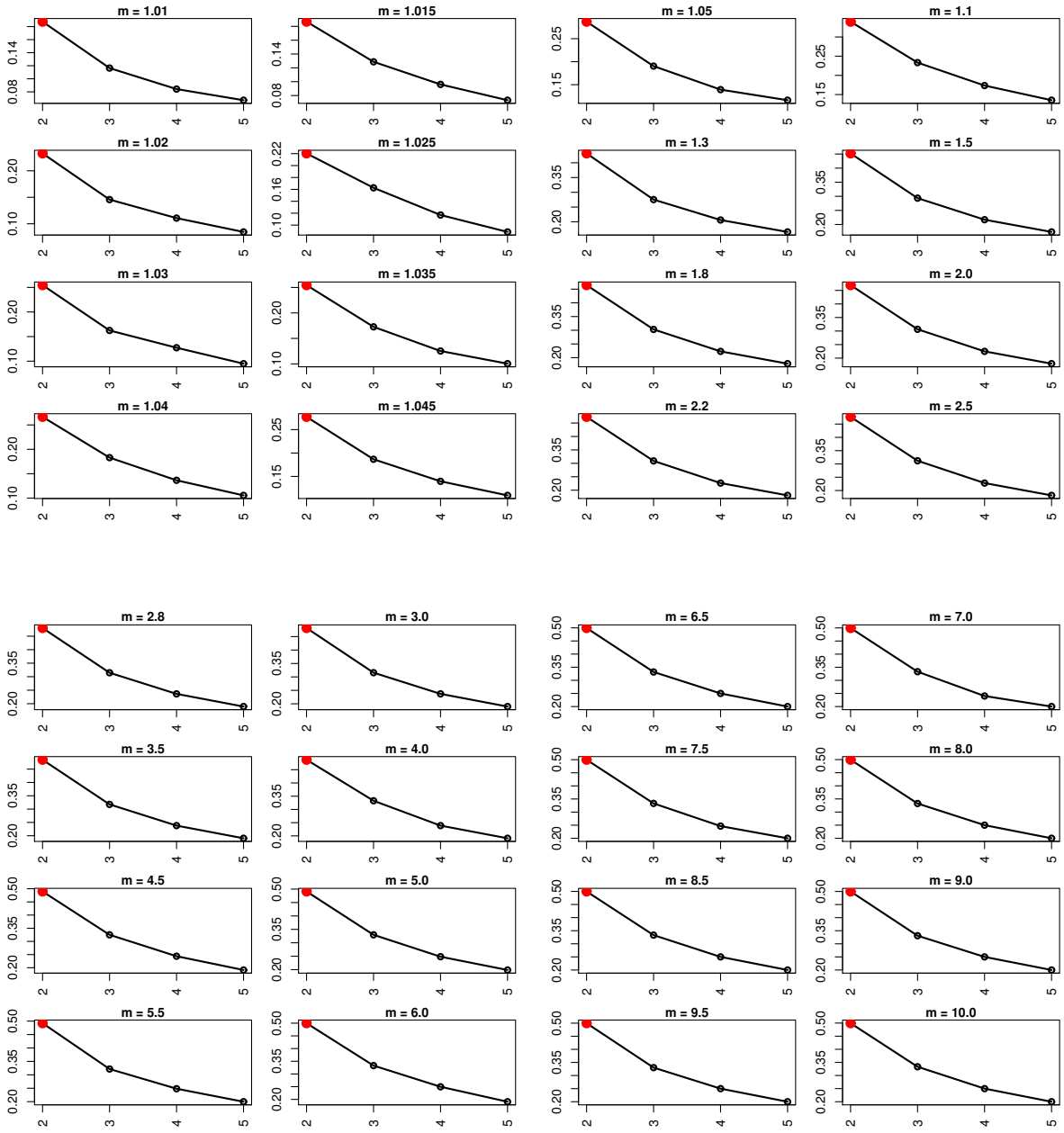


Tabela 140 – CSTR

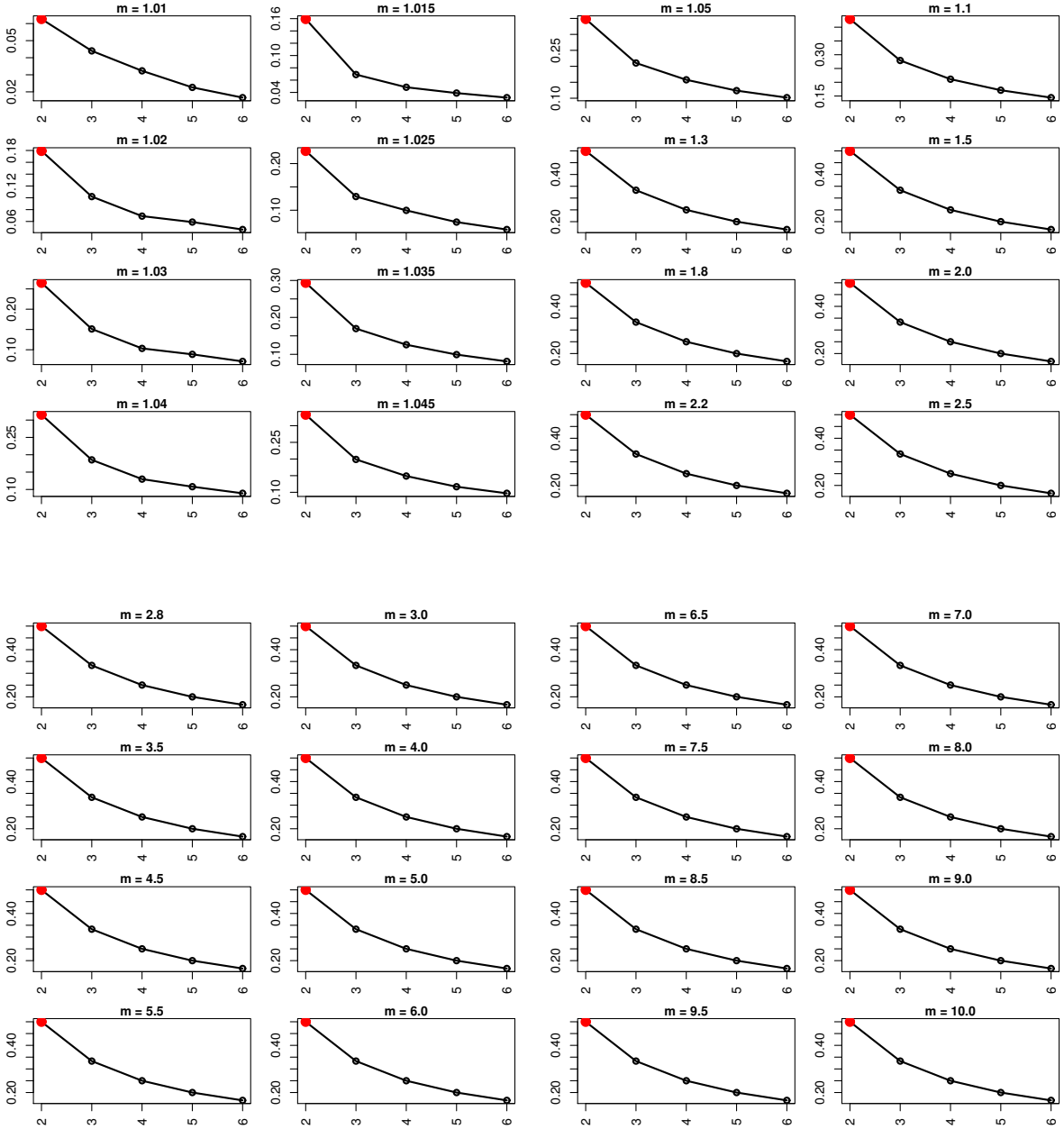


Tabela 141 – SyskillWebert

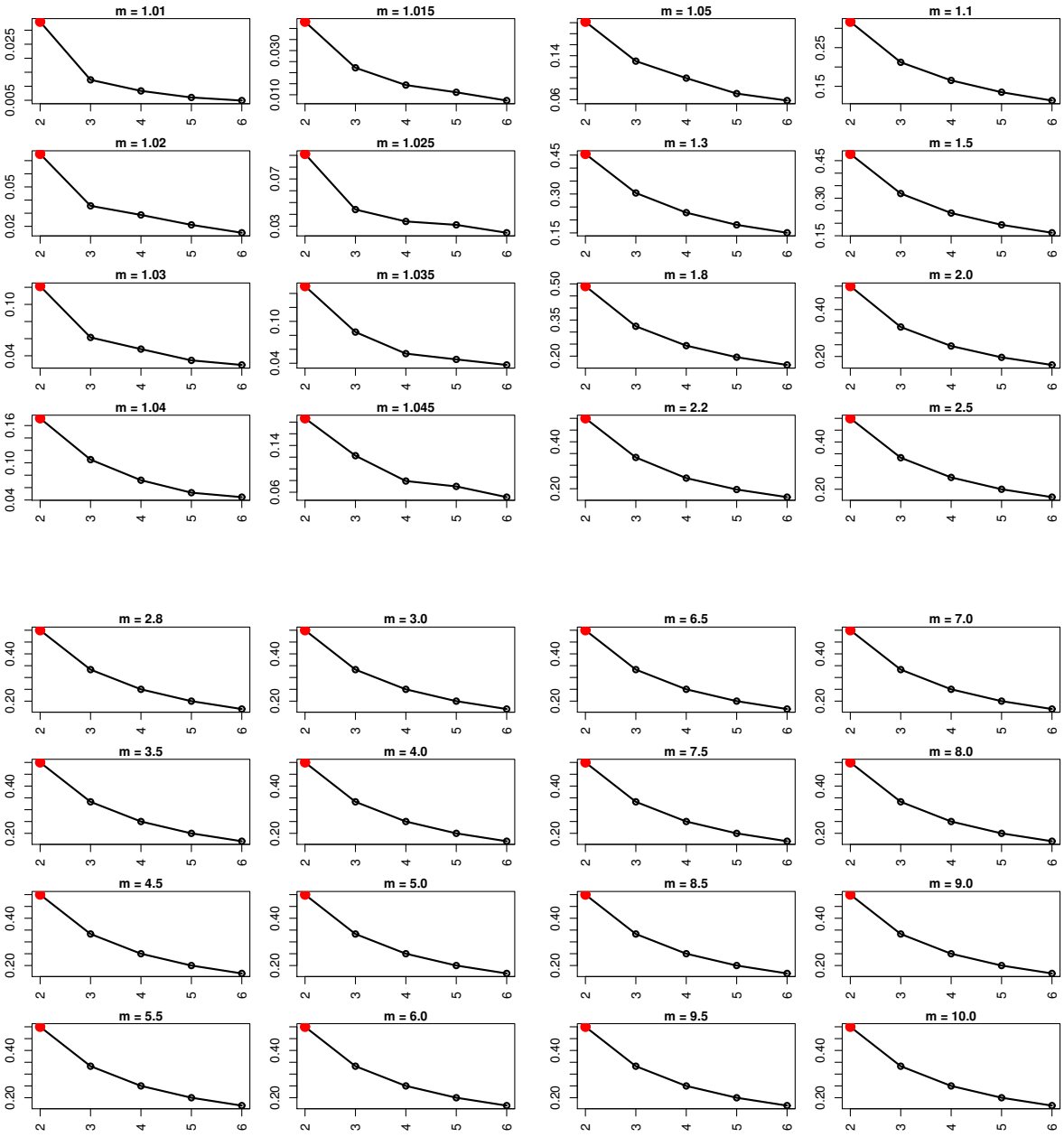


Tabela 142 – Hitech

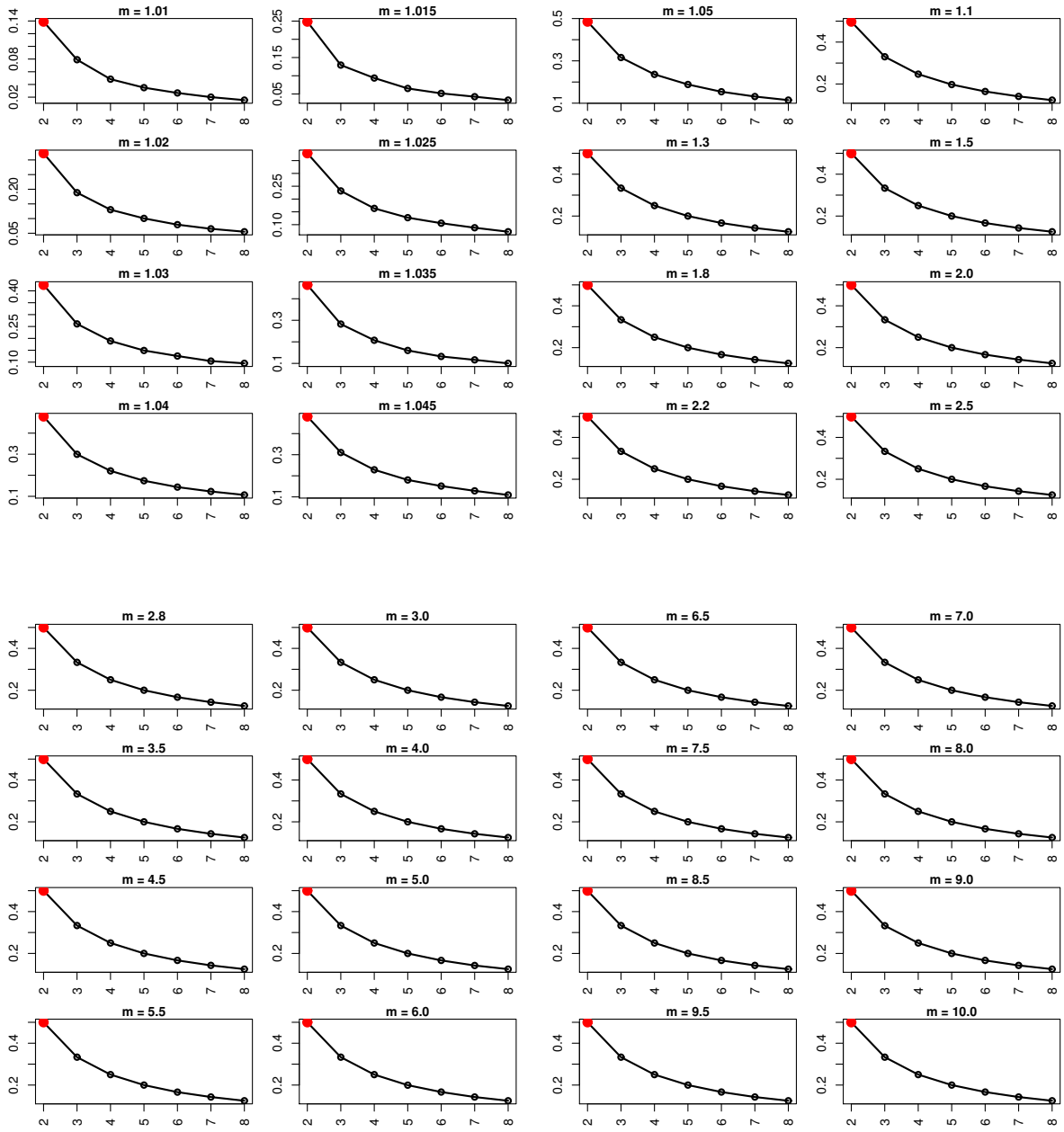


Tabela 143 – WAP

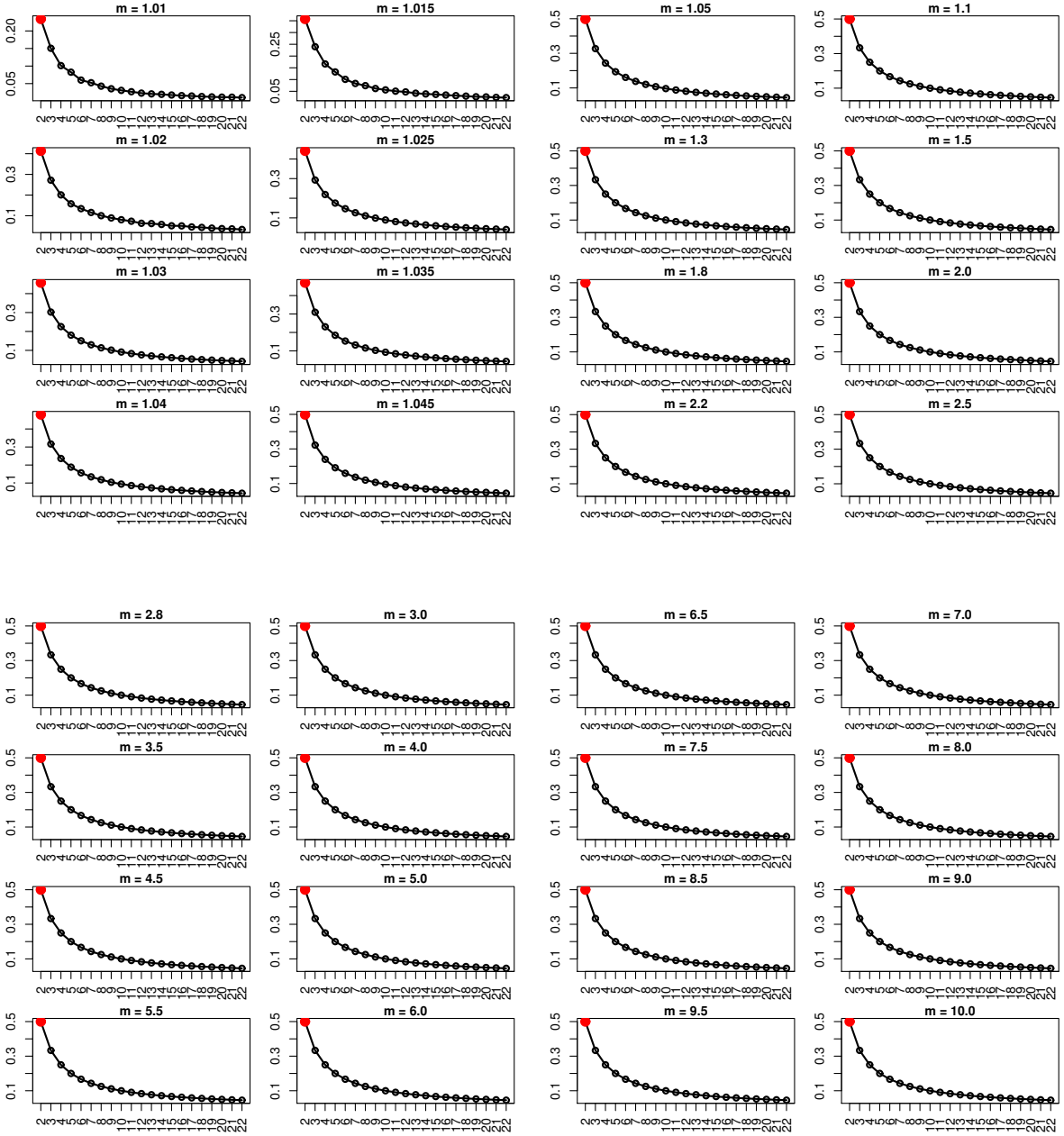


Tabela 144 – NSF

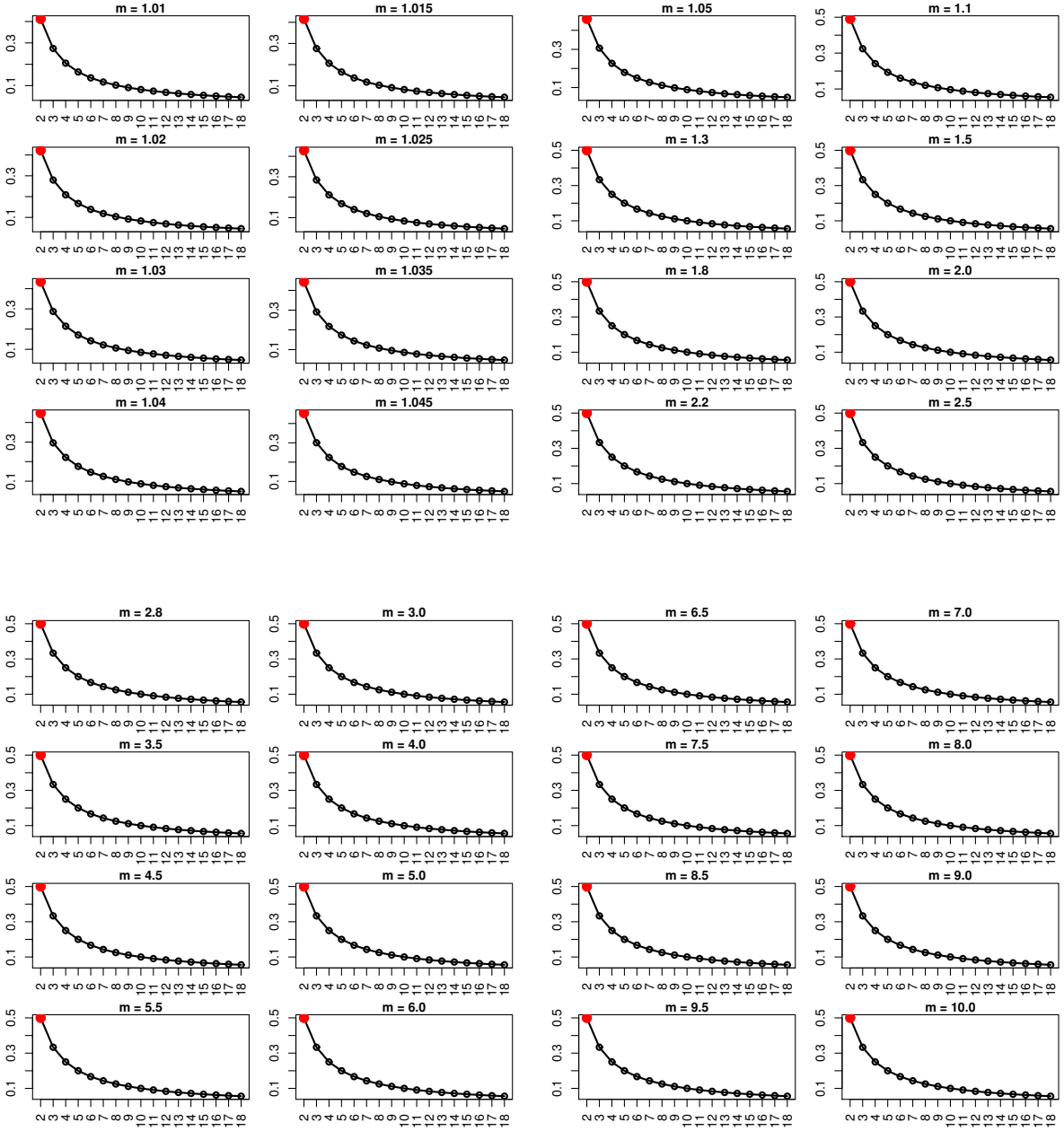


Tabela 145 – Irish-Sentiment

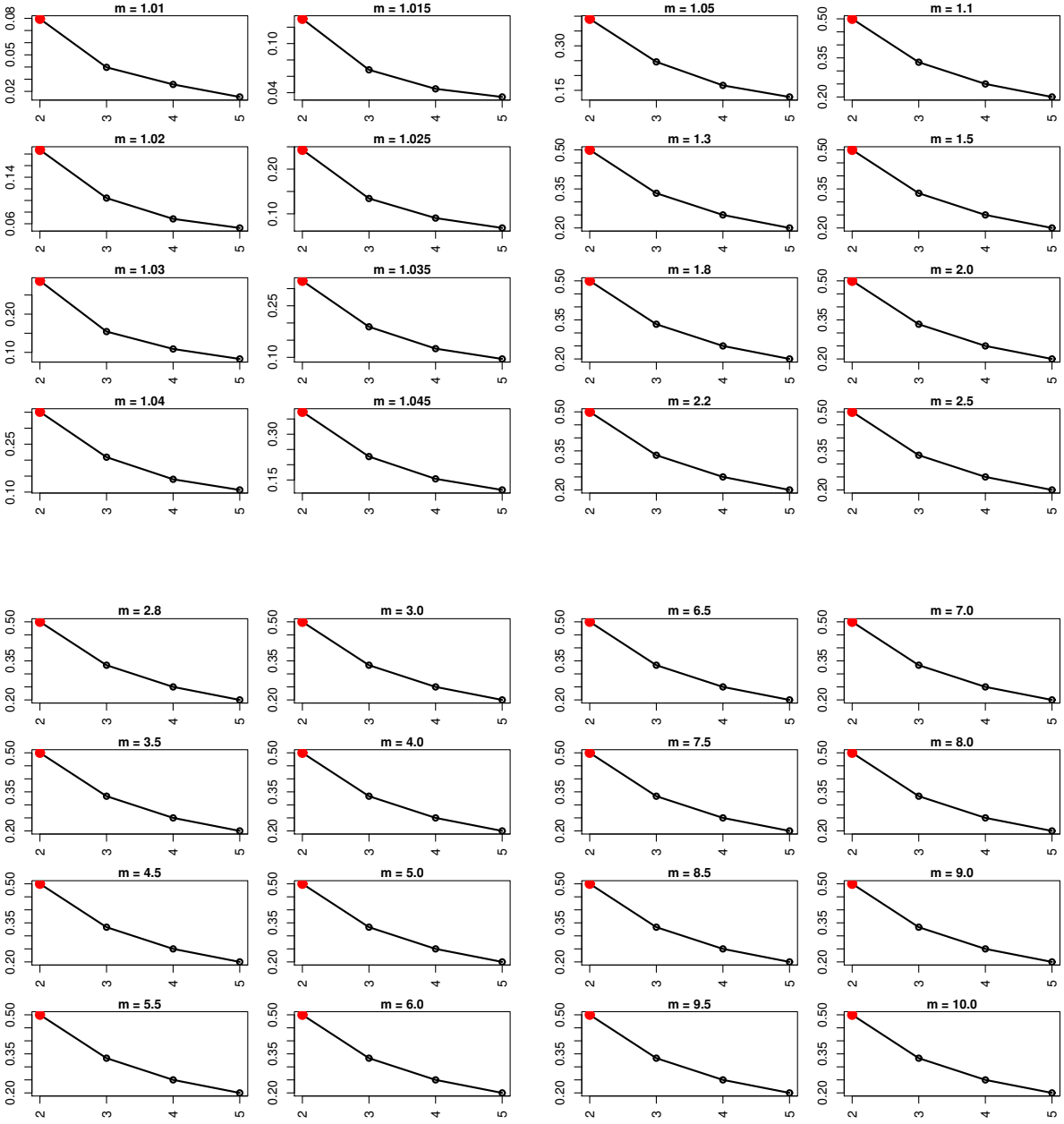


Tabela 146 – 20Newsgroups

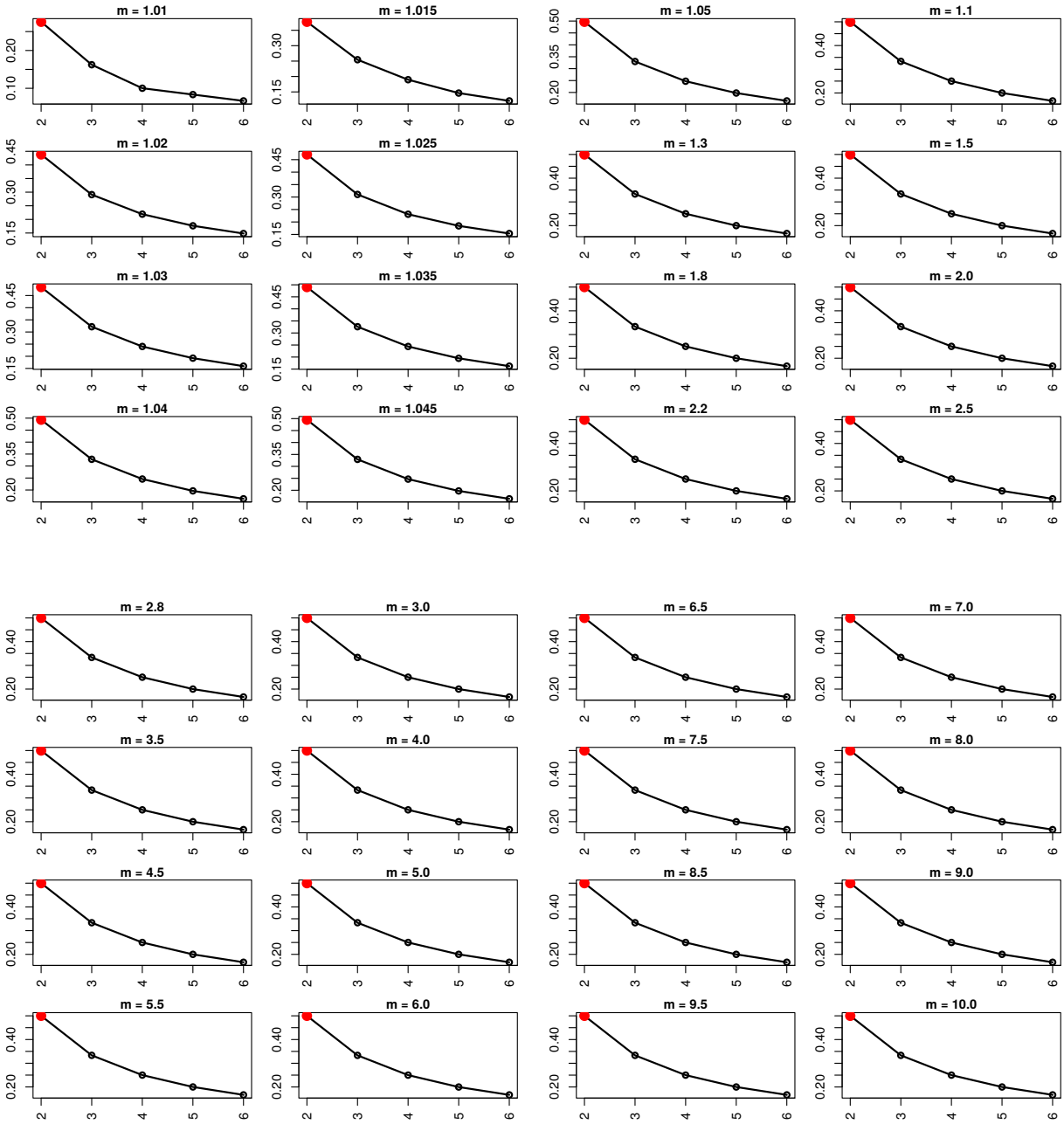


Tabela 147 – La1s

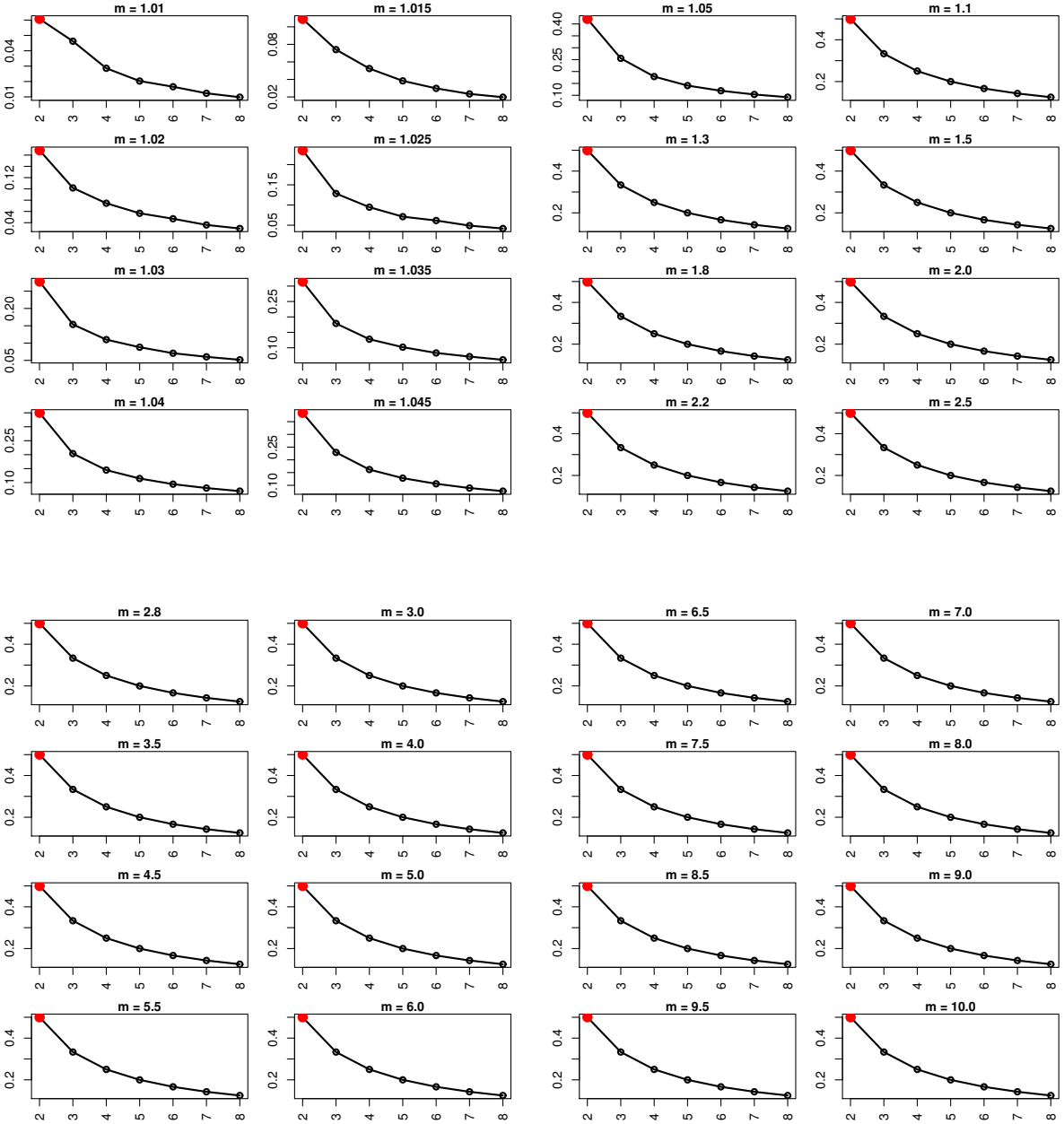
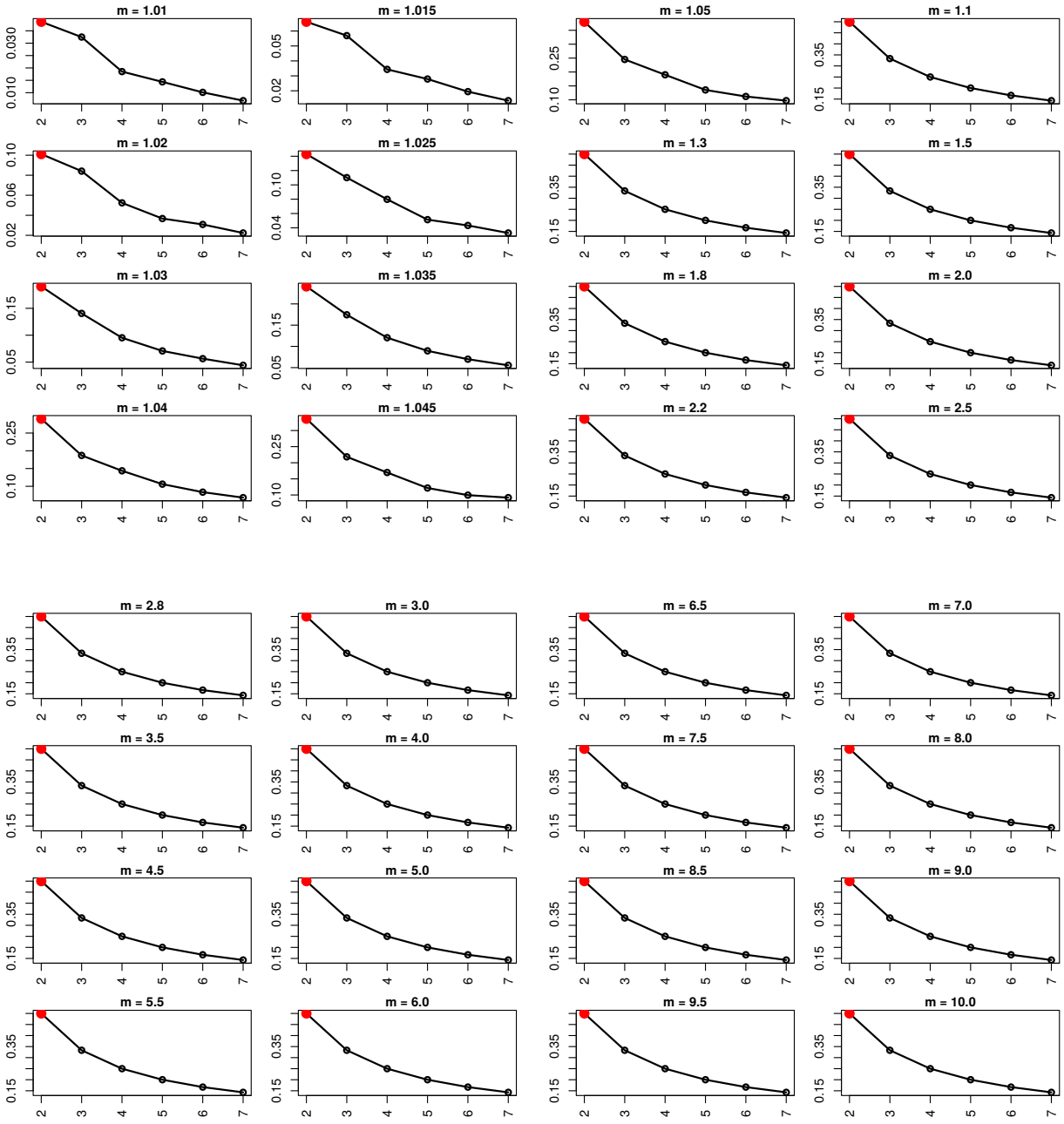


Tabela 148 – Reviews



ANEXO M – PBMF

Tabela 149 – NewYorkTimes

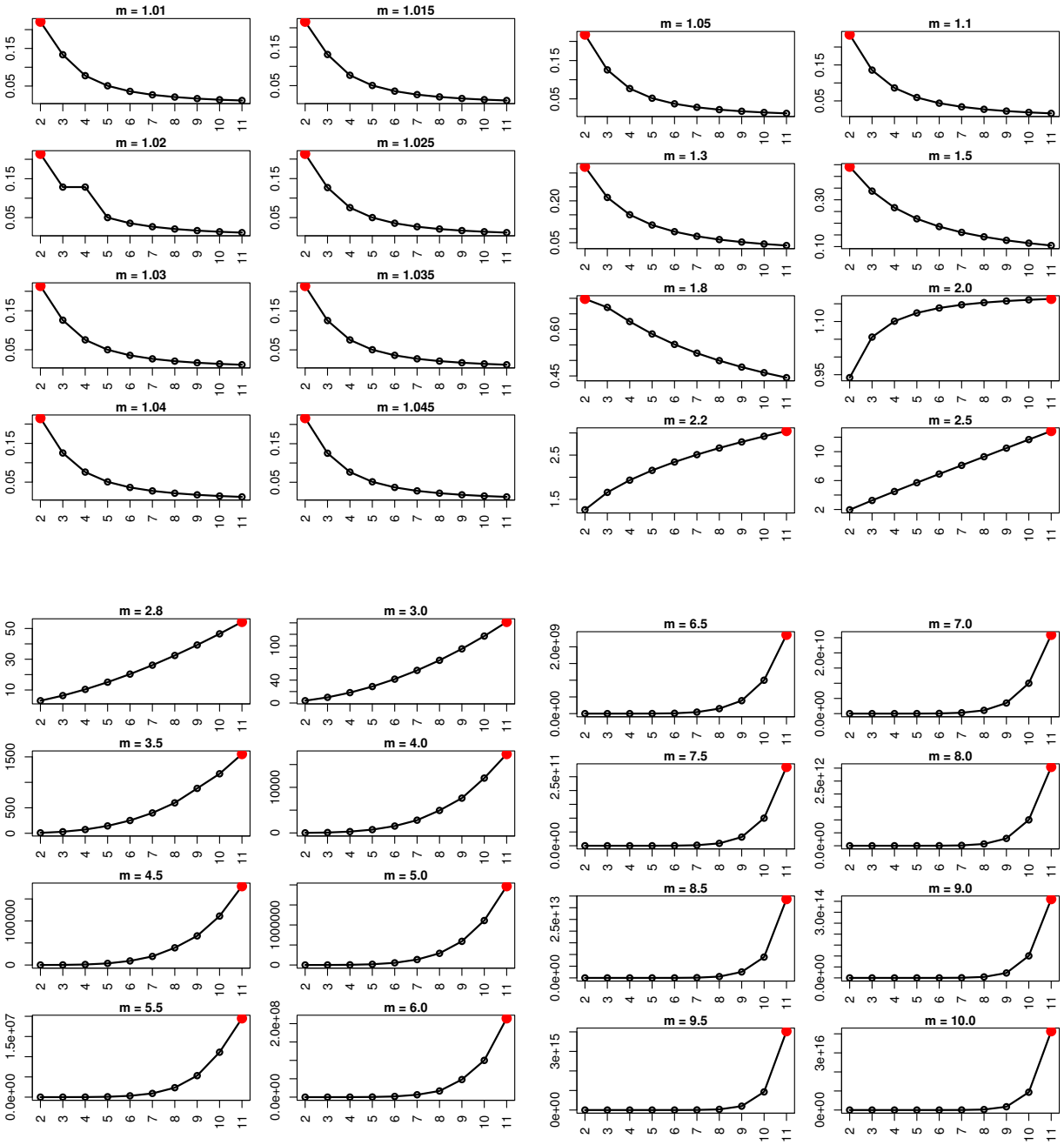


Tabela 150 – IAarticles

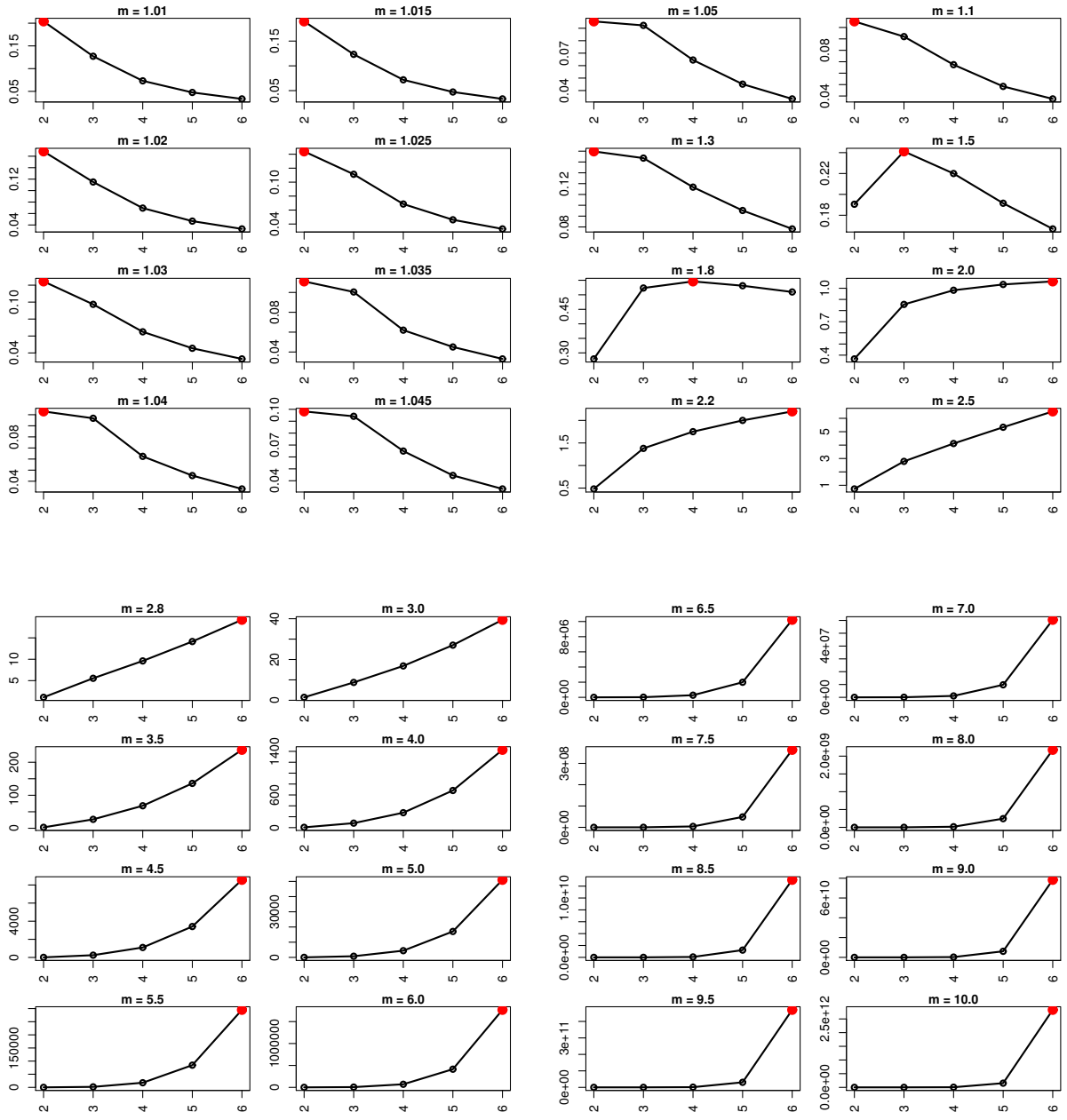


Tabela 151 – Opinosis

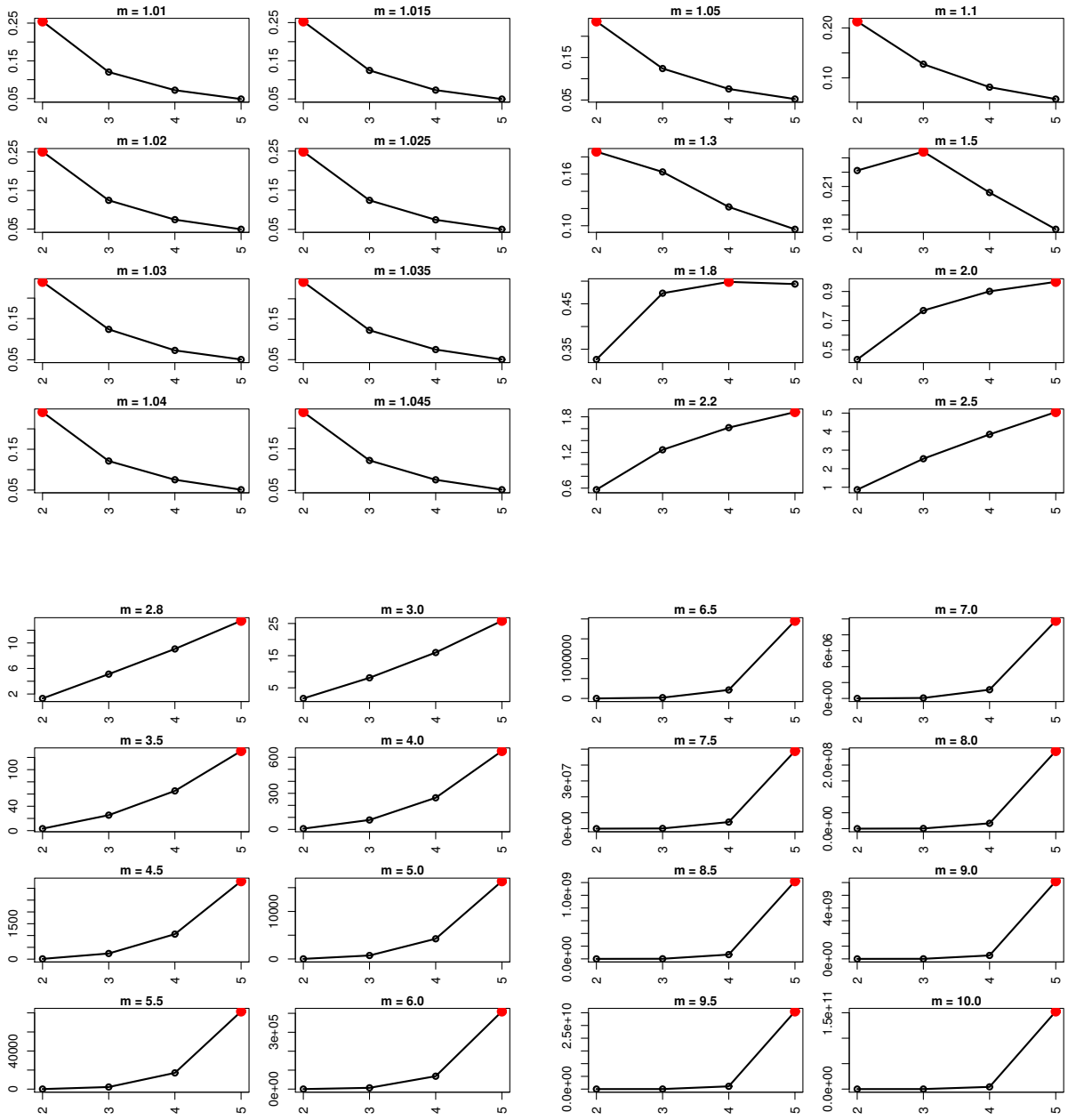


Tabela 152 – CSTR

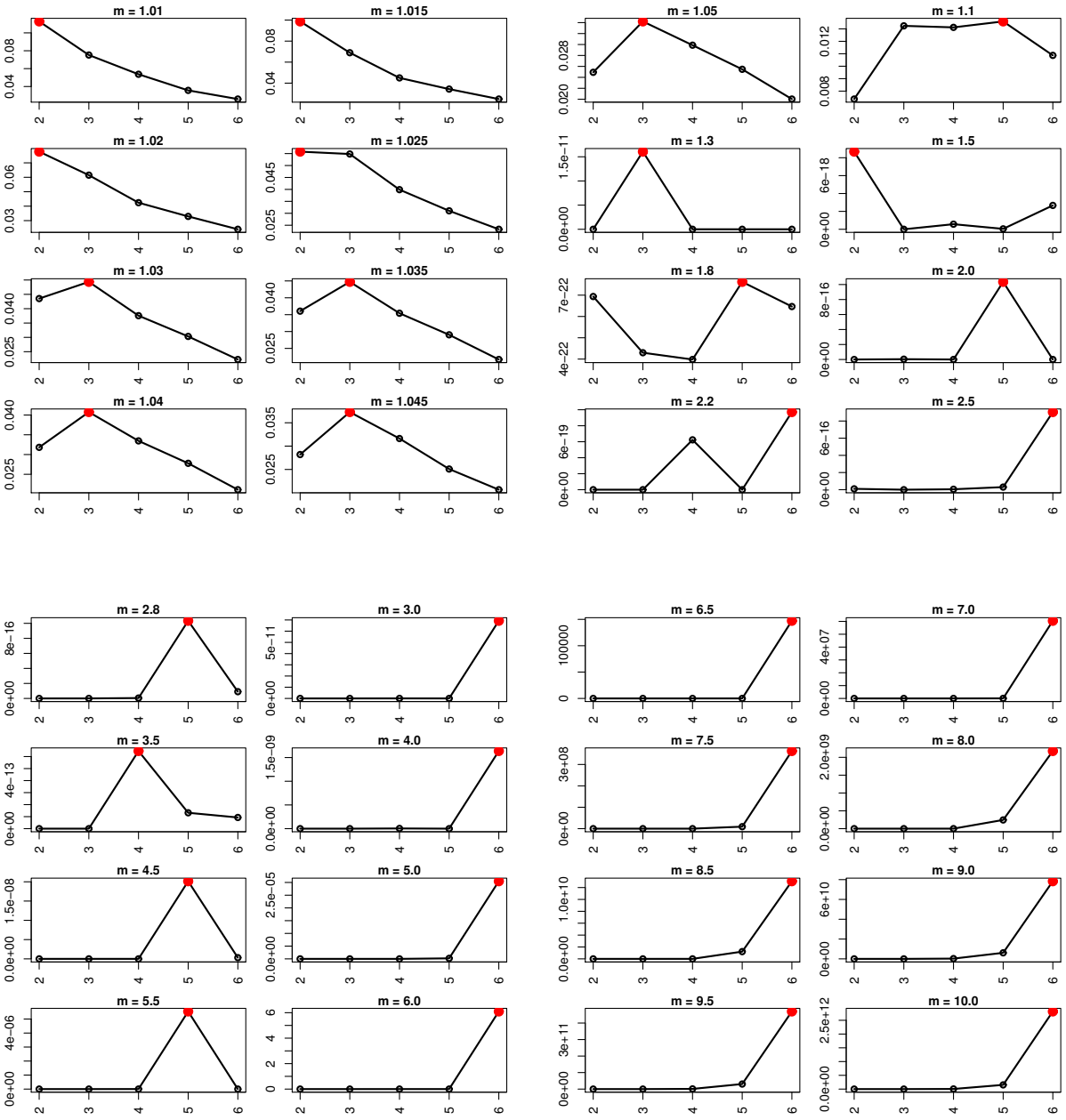


Tabela 153 – SyskillWebert

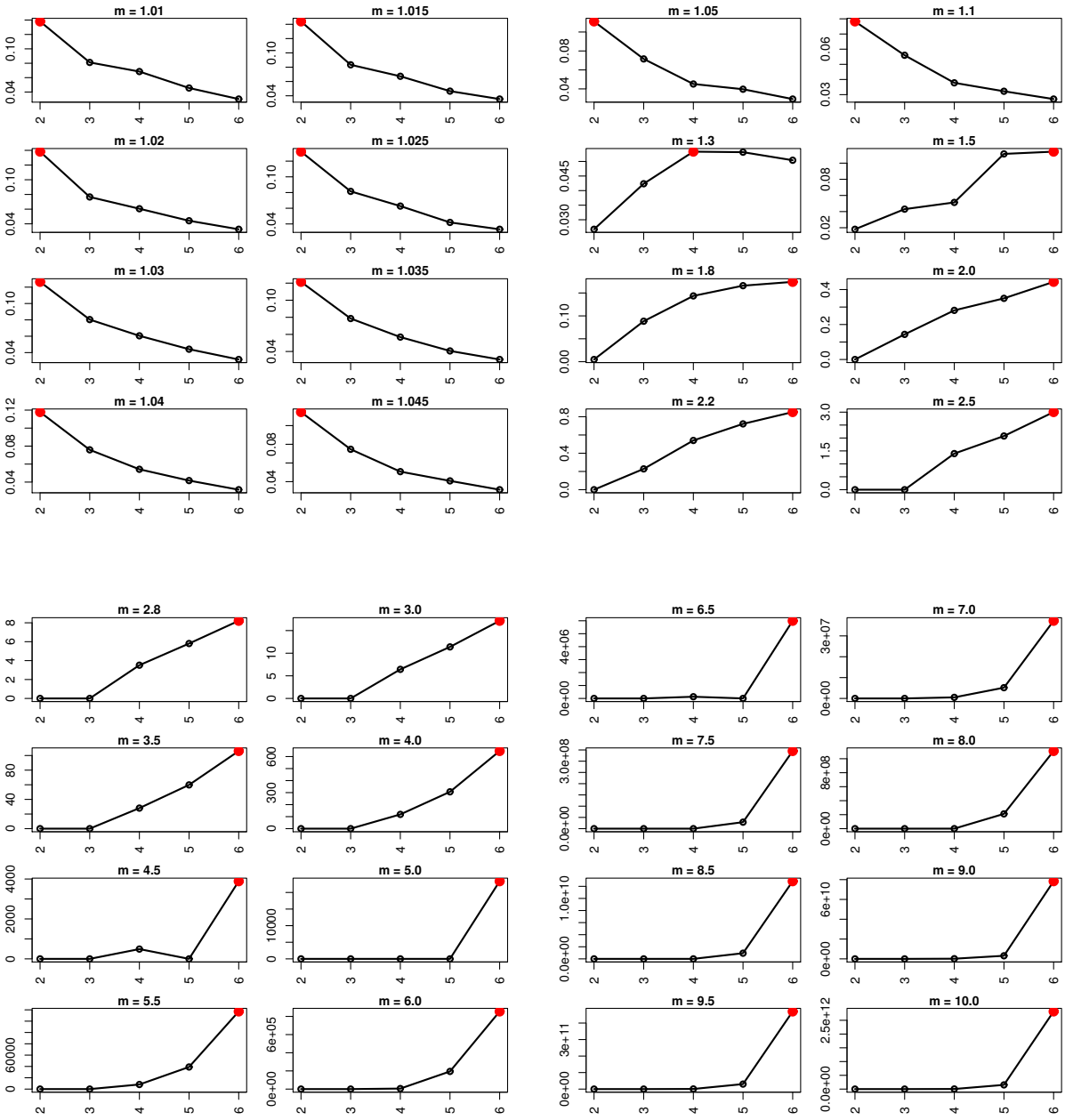


Tabela 154 – Hitech

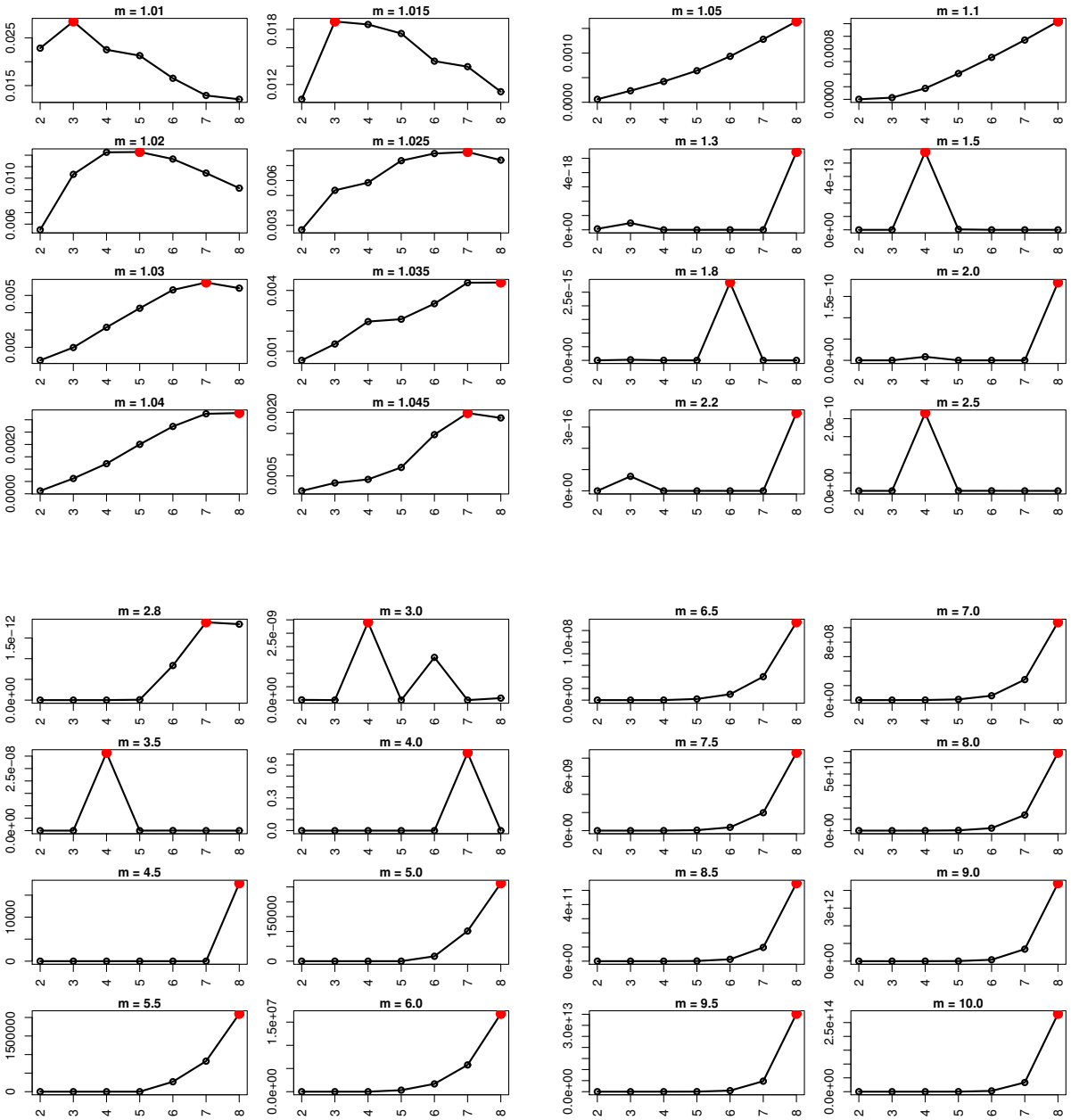


Tabela 155 – WAP

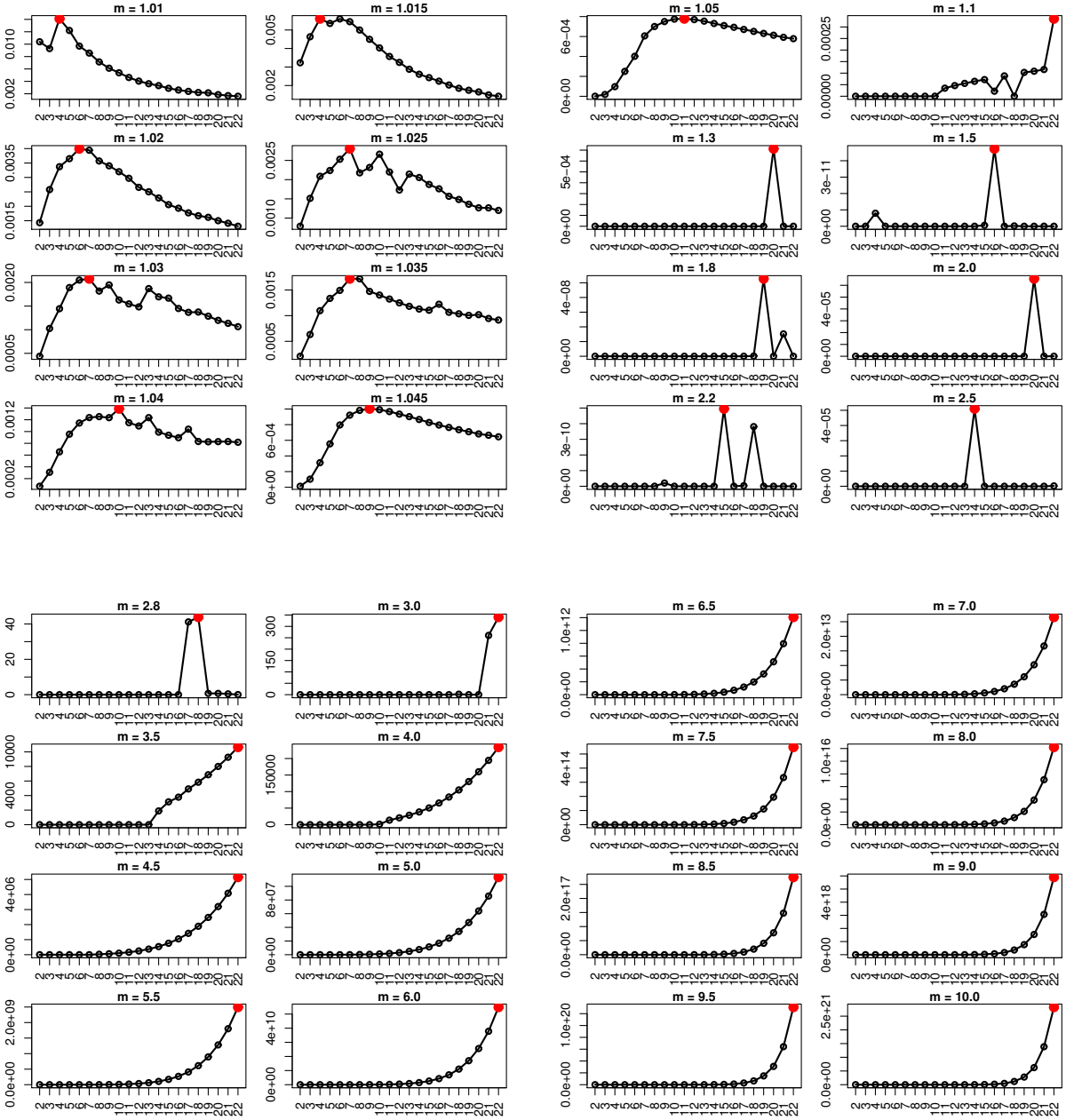


Tabela 156 – NSF

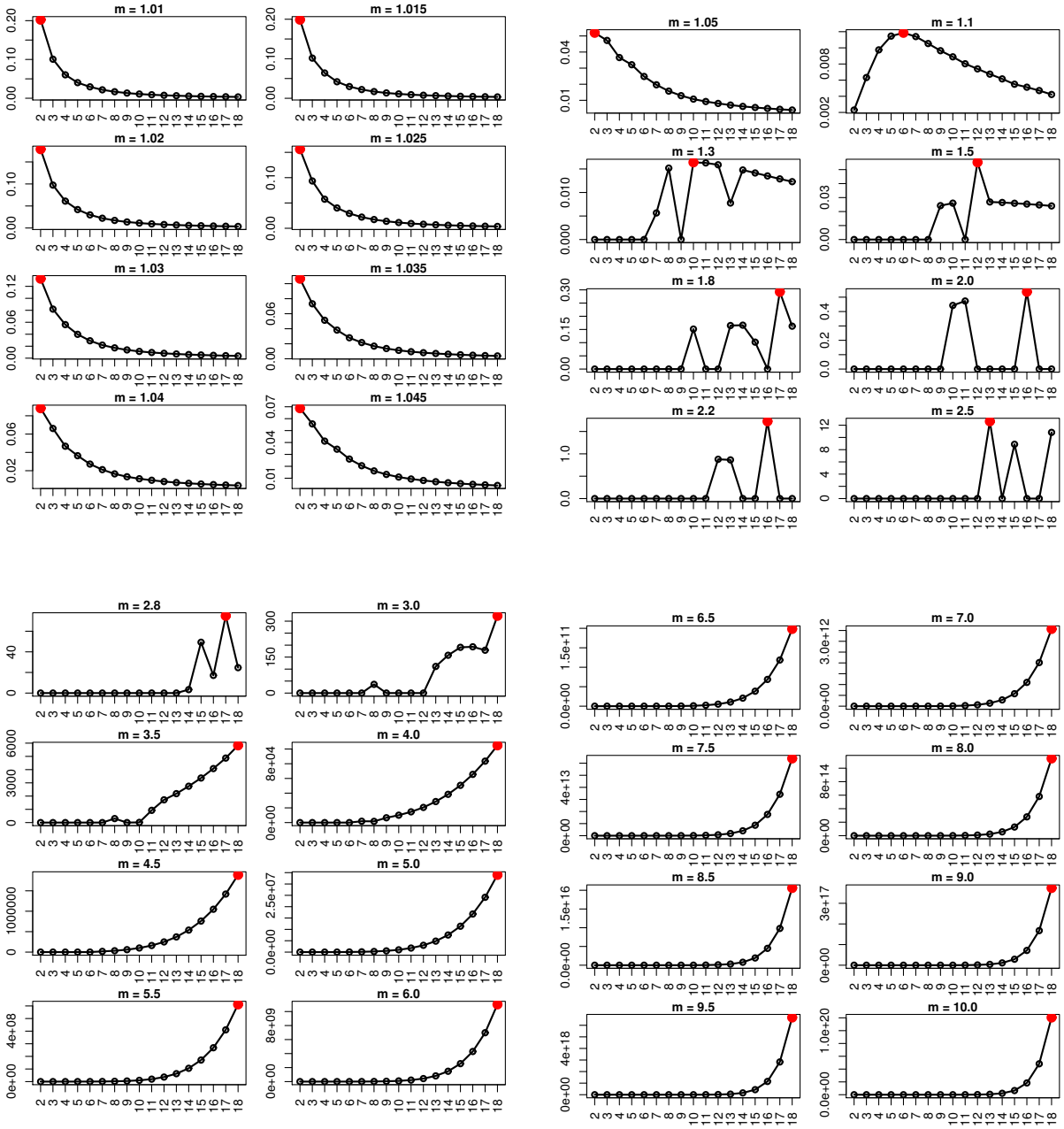


Tabela 157 – Irish-Sentiment

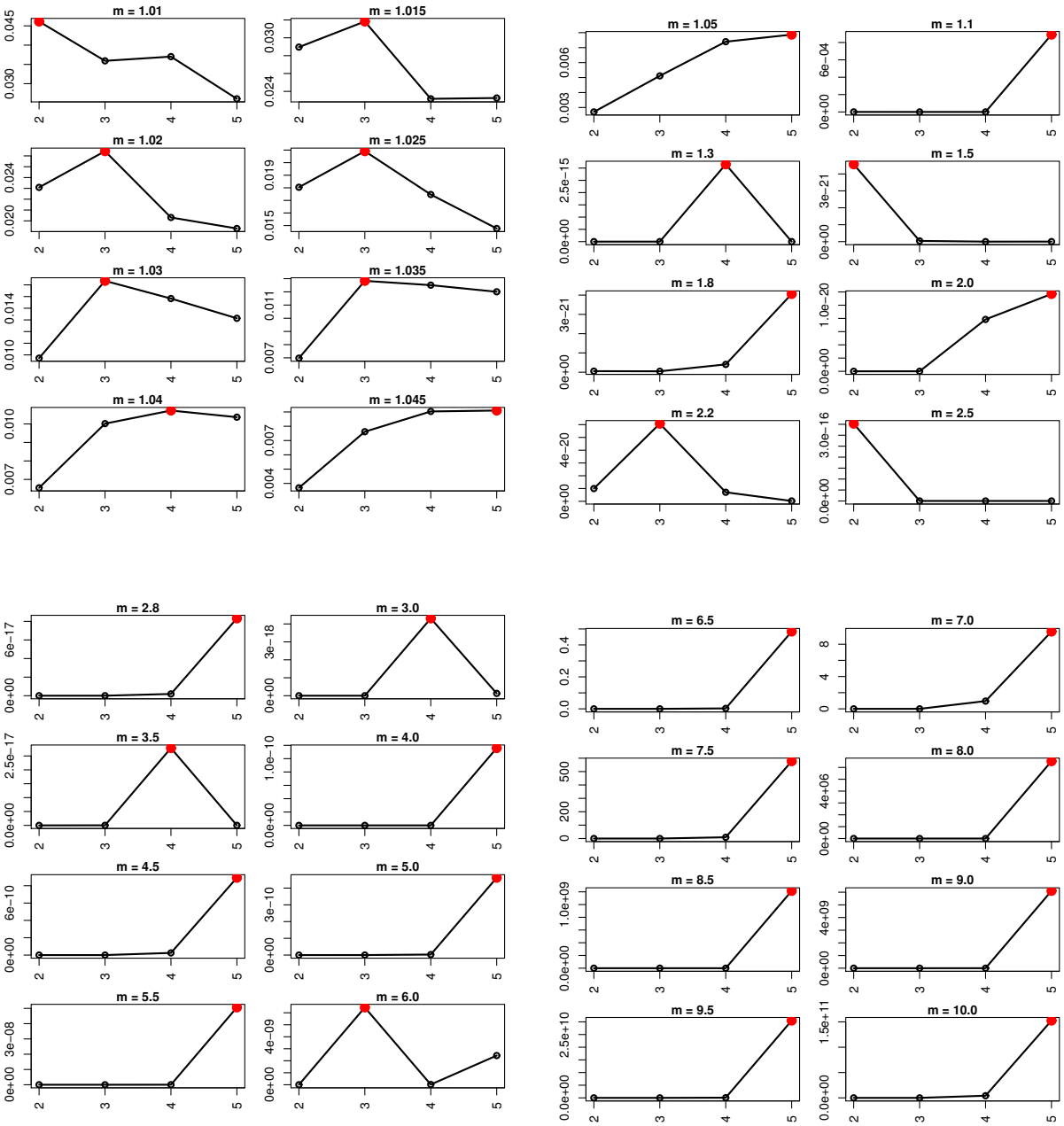


Tabela 158 – 20Newsgroups

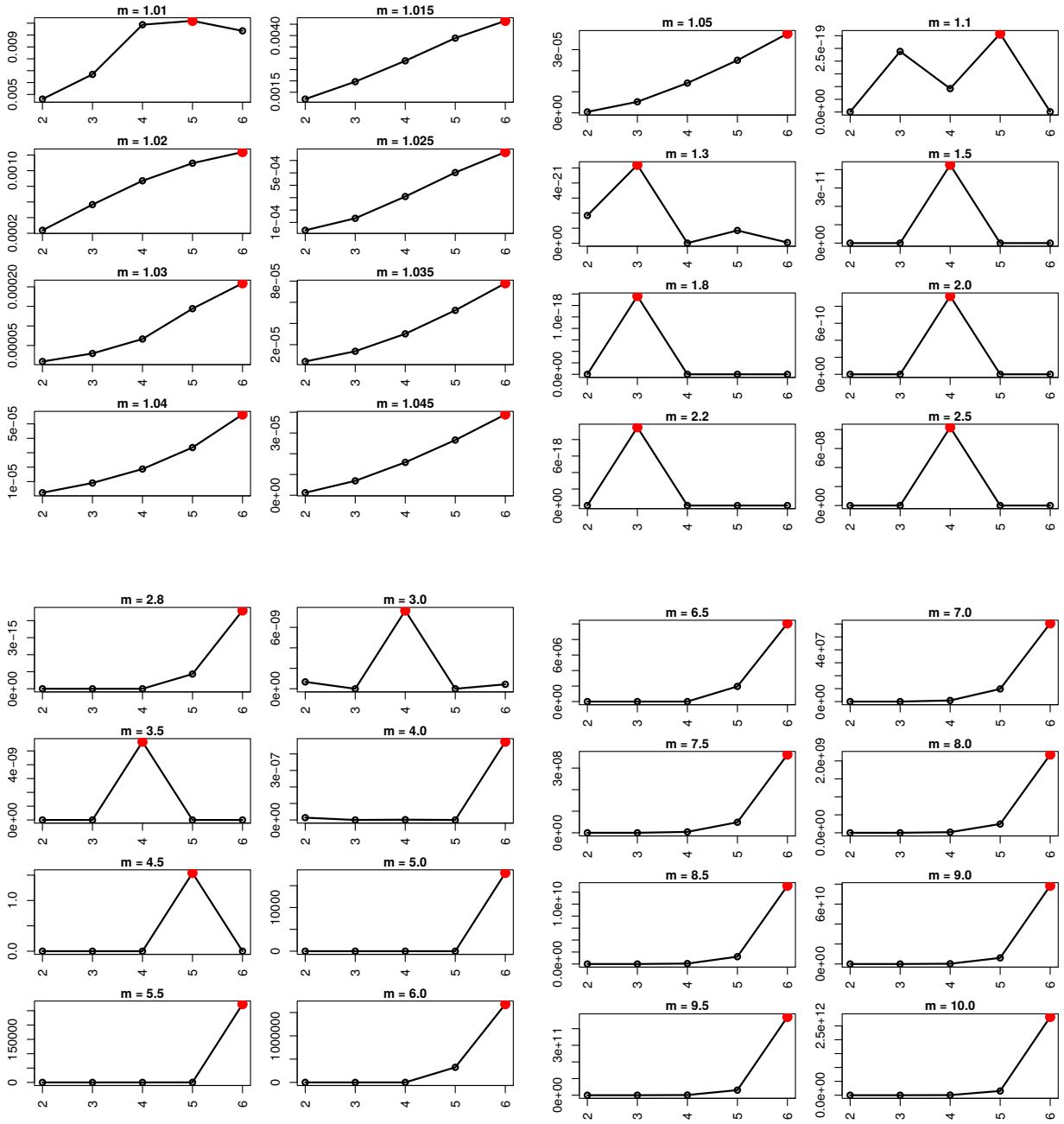


Tabela 159 – La1s

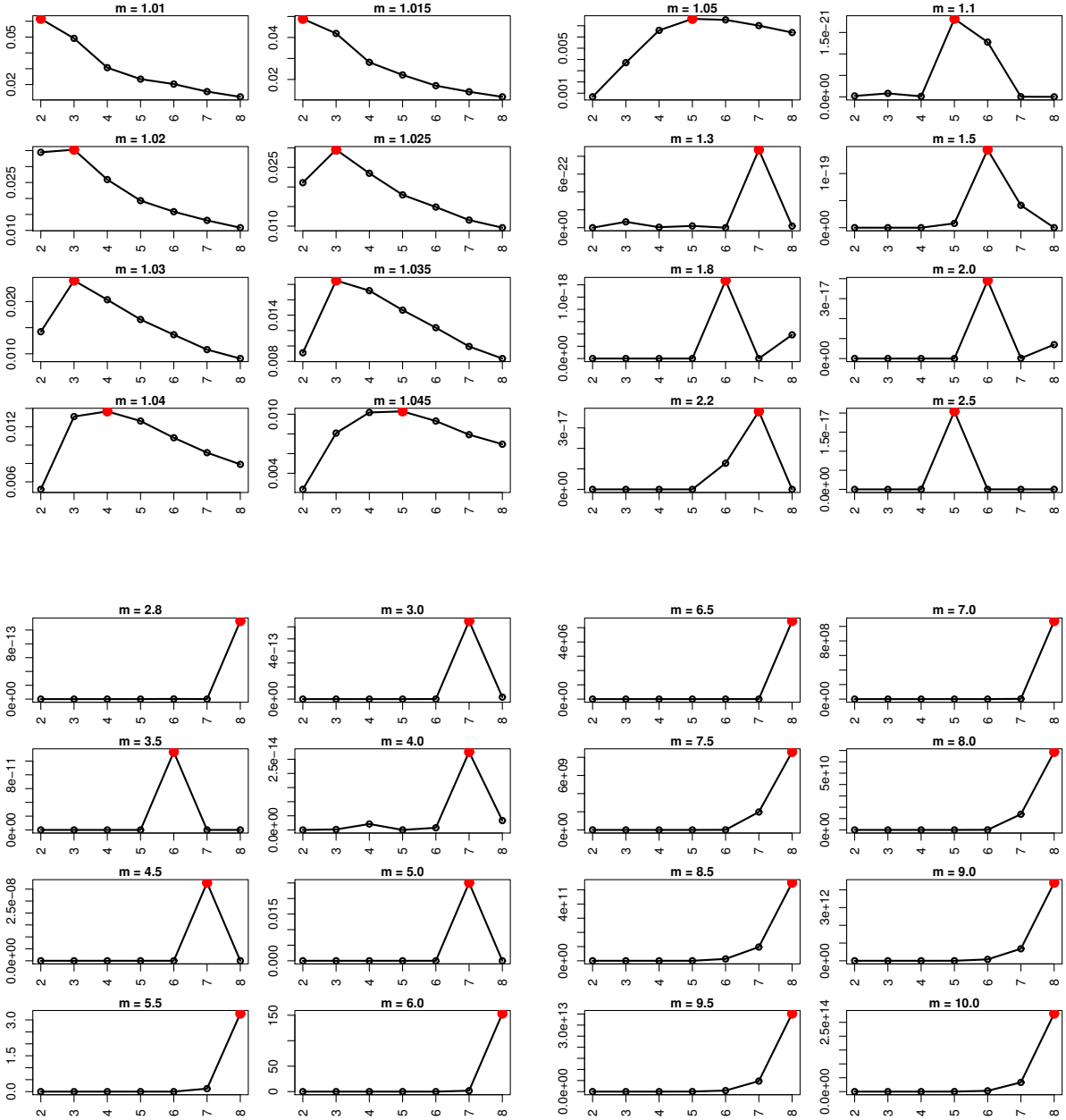
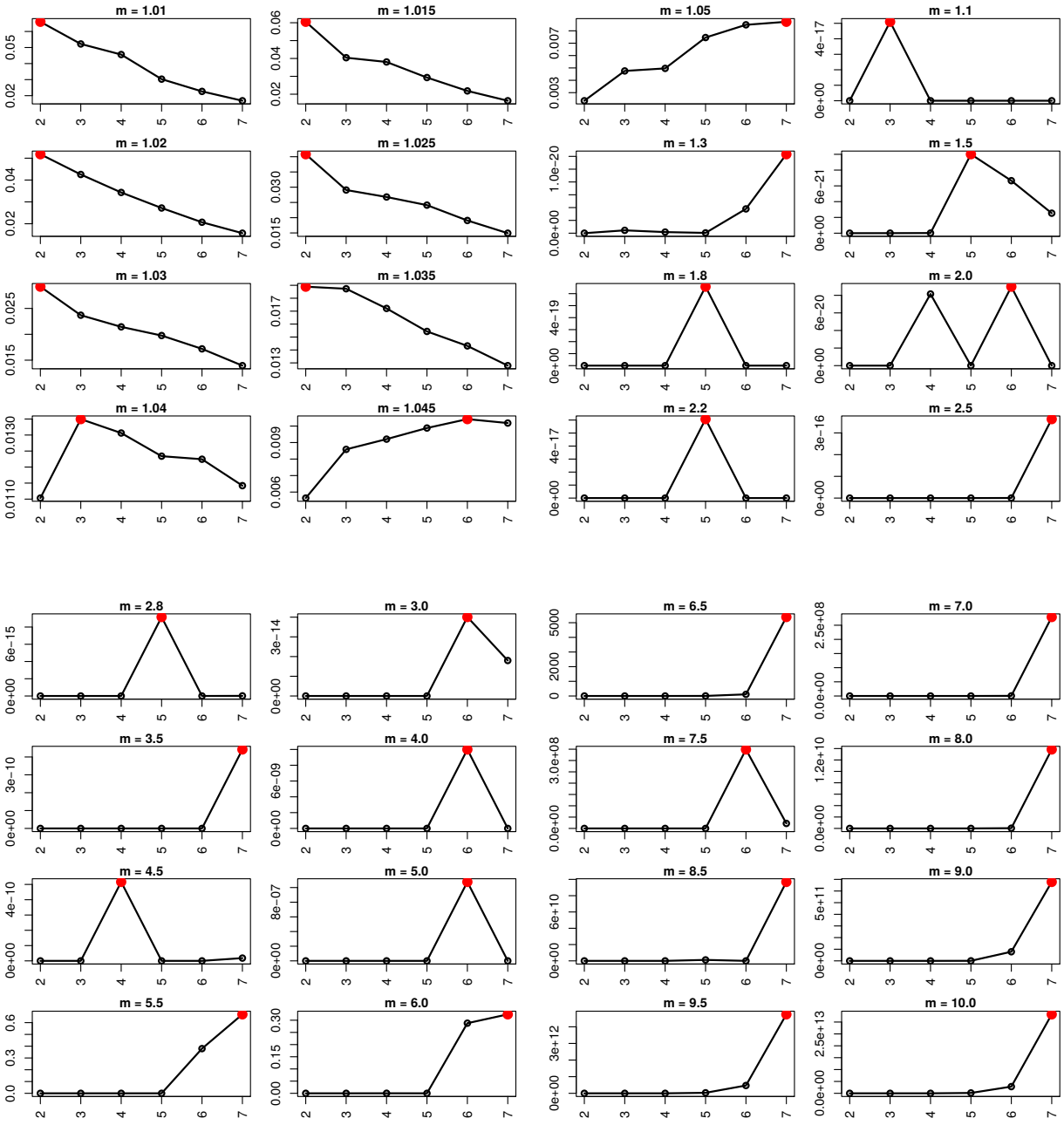


Tabela 160 – Reviews



ANEXO N – PCAES

Tabela 161 – NewYorkTimes

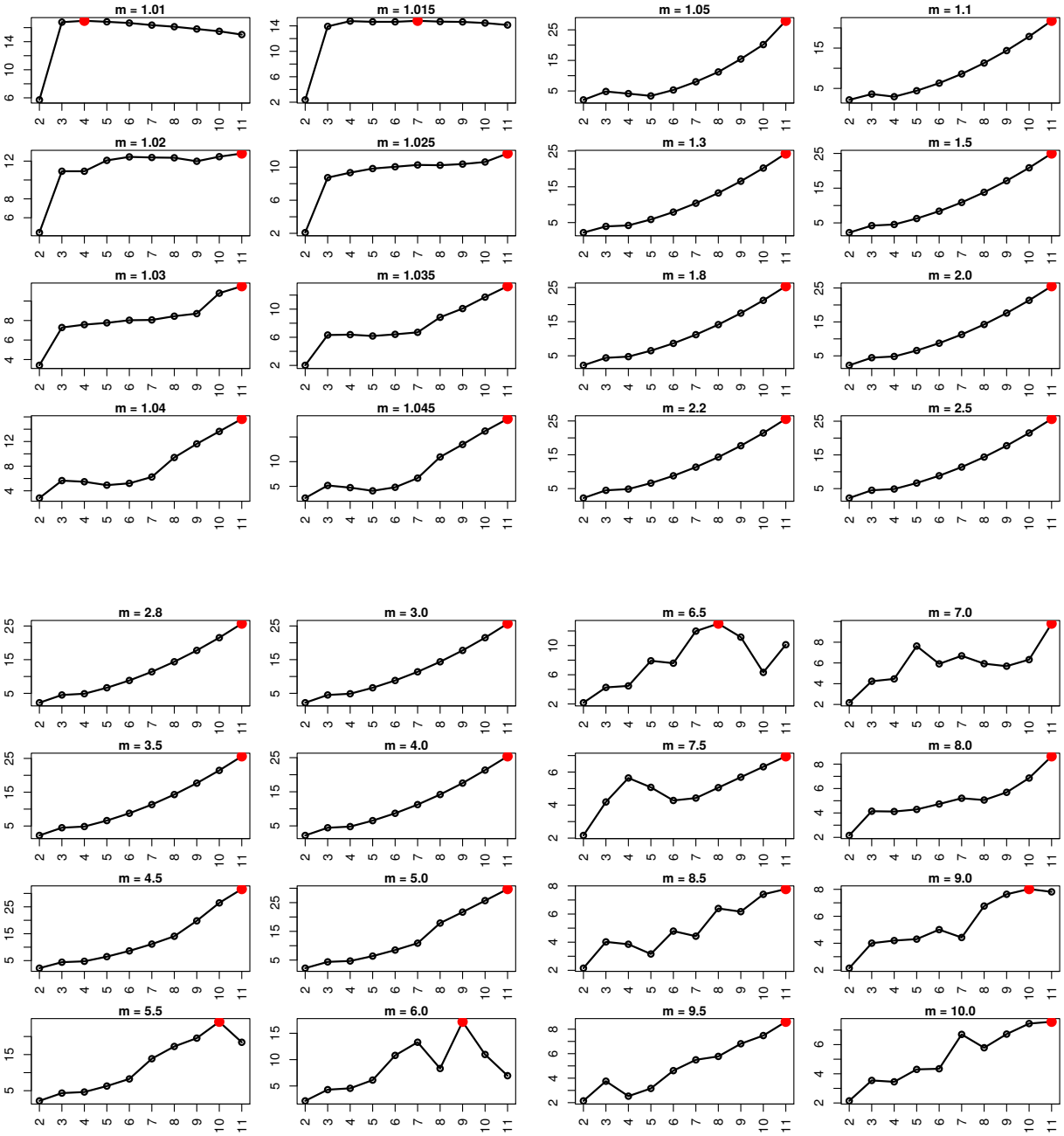


Tabela 162 – IAarticles

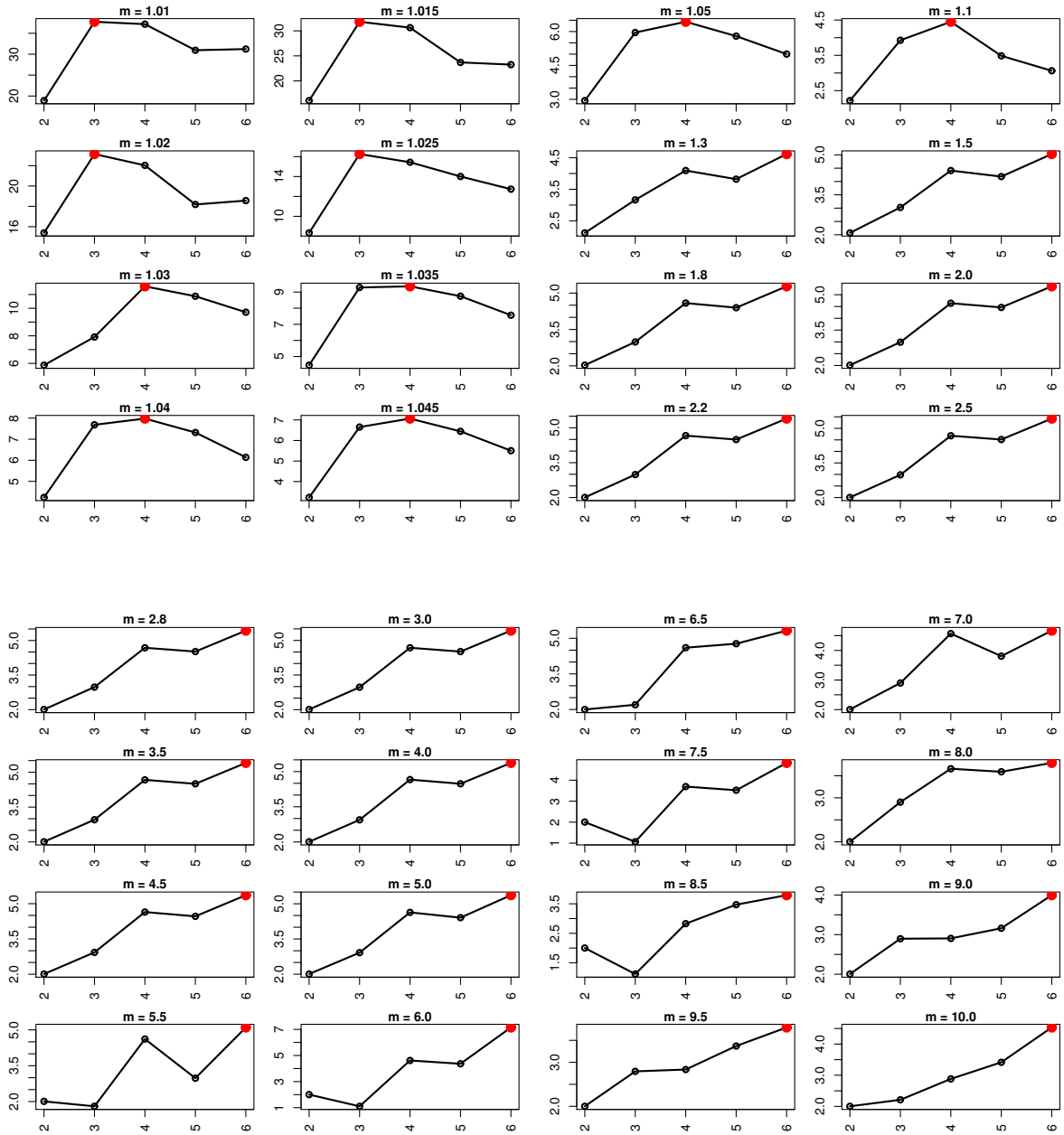


Tabela 163 – Opinosis

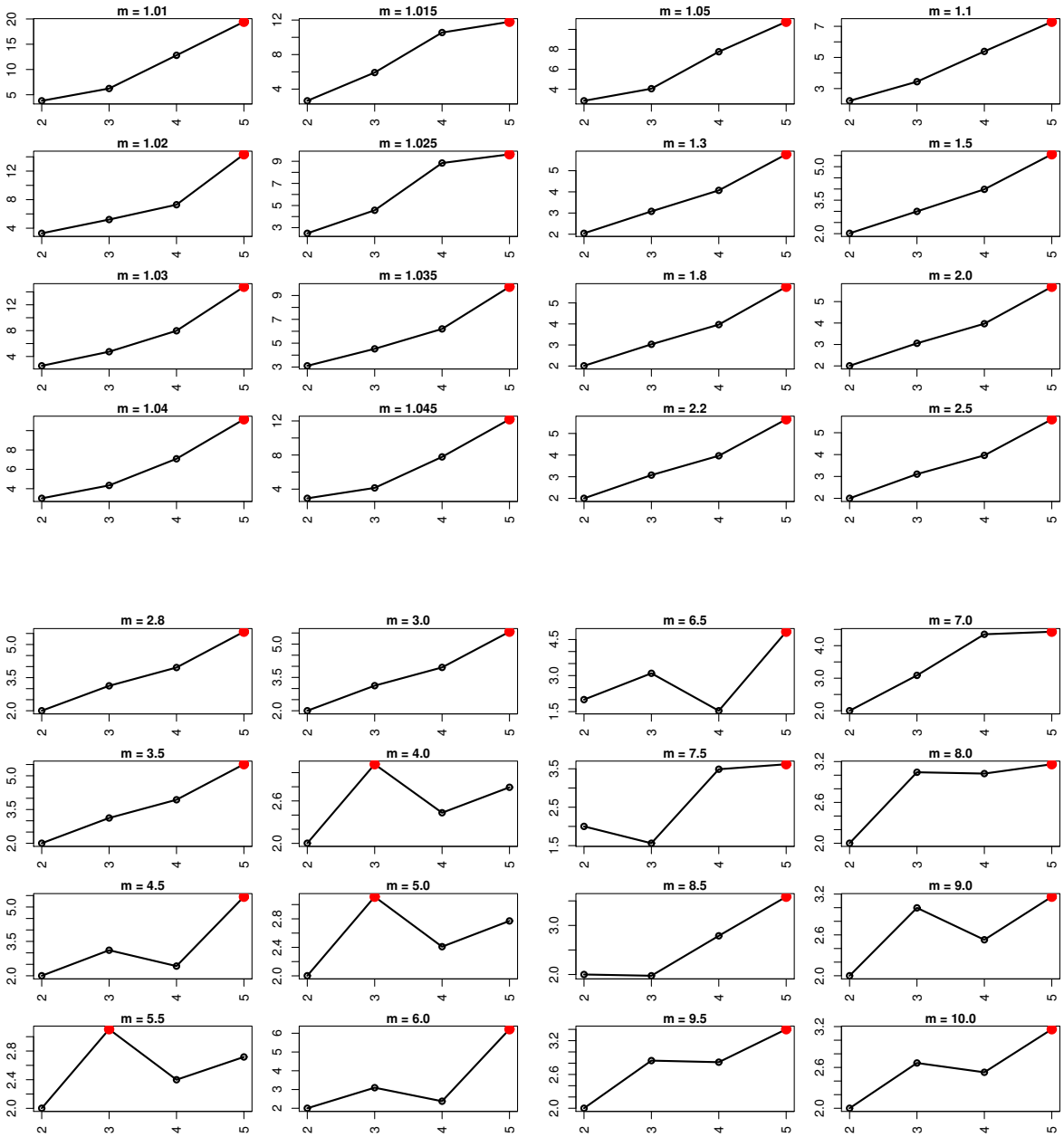


Tabela 164 – CSTR

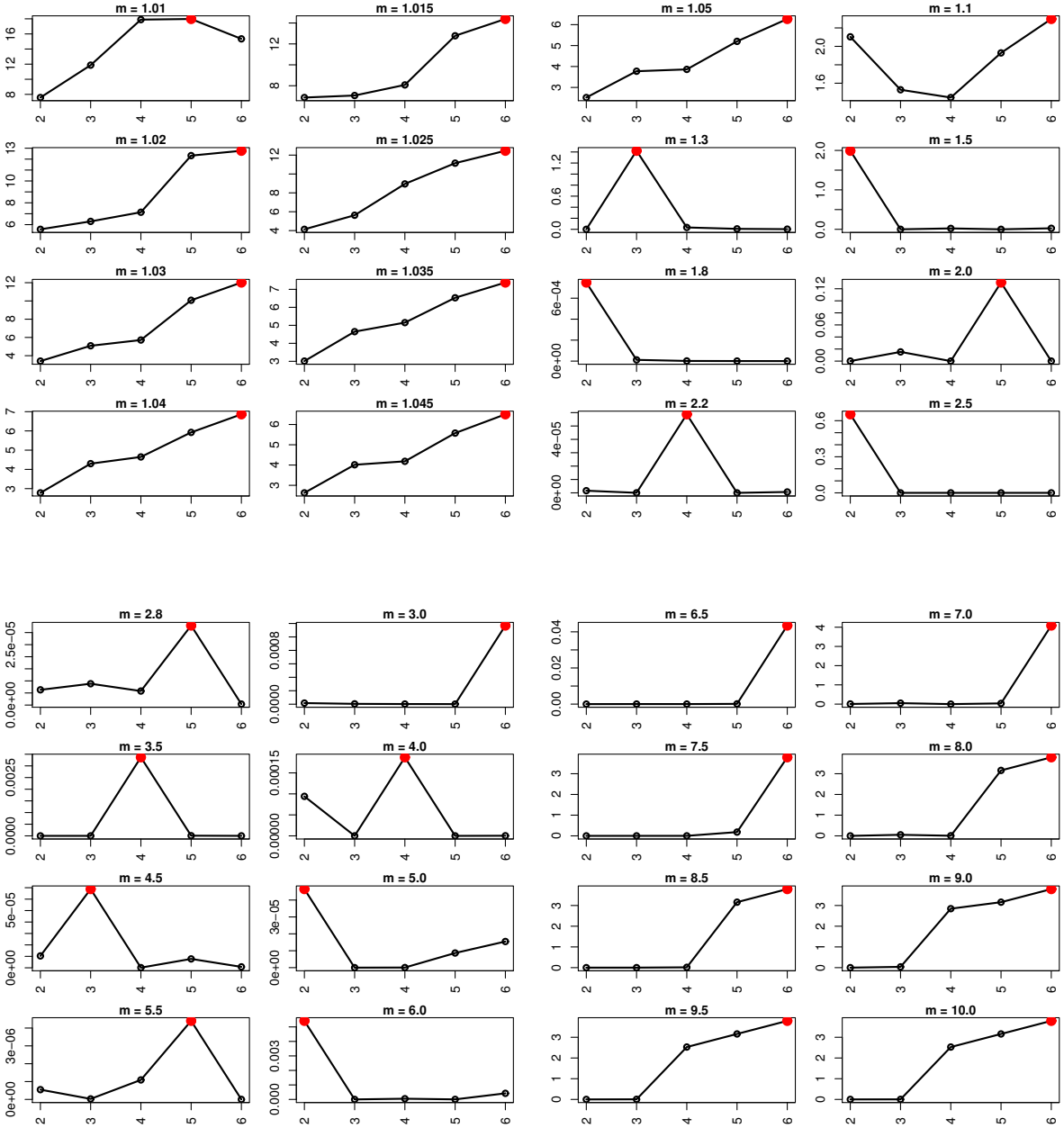


Tabela 165 – SyskillWilbert

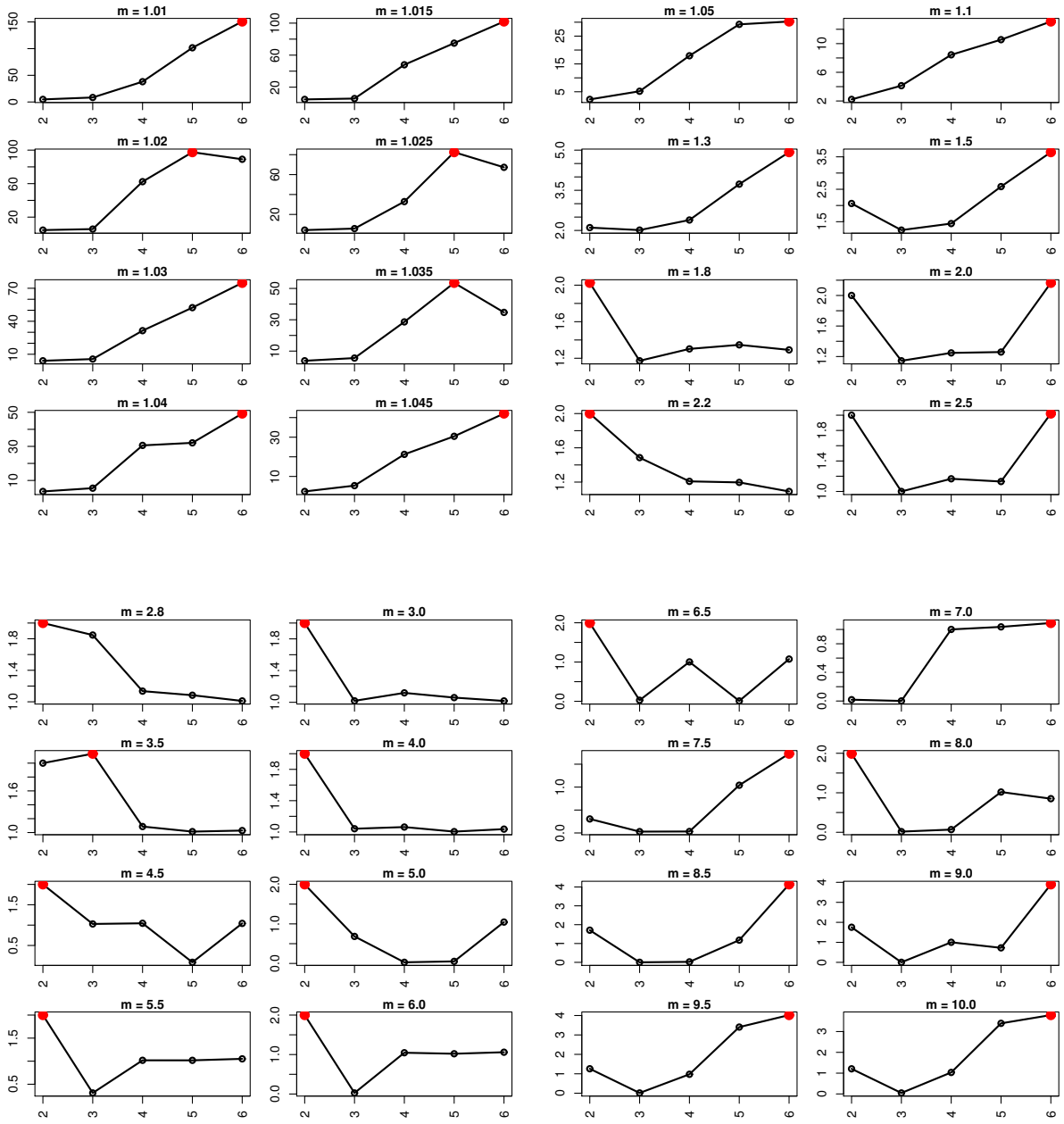


Tabela 166 – Hitech

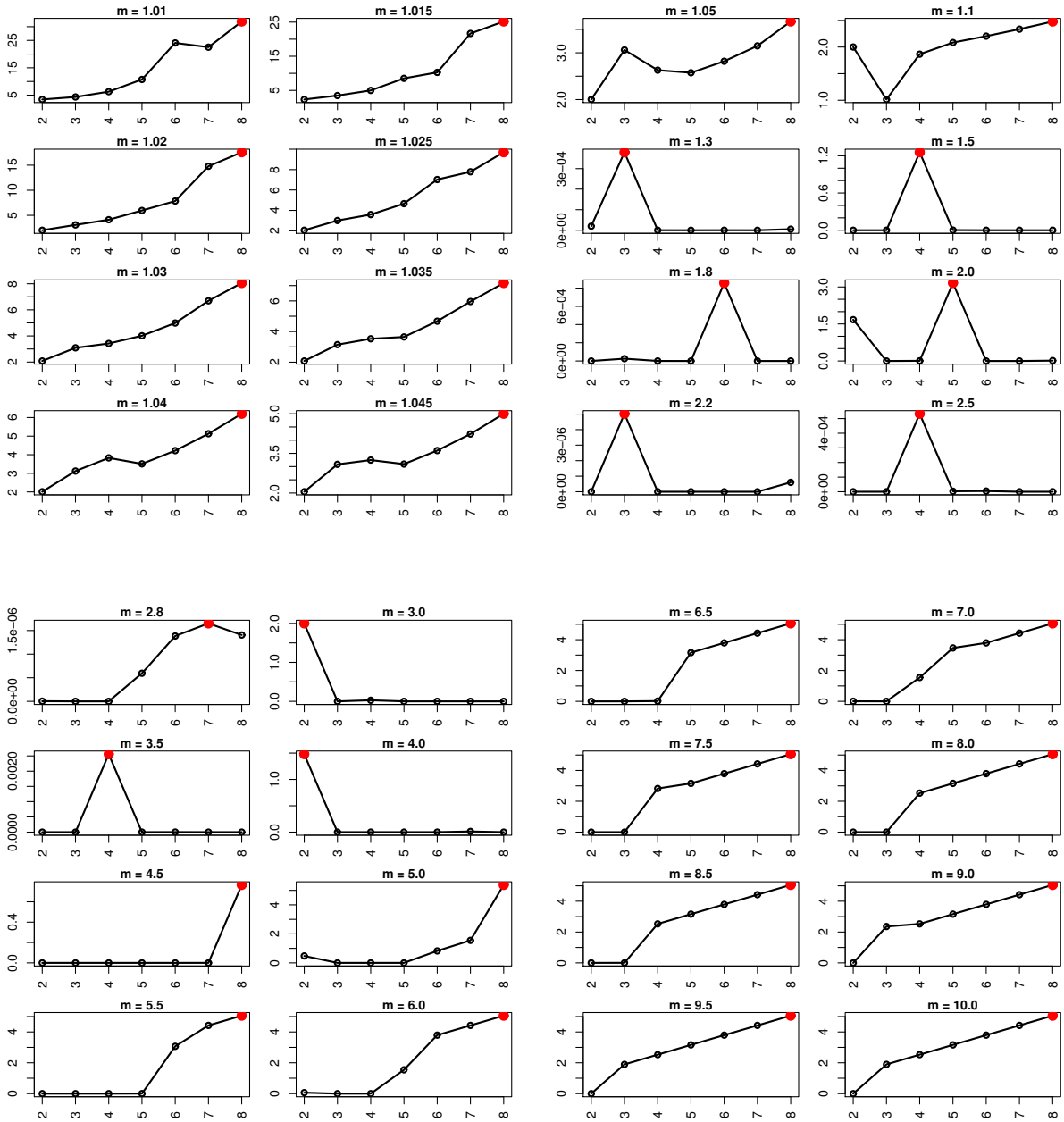


Tabela 167 – WAP

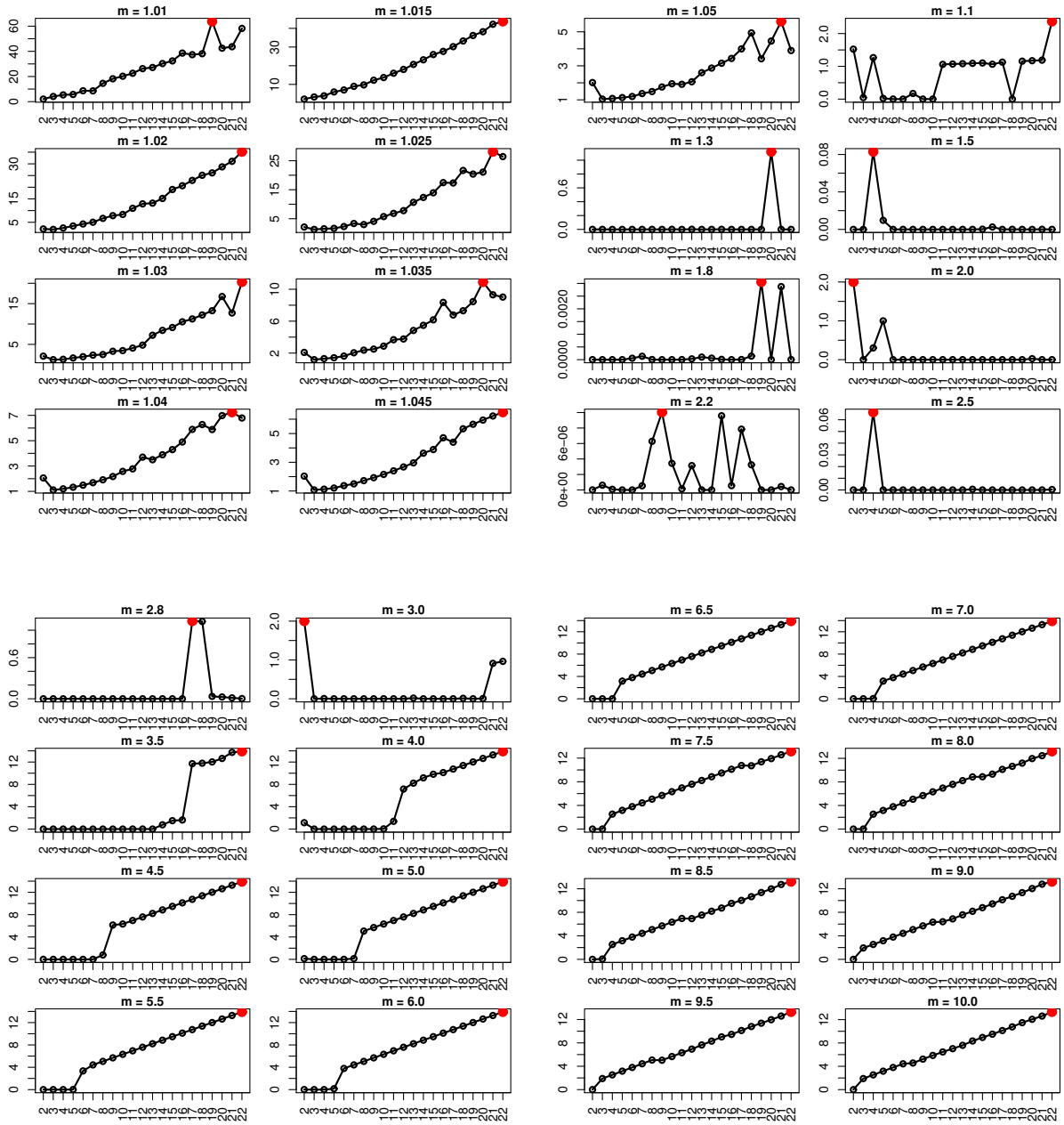


Tabela 168 – NSF

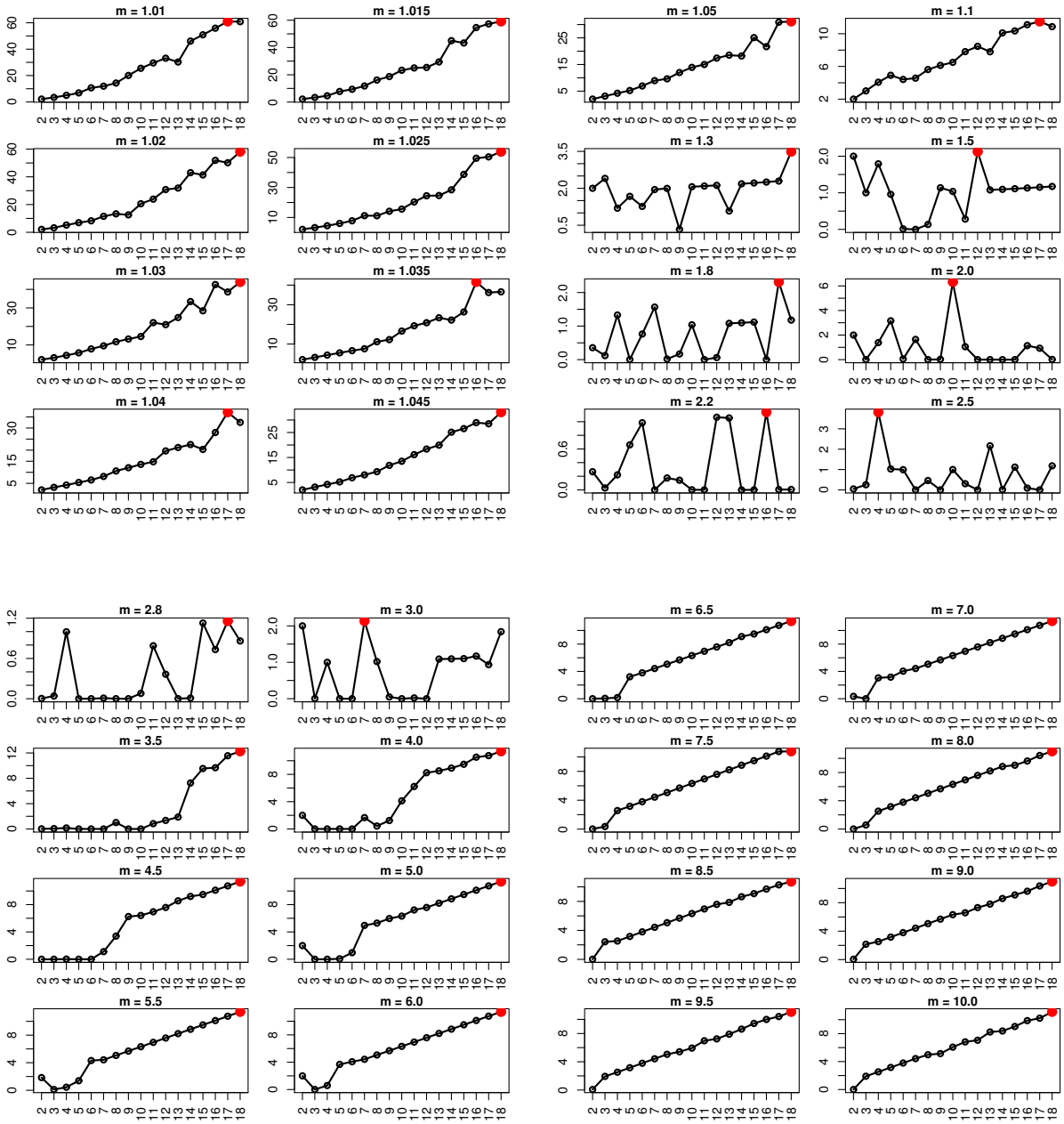


Tabela 169 – Irish-Sentiment

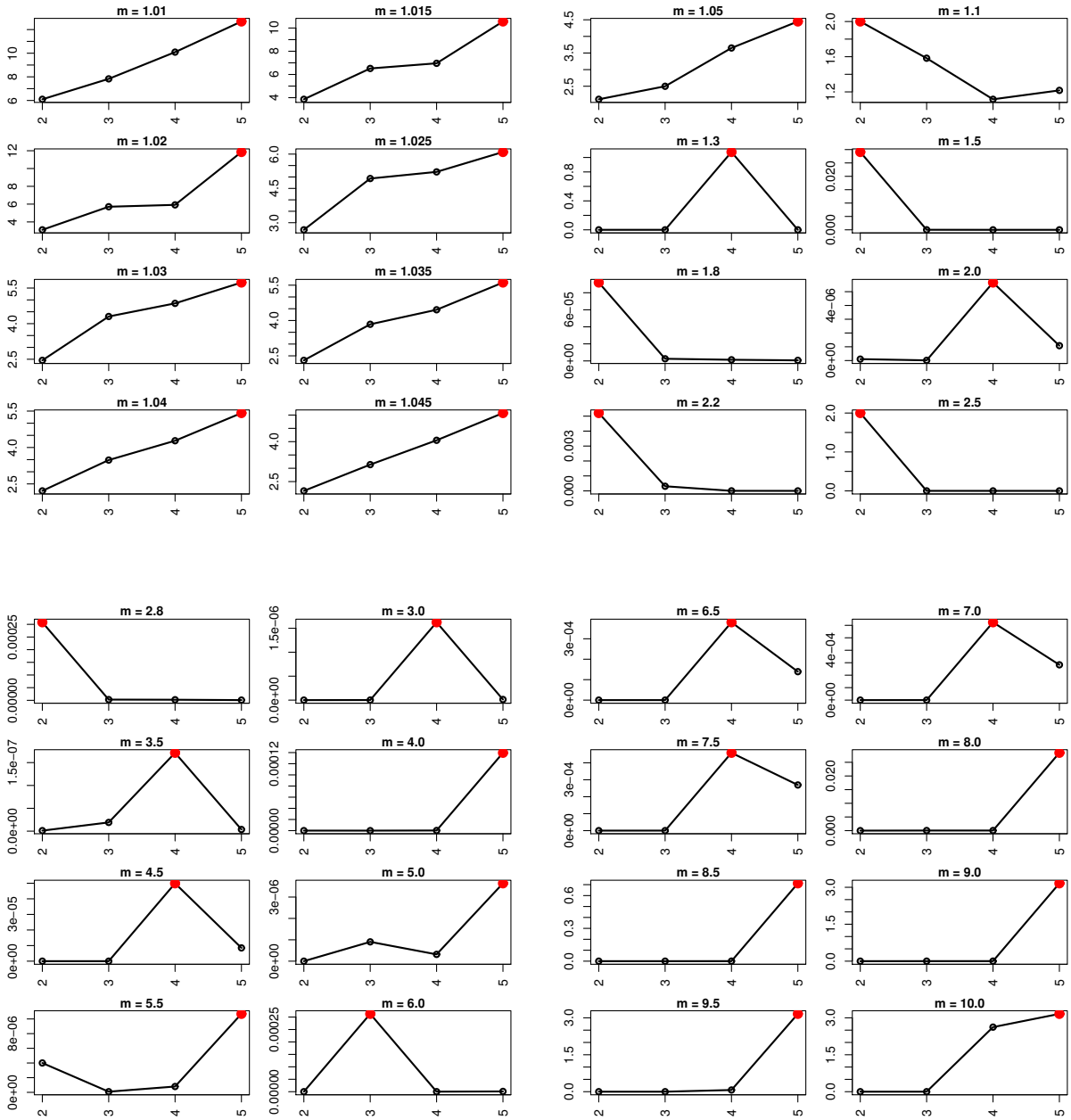


Tabela 170 – 20Newsgroups

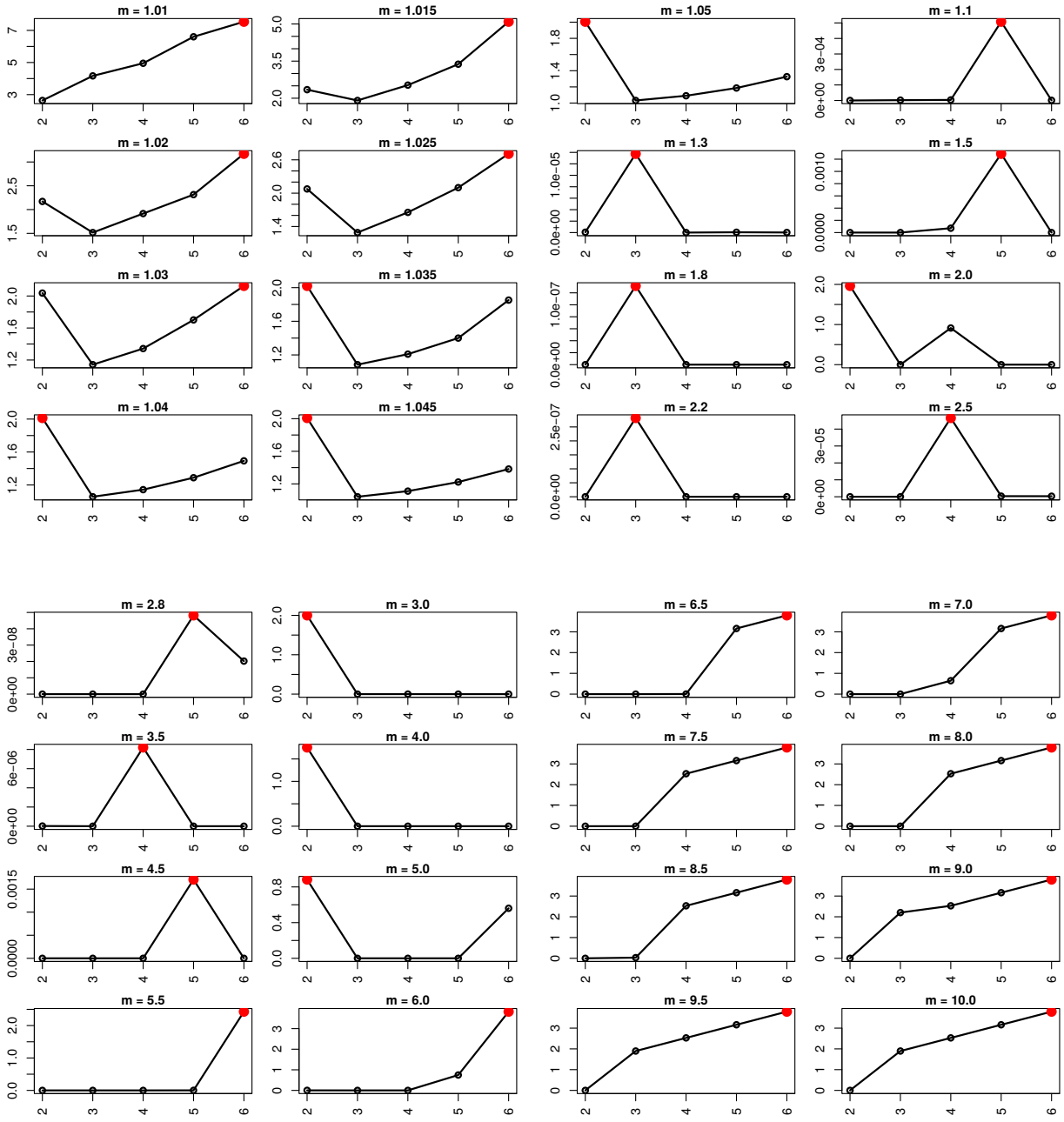


Tabela 171 – La1s

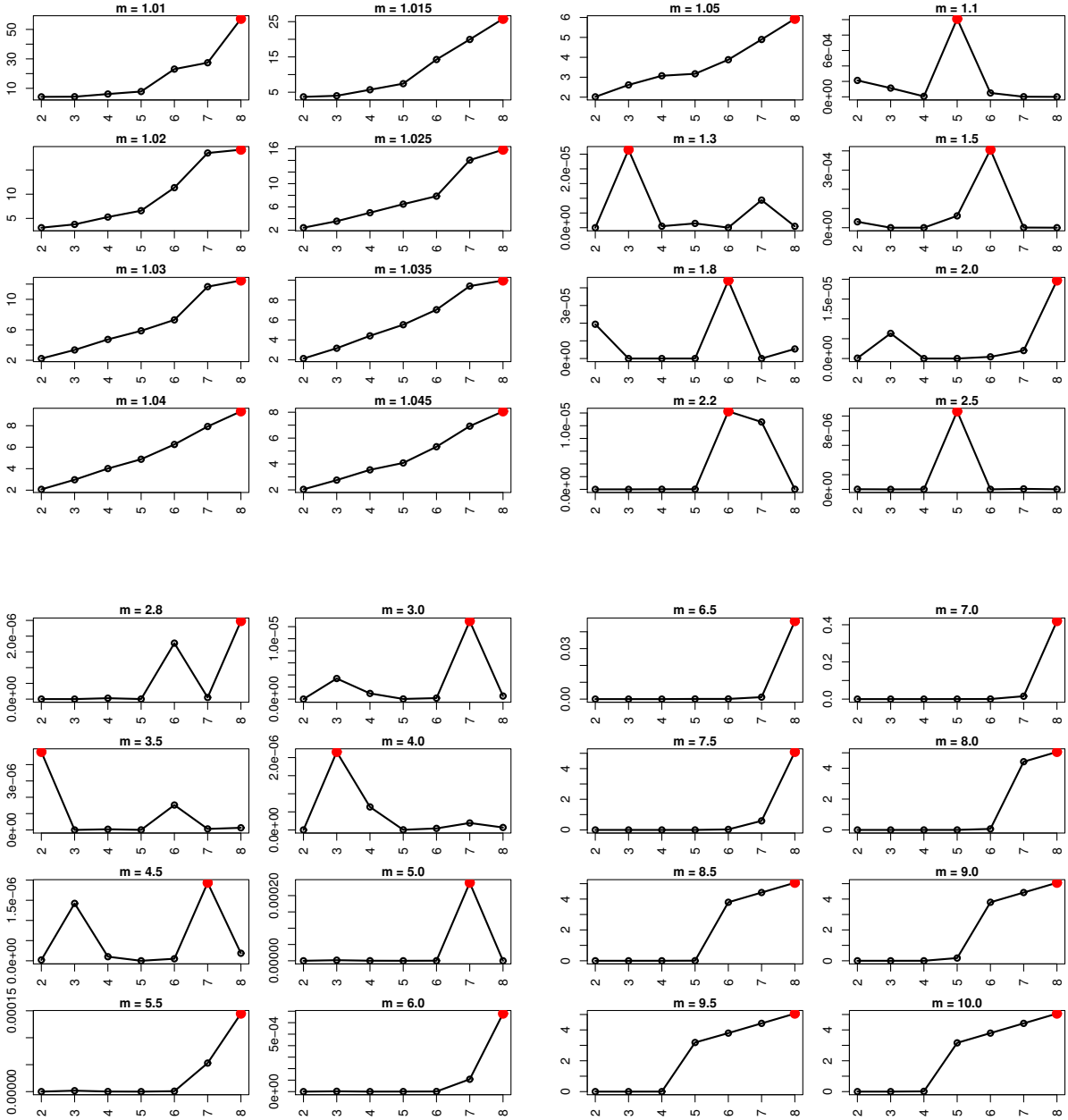
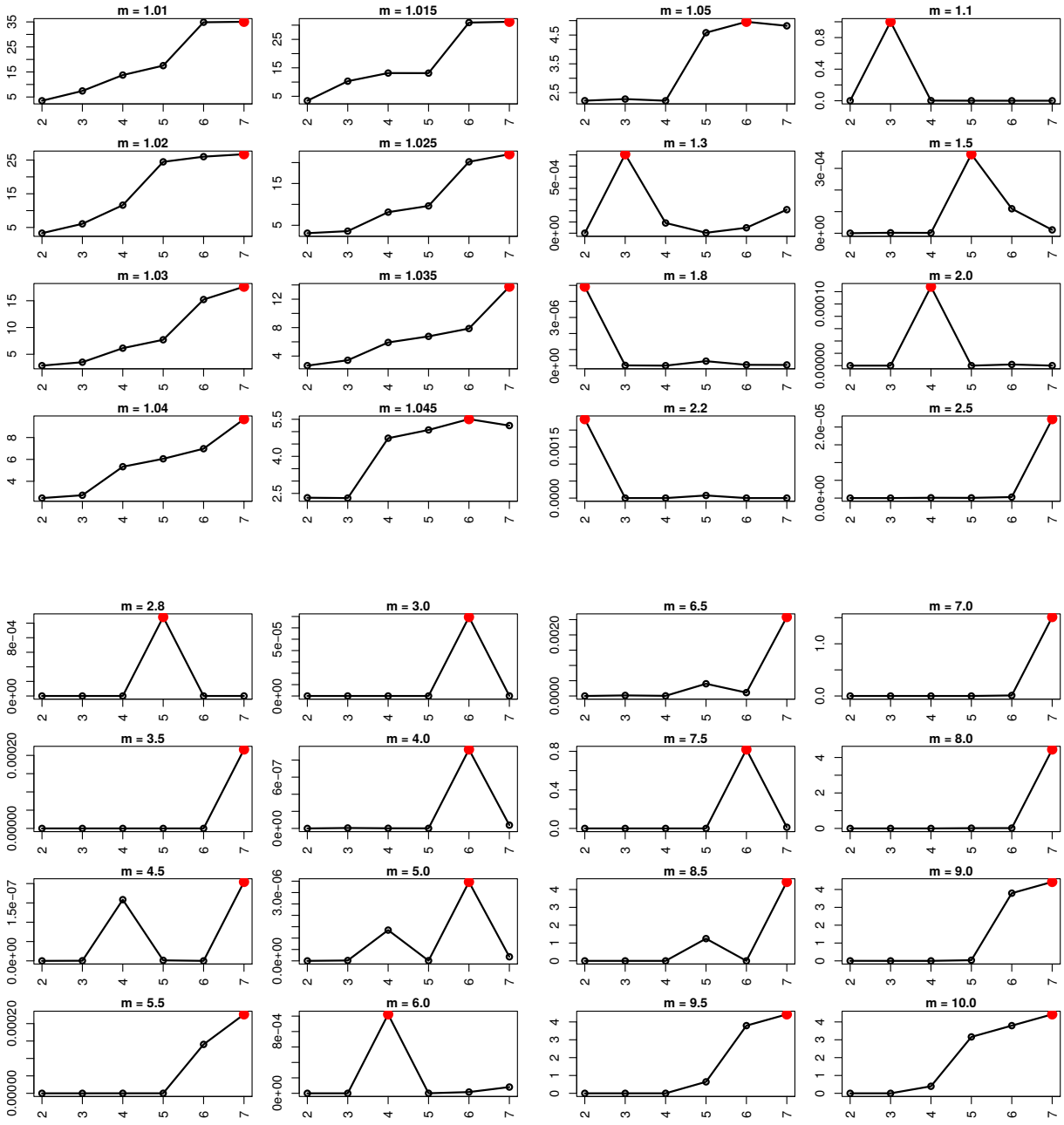


Tabela 172 – Reviews



ANEXO O – T

Tabela 173 – NewYorkTimes

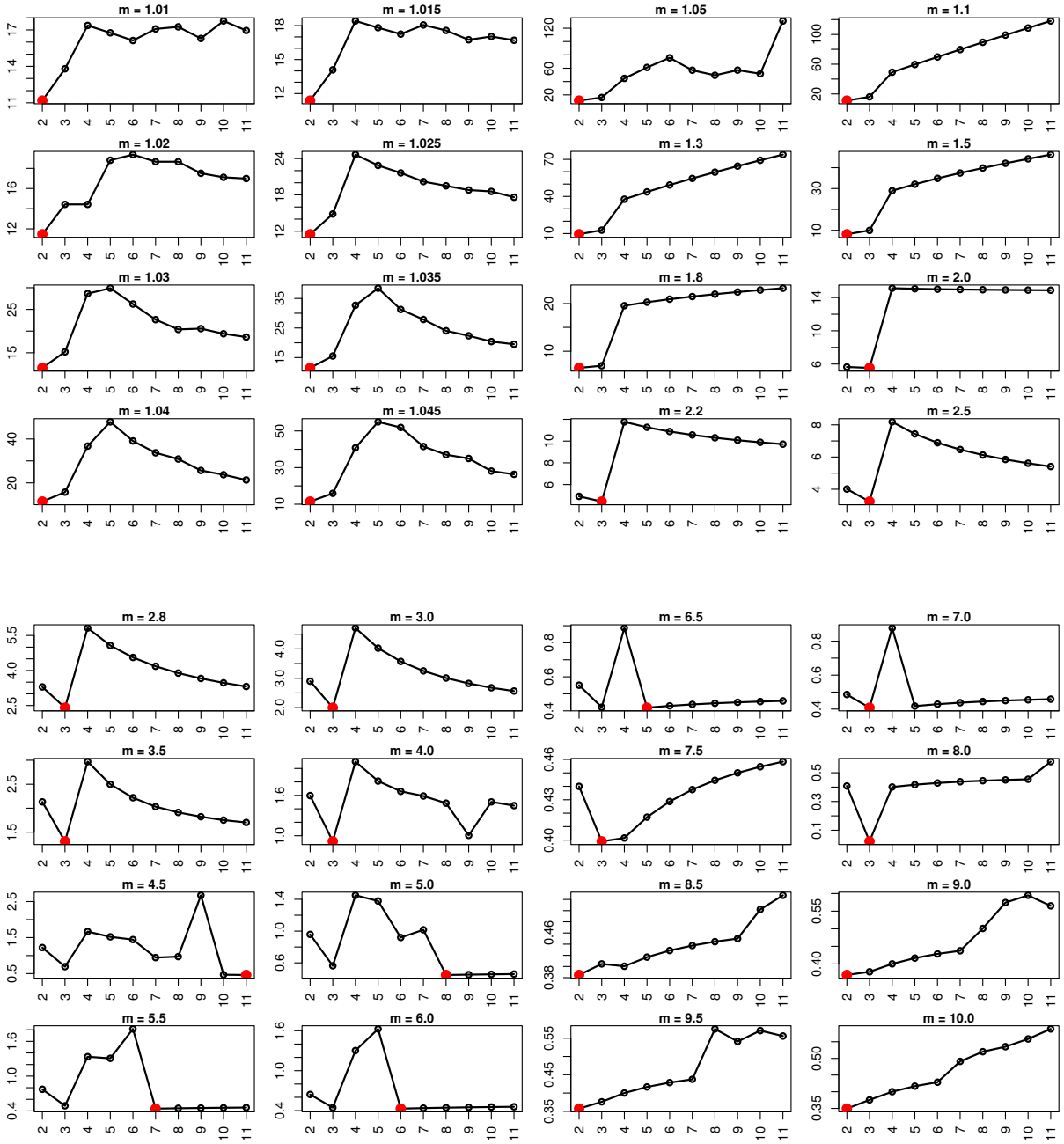


Tabela 174 – IAarticles

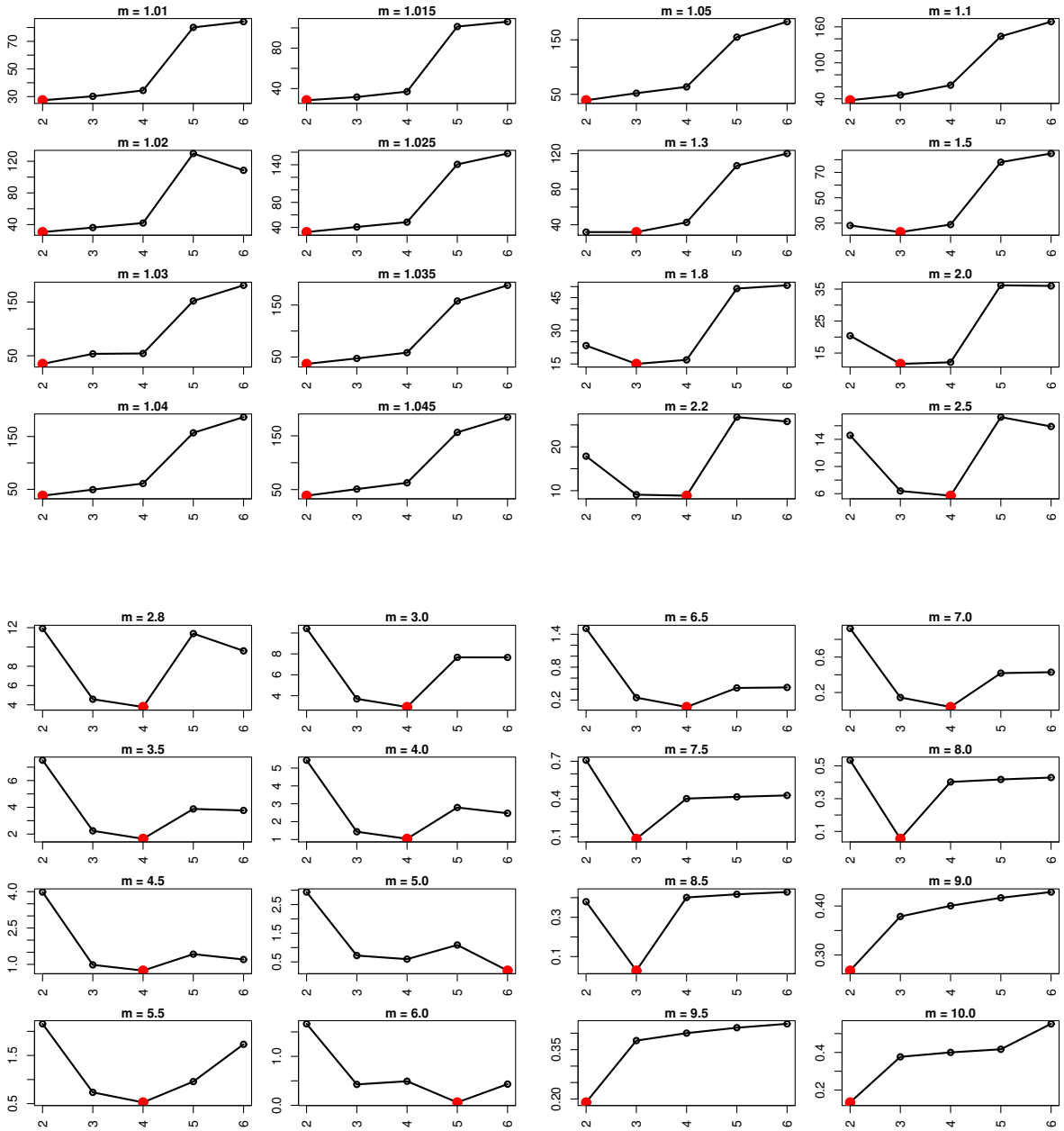


Tabela 175 – Opinosis

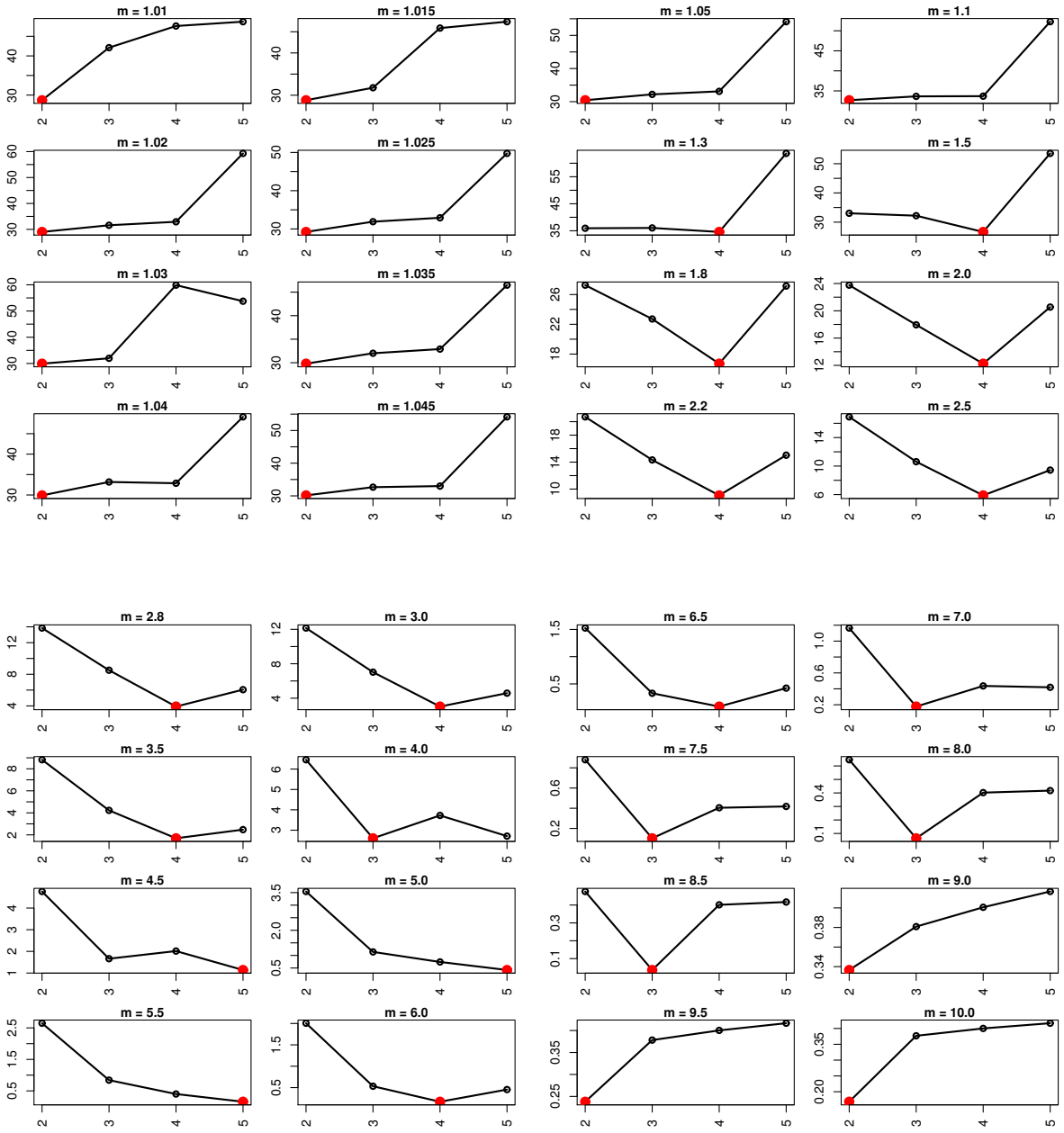


Tabela 176 – CSTR

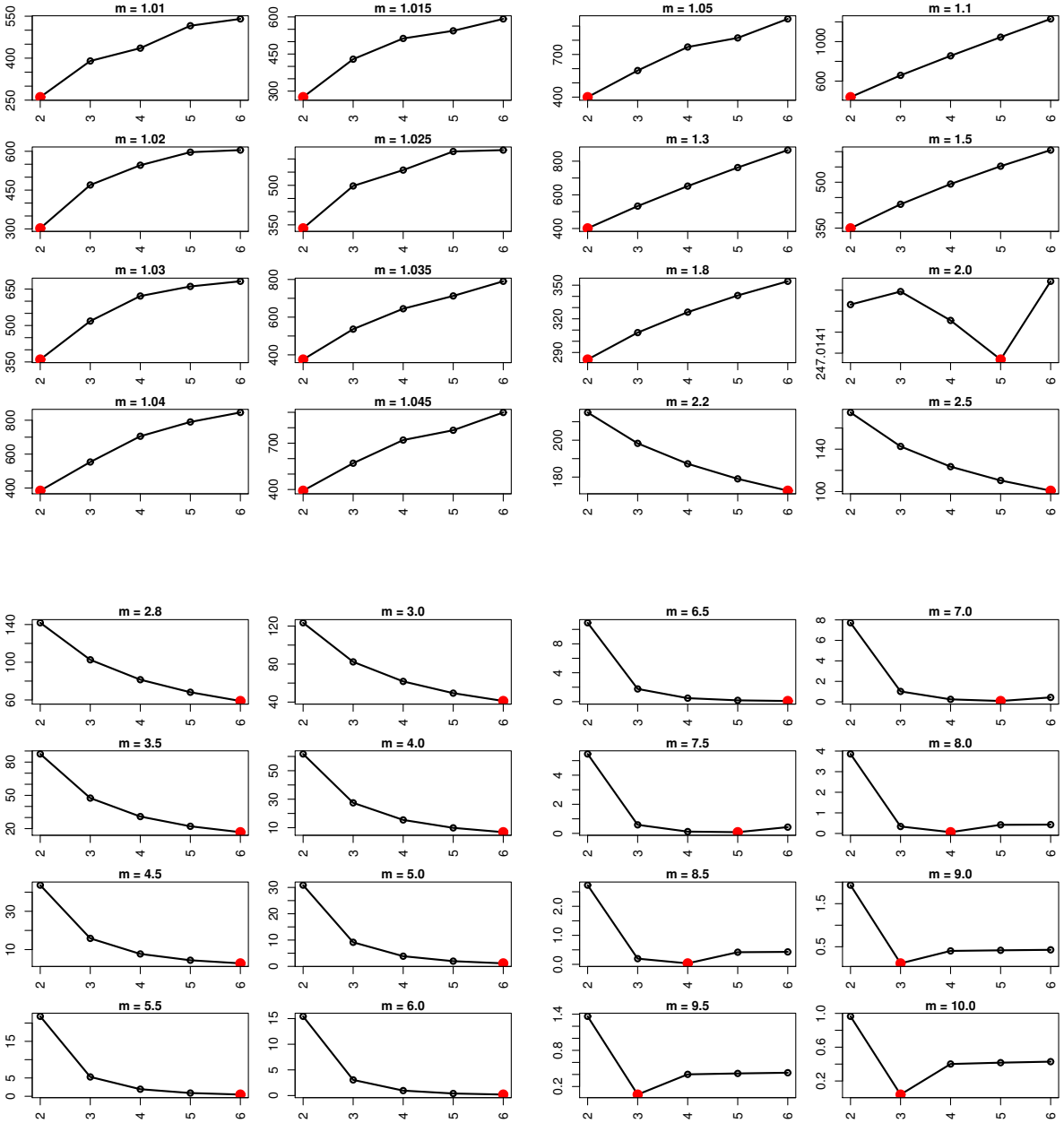


Tabela 177 – SyskillWebert

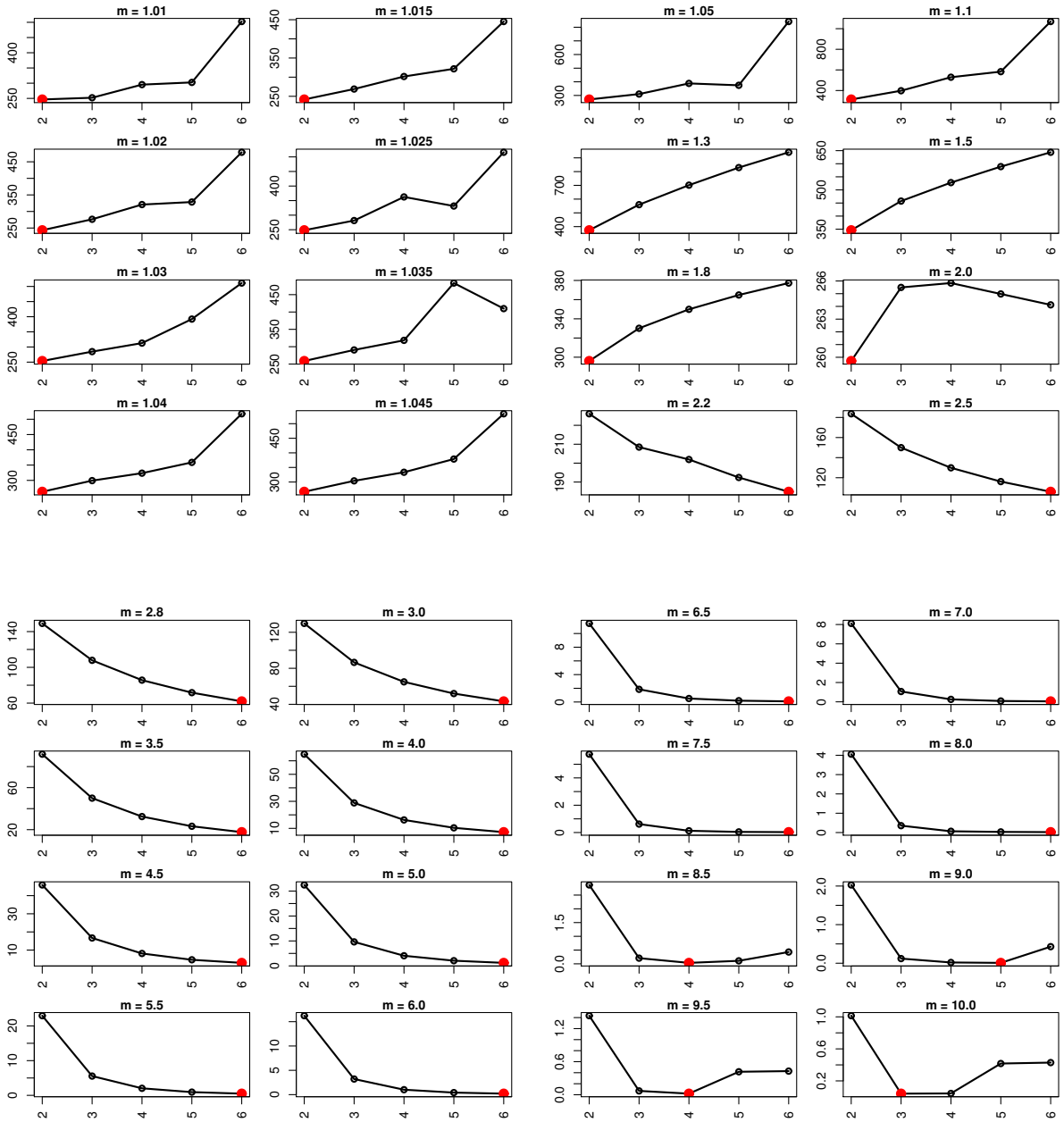


Tabela 178 – Hitech

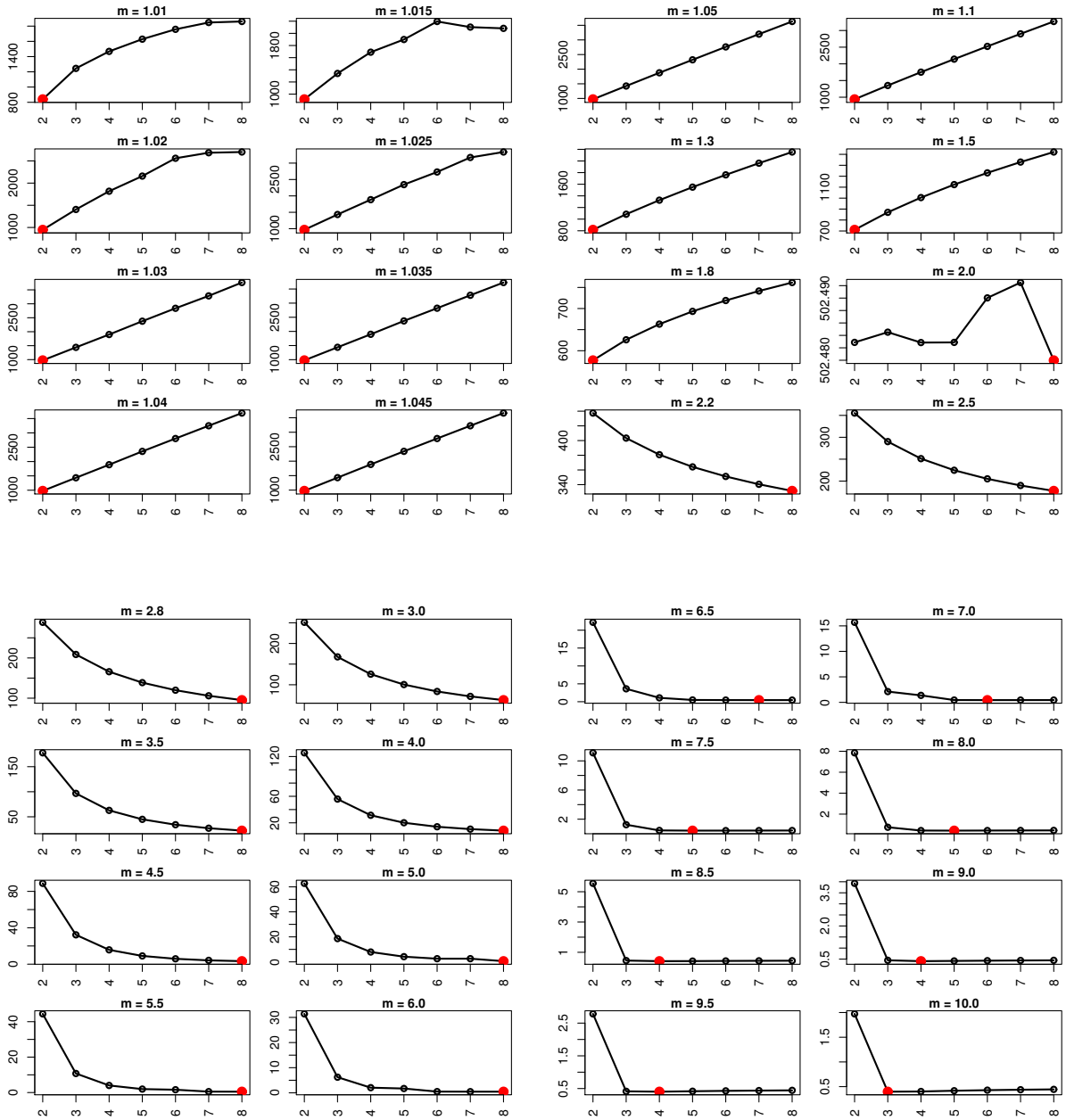


Tabela 179 – WAP

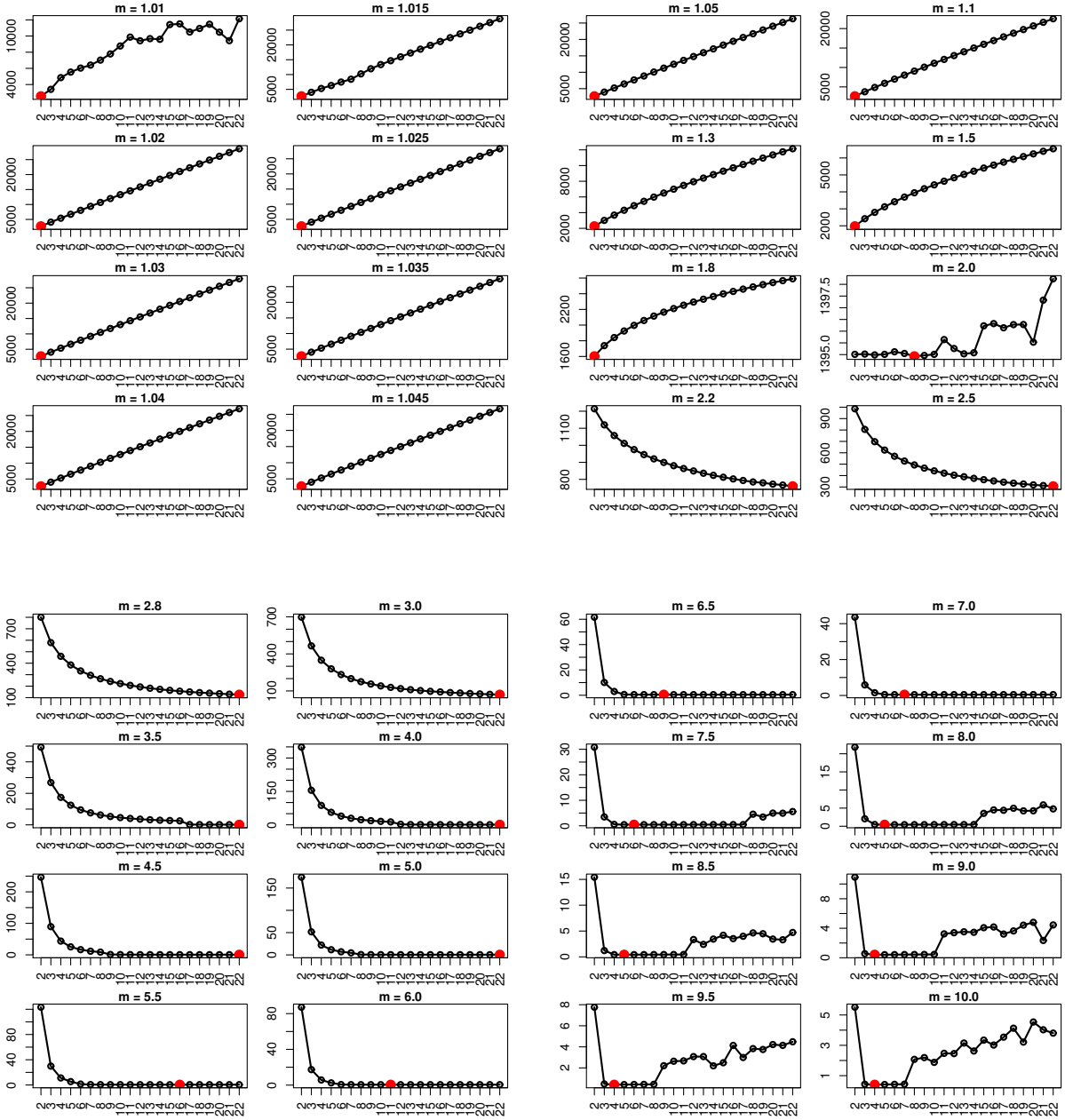


Tabela 180 – NSF

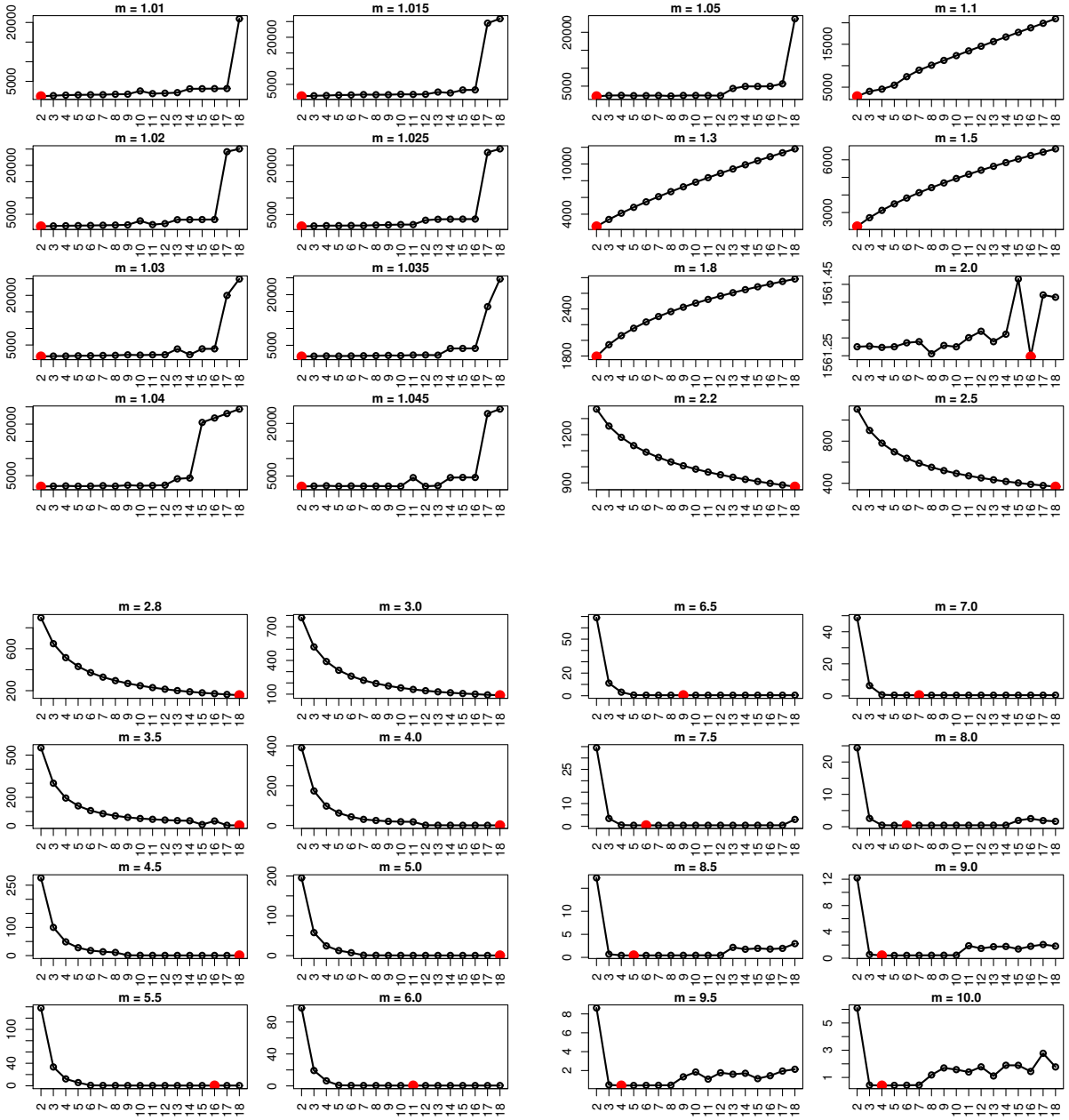


Tabela 181 – Irish-Sentiment

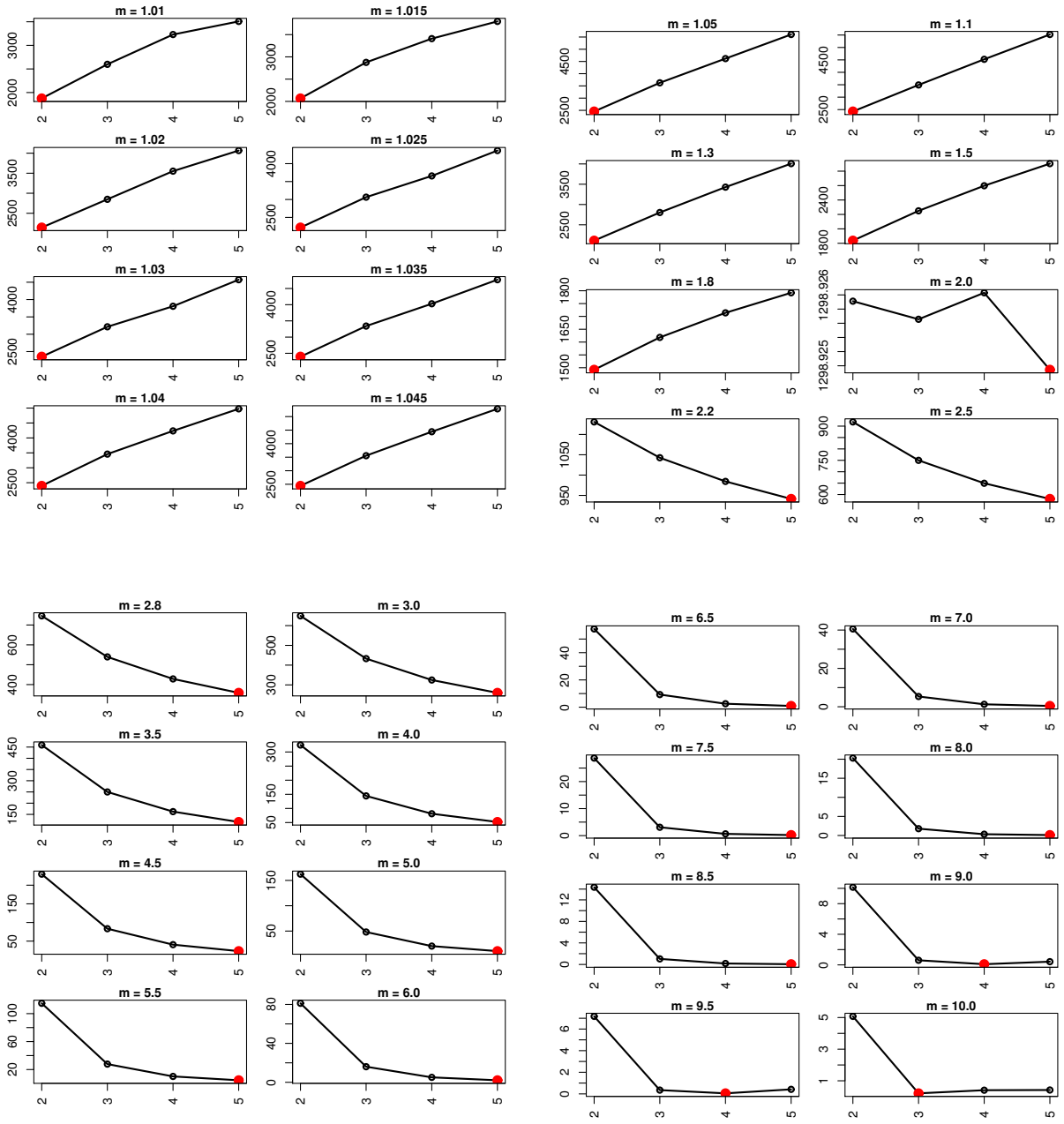


Tabela 182 – 20Newsgroups

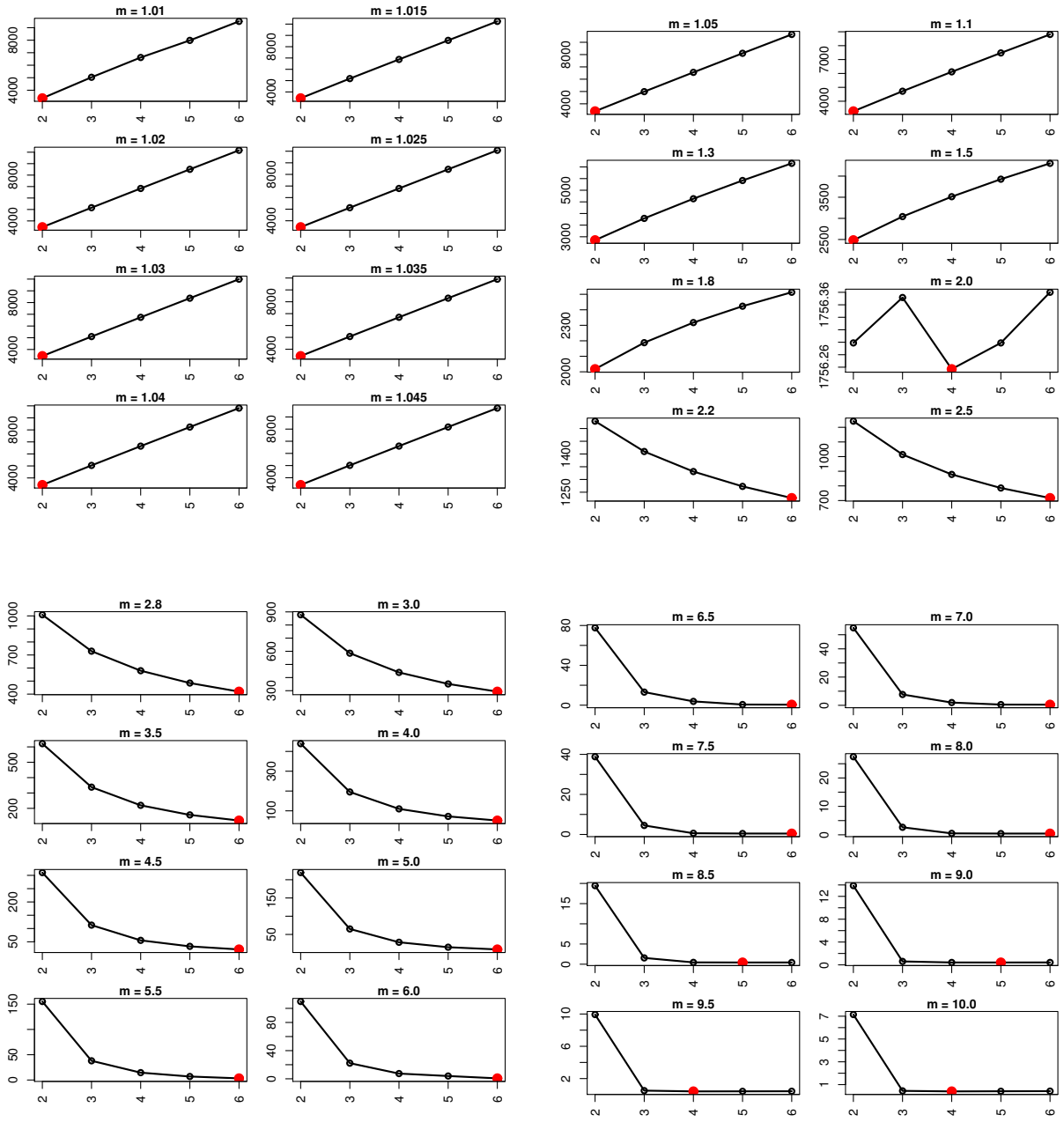


Tabela 183 – La1s

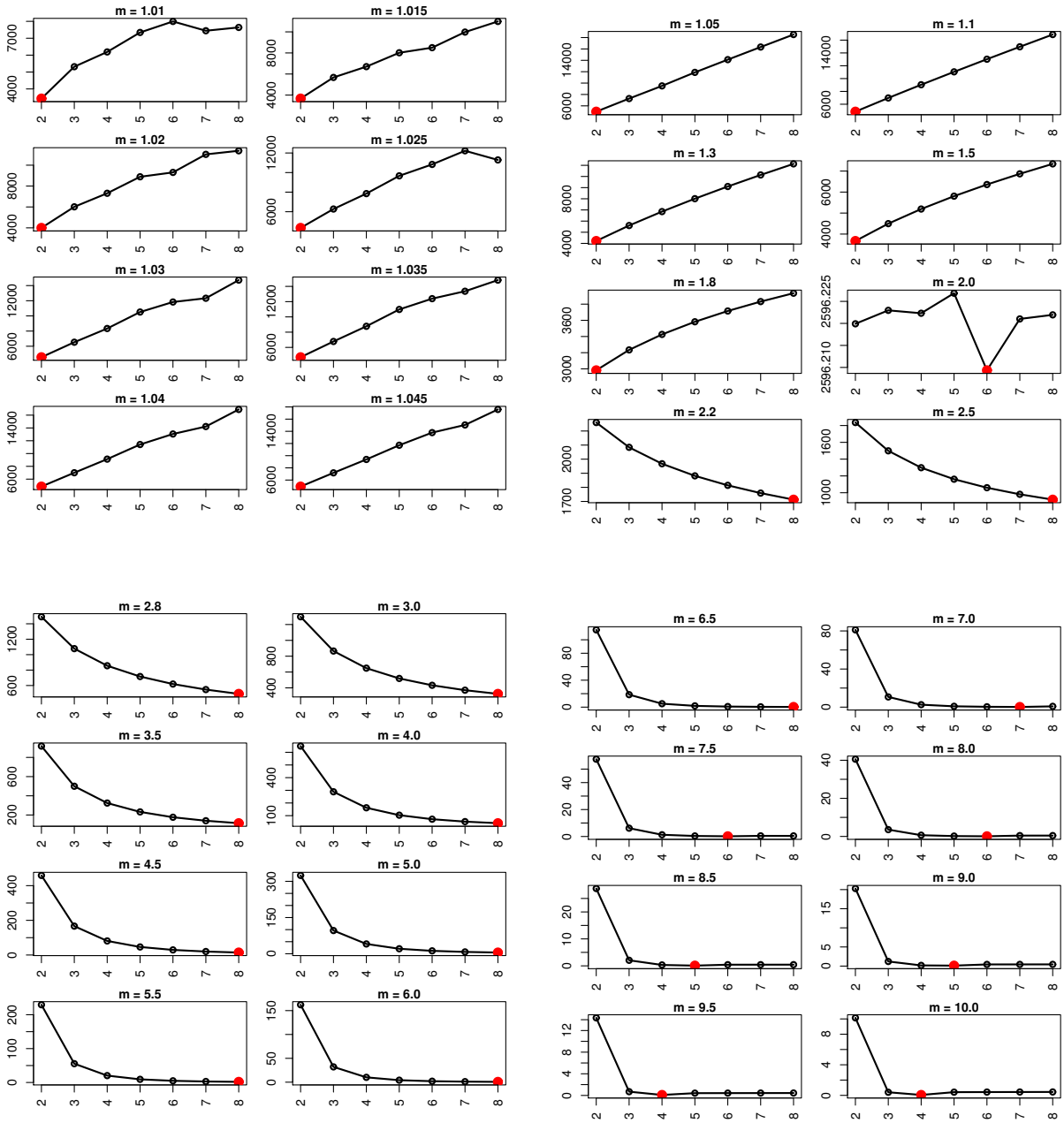
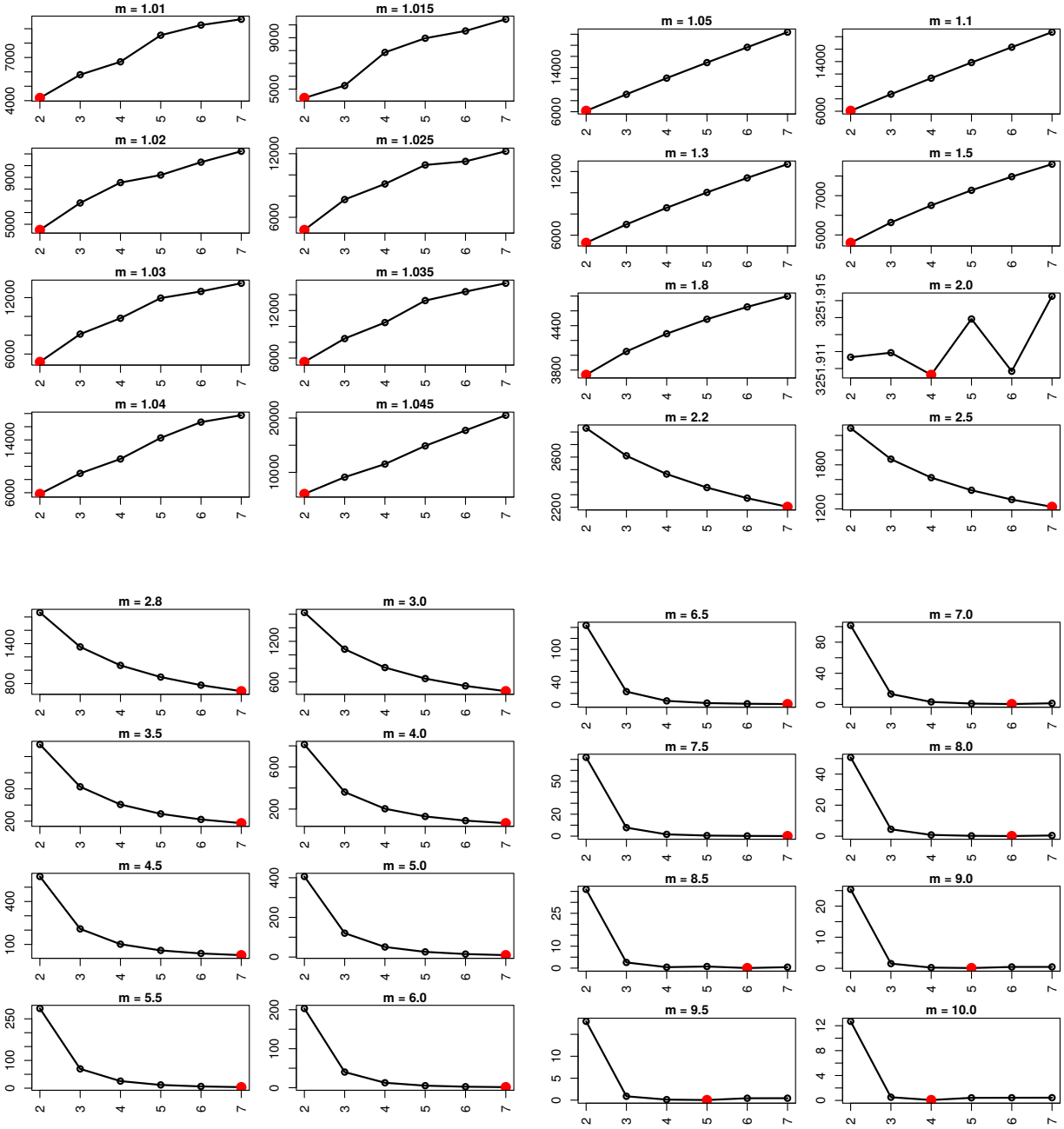


Tabela 184 – Reviews



ANEXO P – SF

Tabela 185 – NewYorkTimes

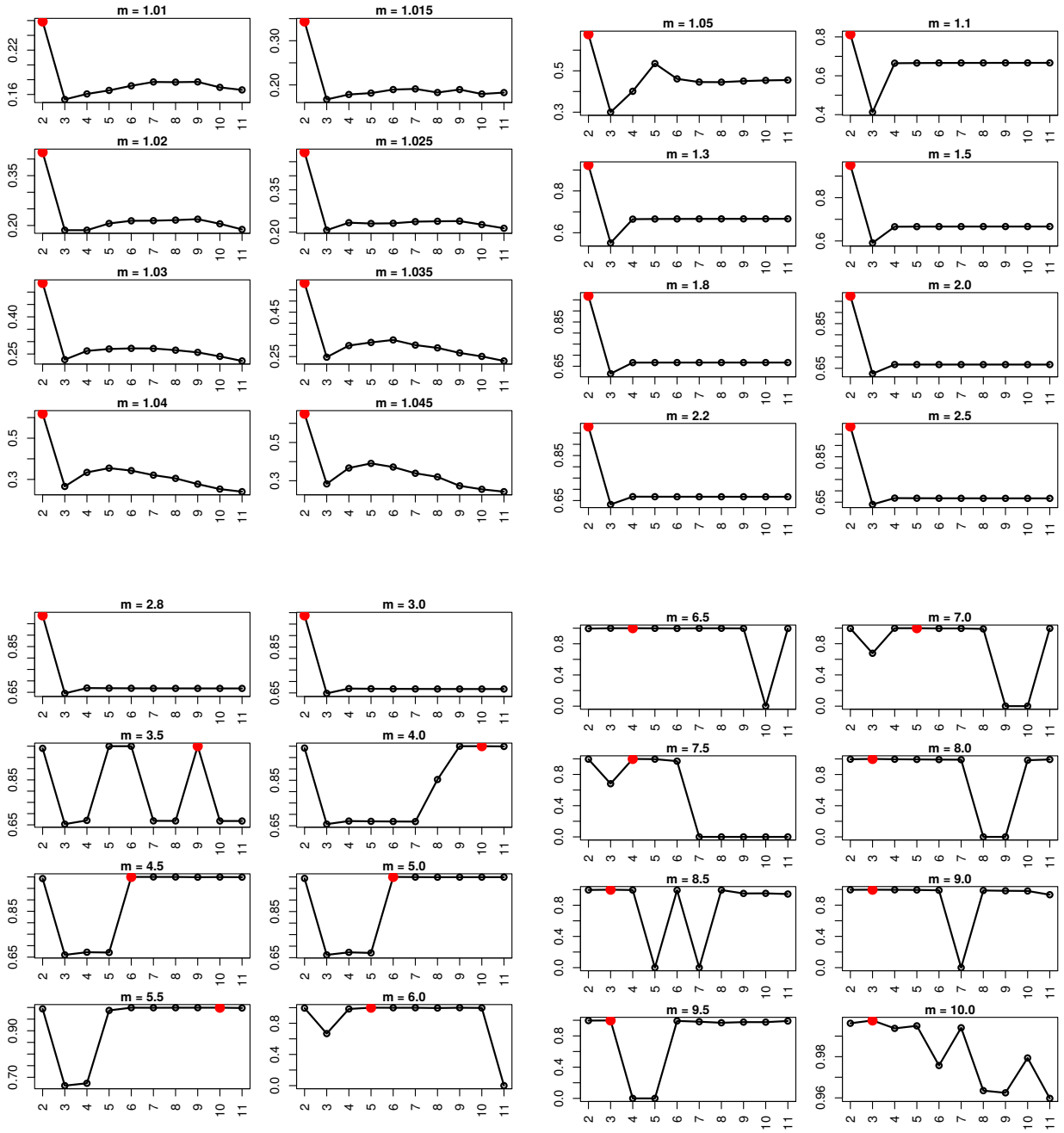


Tabela 186 – IAarticles

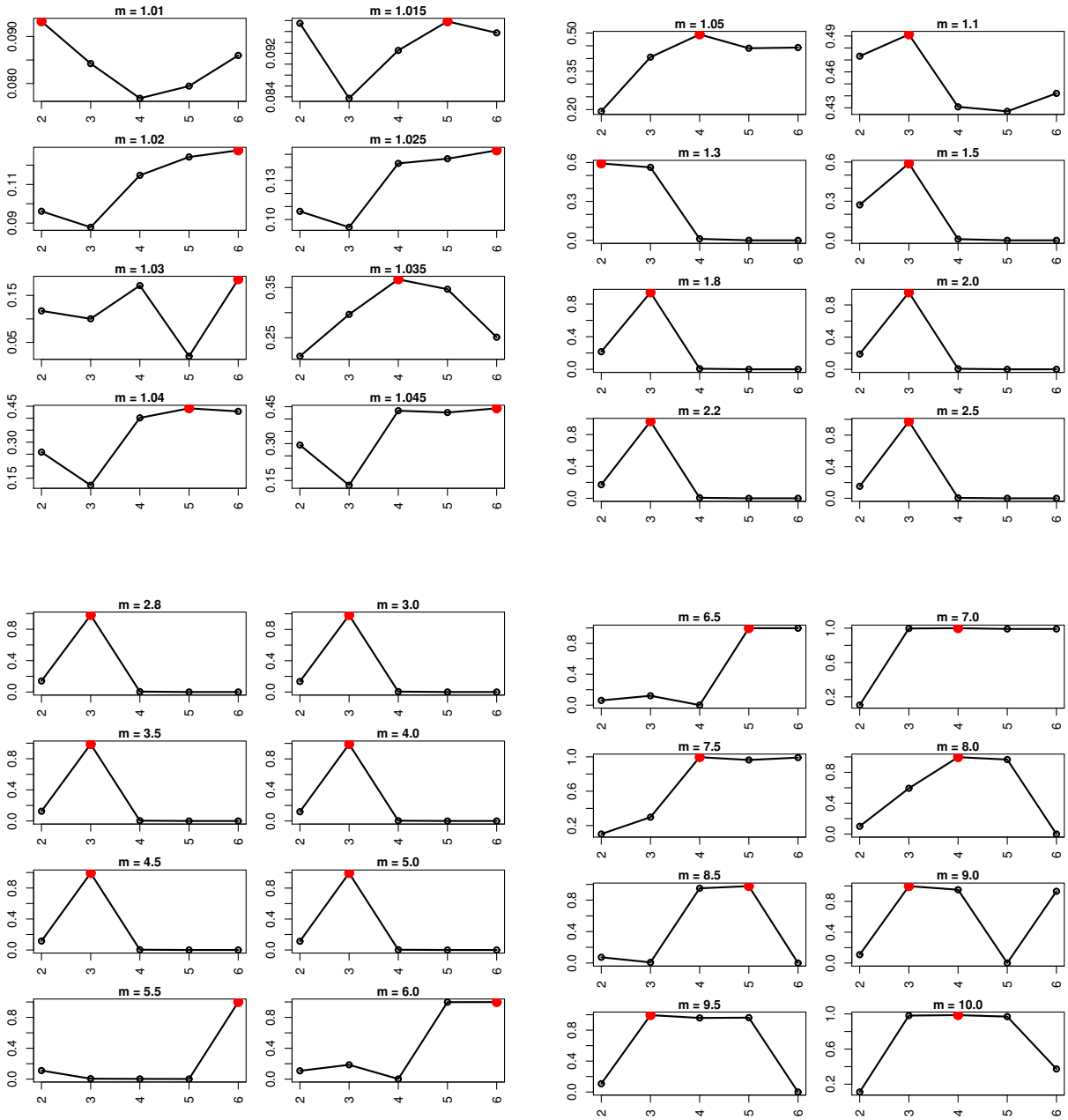


Tabela 187 – Opinosis

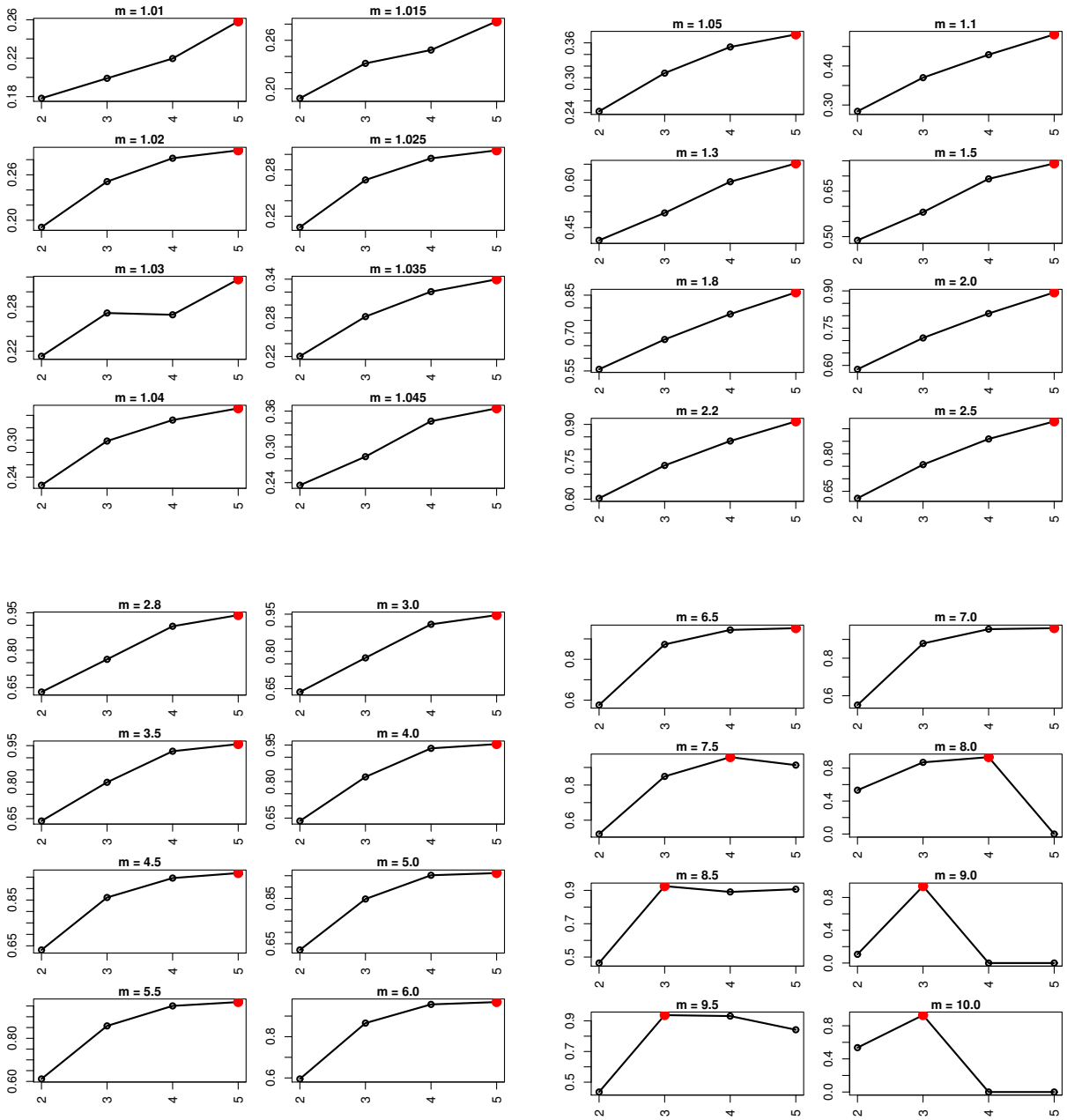


Tabela 188 – CSTR

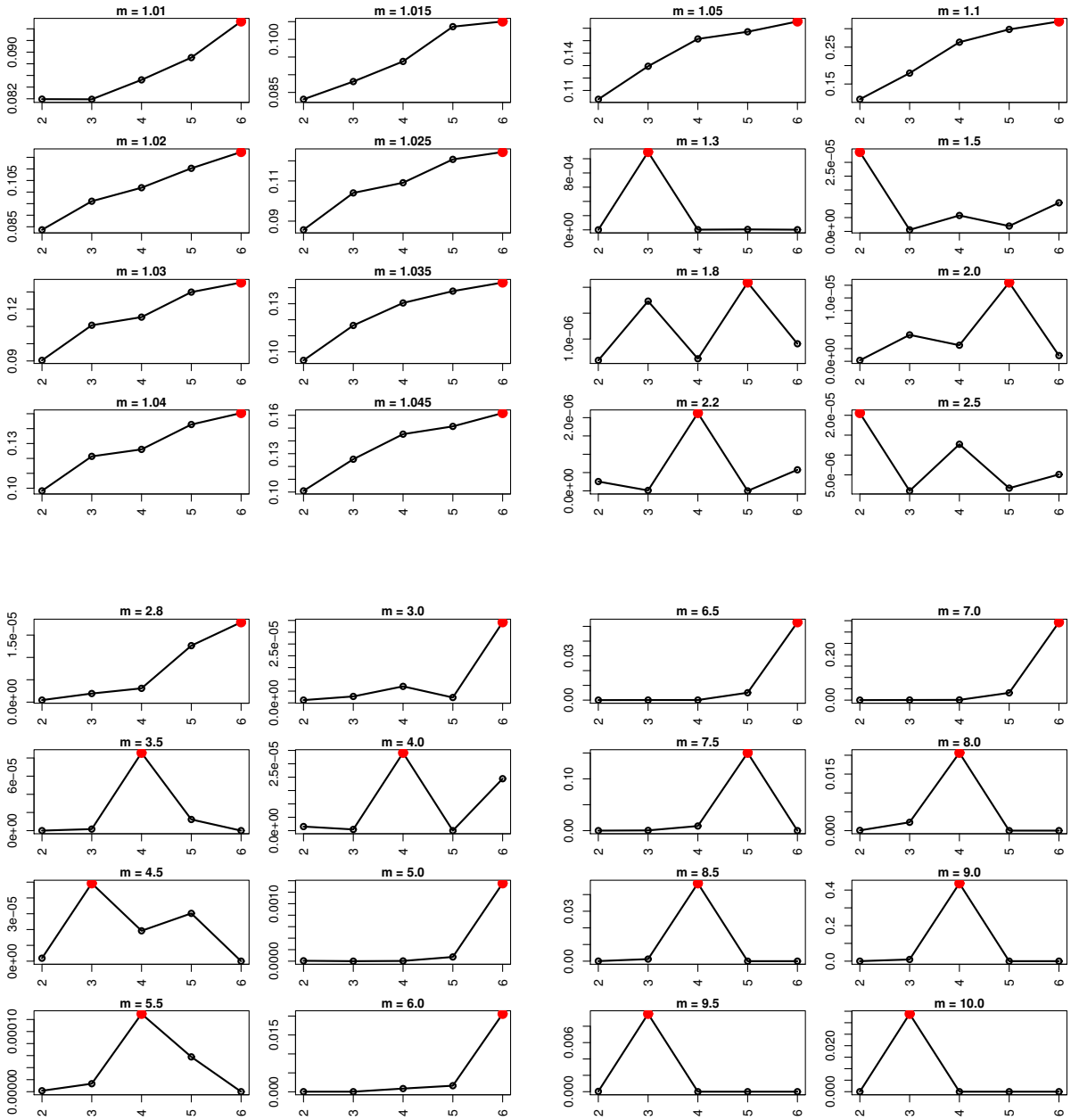


Tabela 189 – SyskillWebert

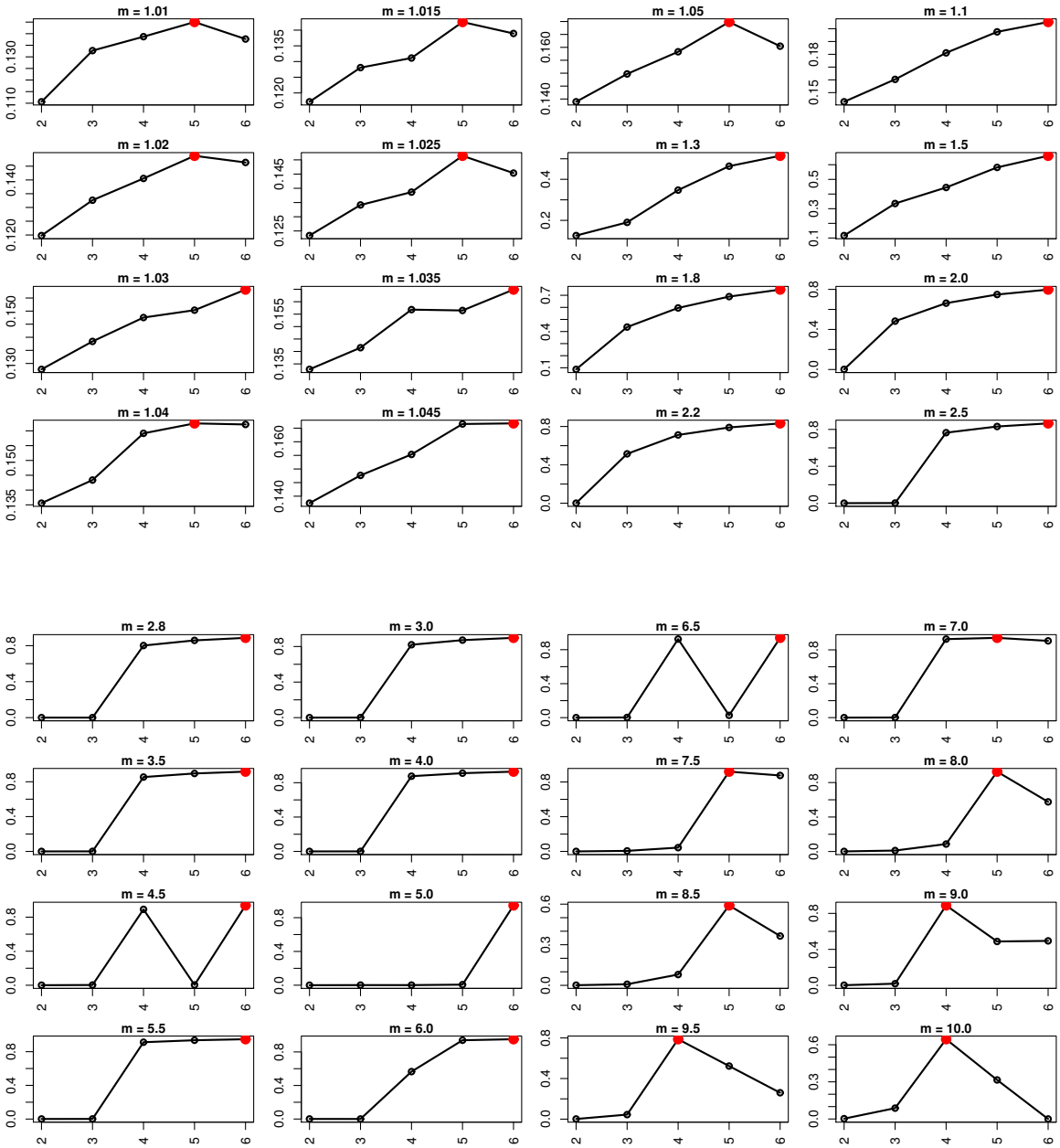


Tabela 190 – Hitech

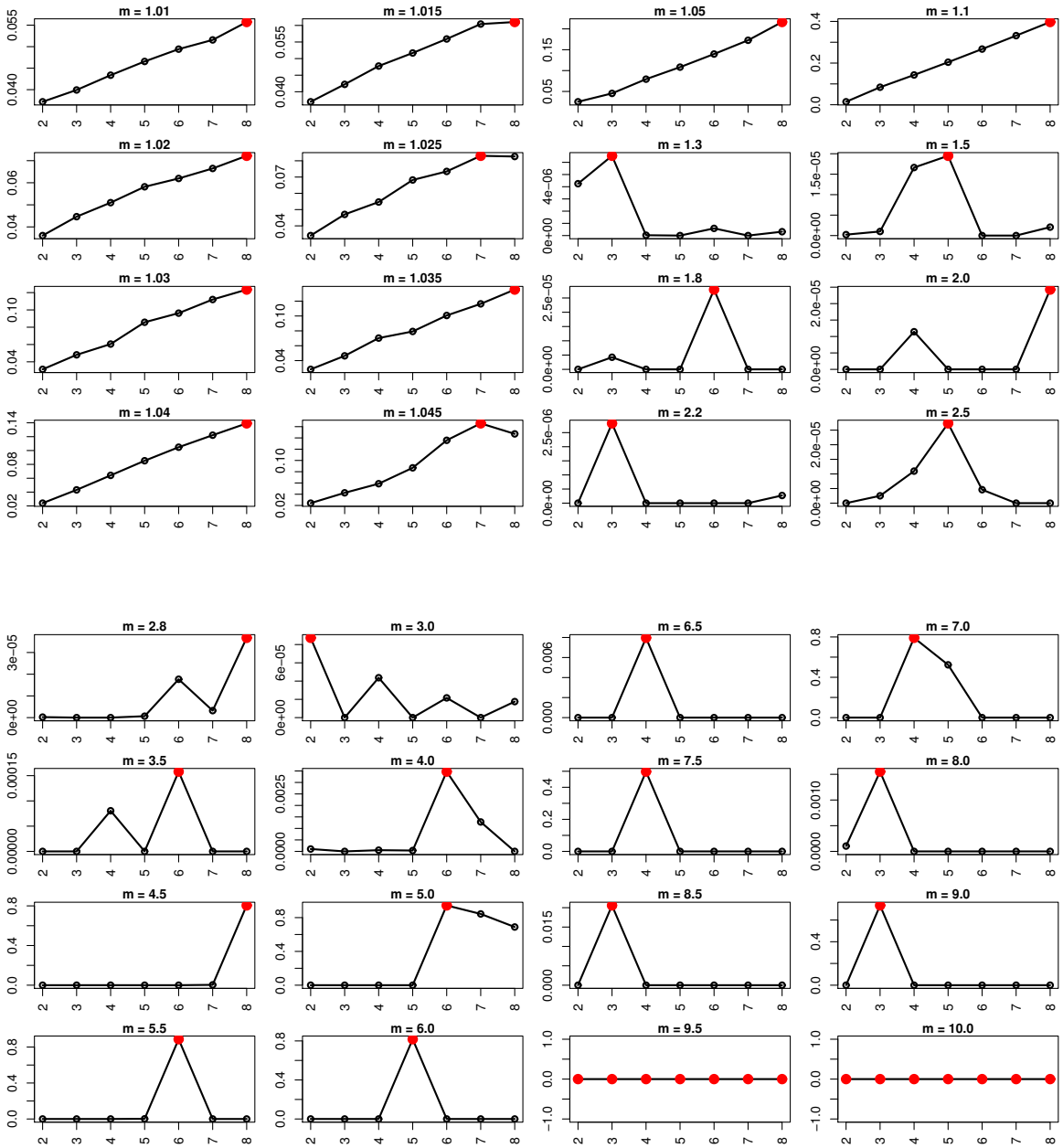


Tabela 191 – WAP

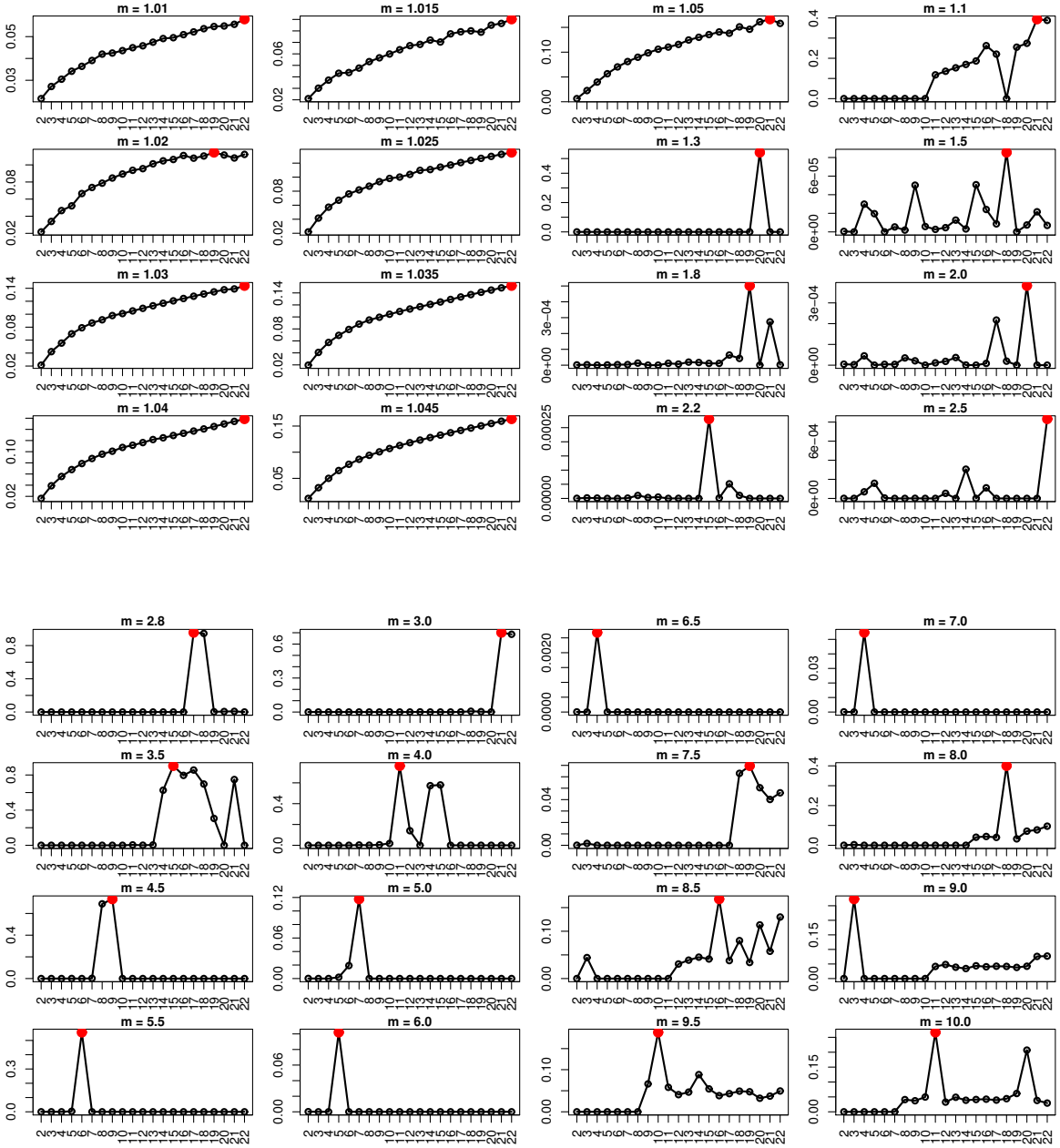


Tabela 192 – NSF

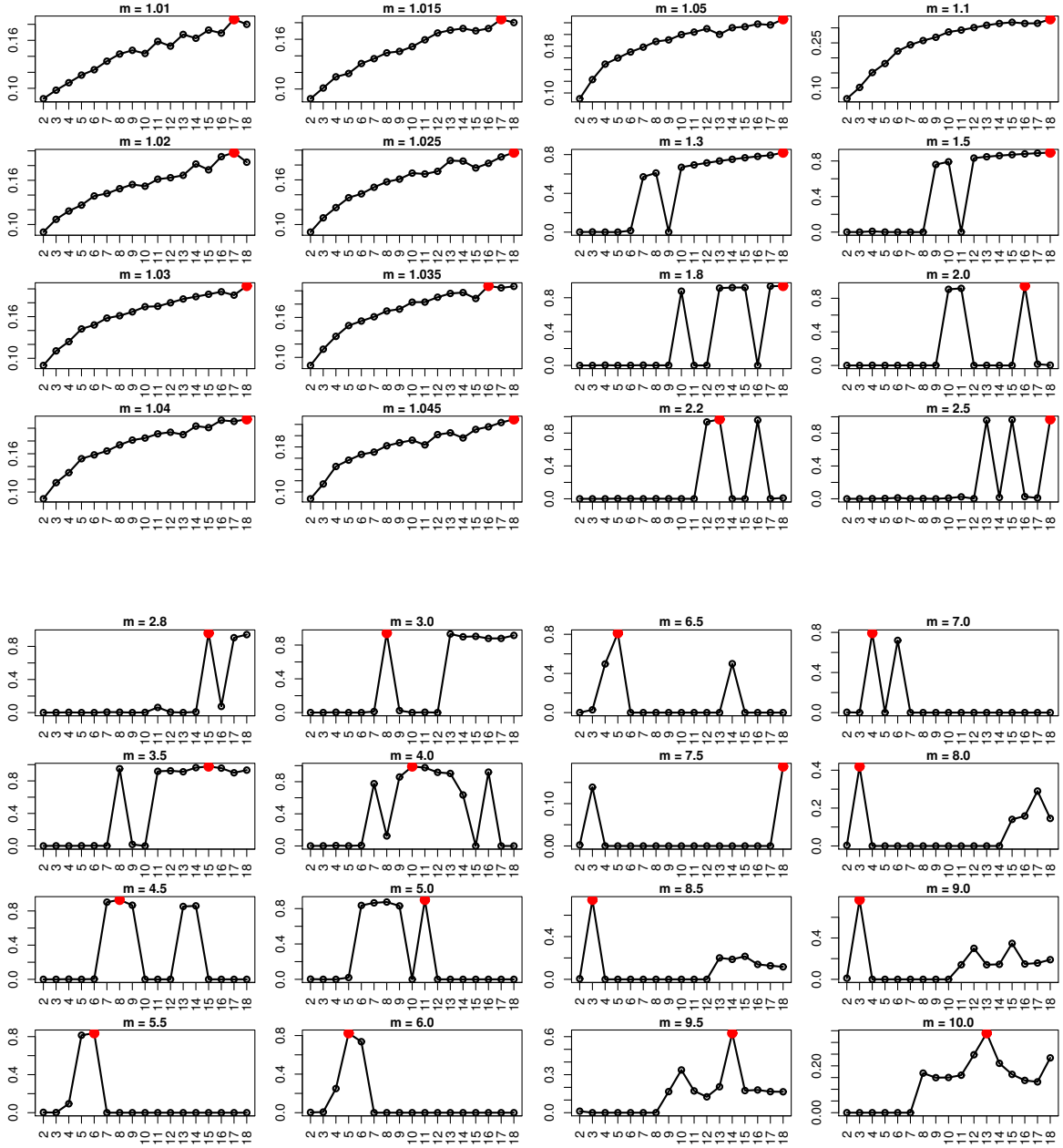


Tabela 193 – Irish-Sentiment

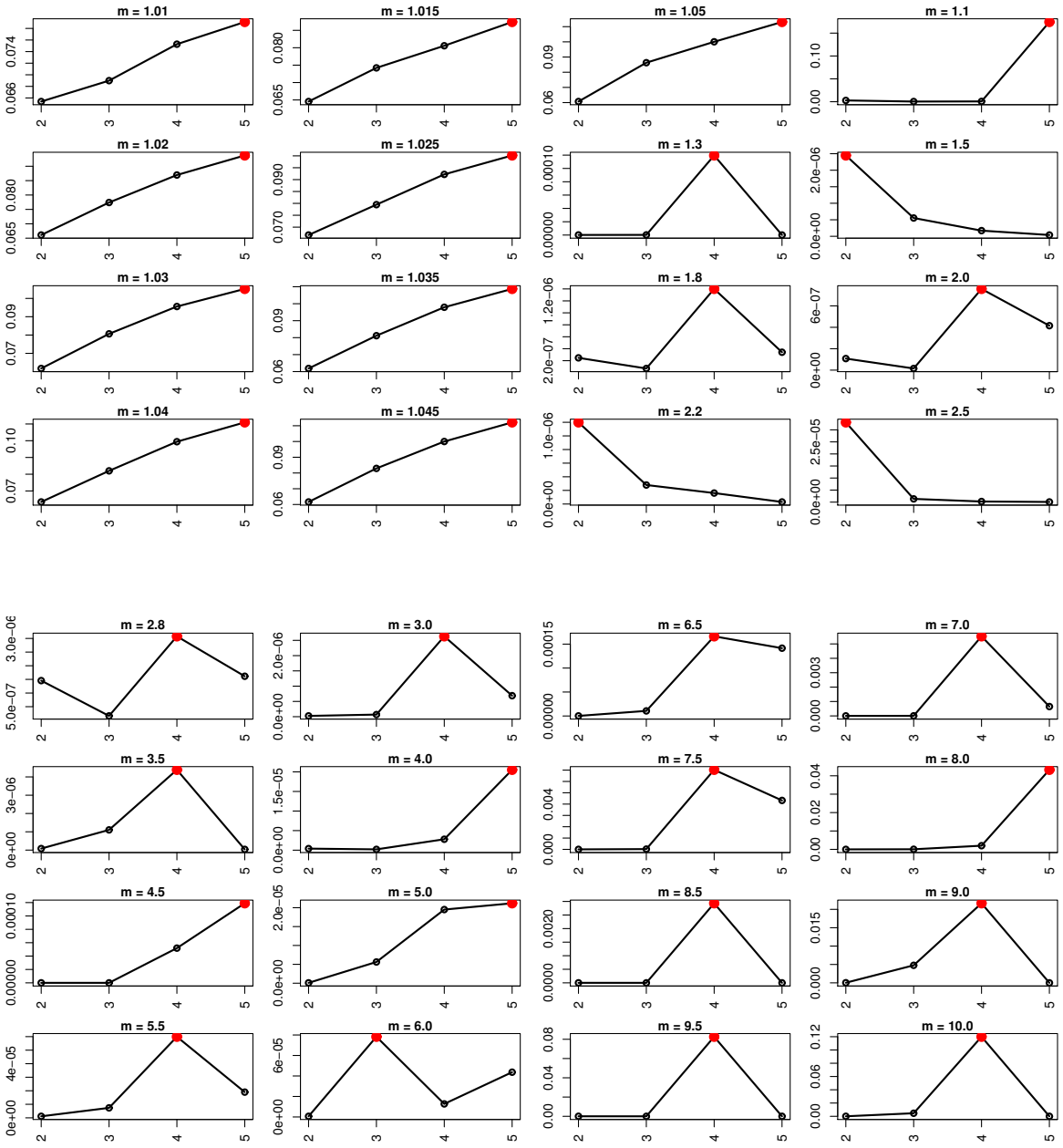


Tabela 194 – 20Newsgroups

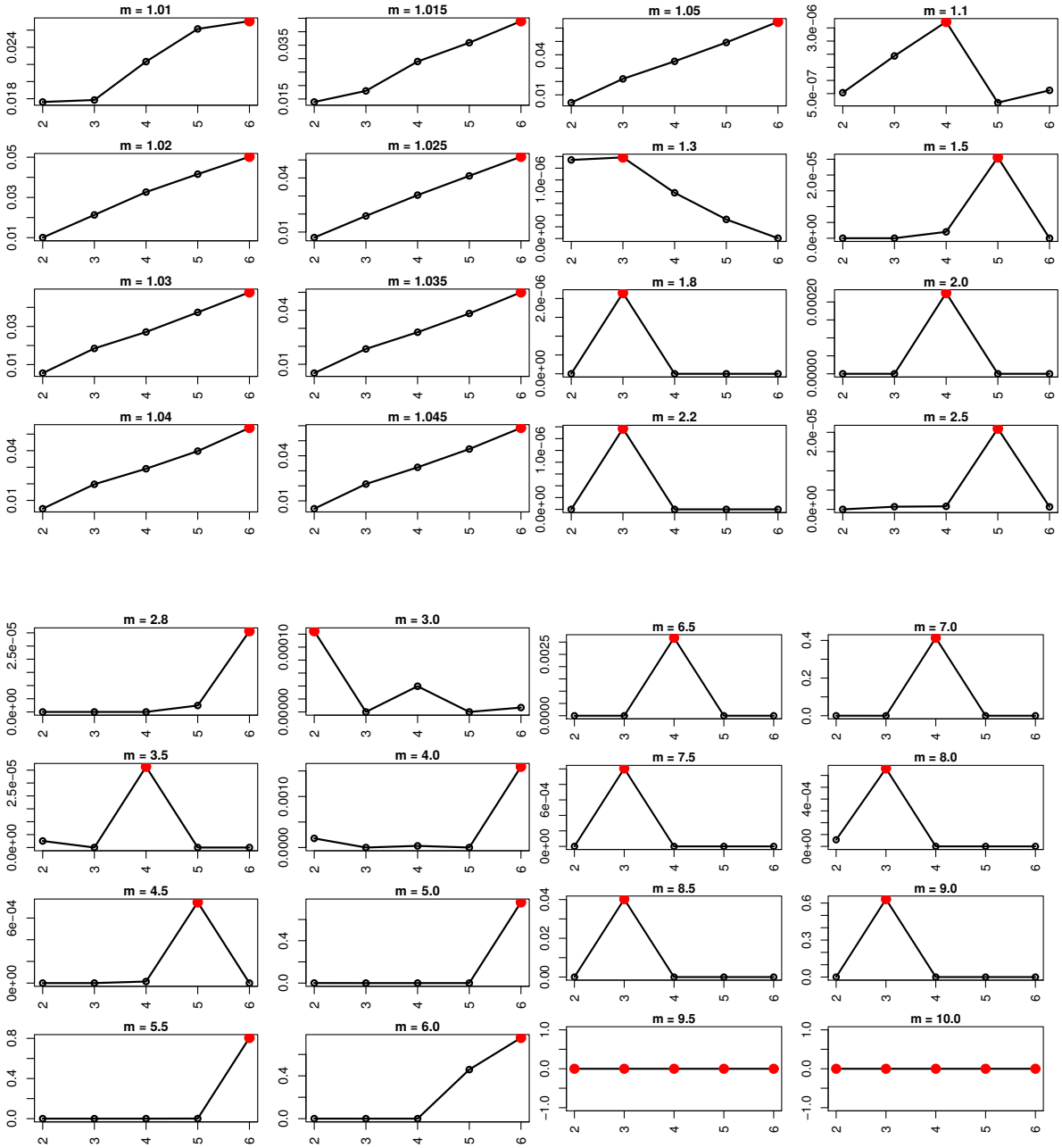


Tabela 195 – La1s

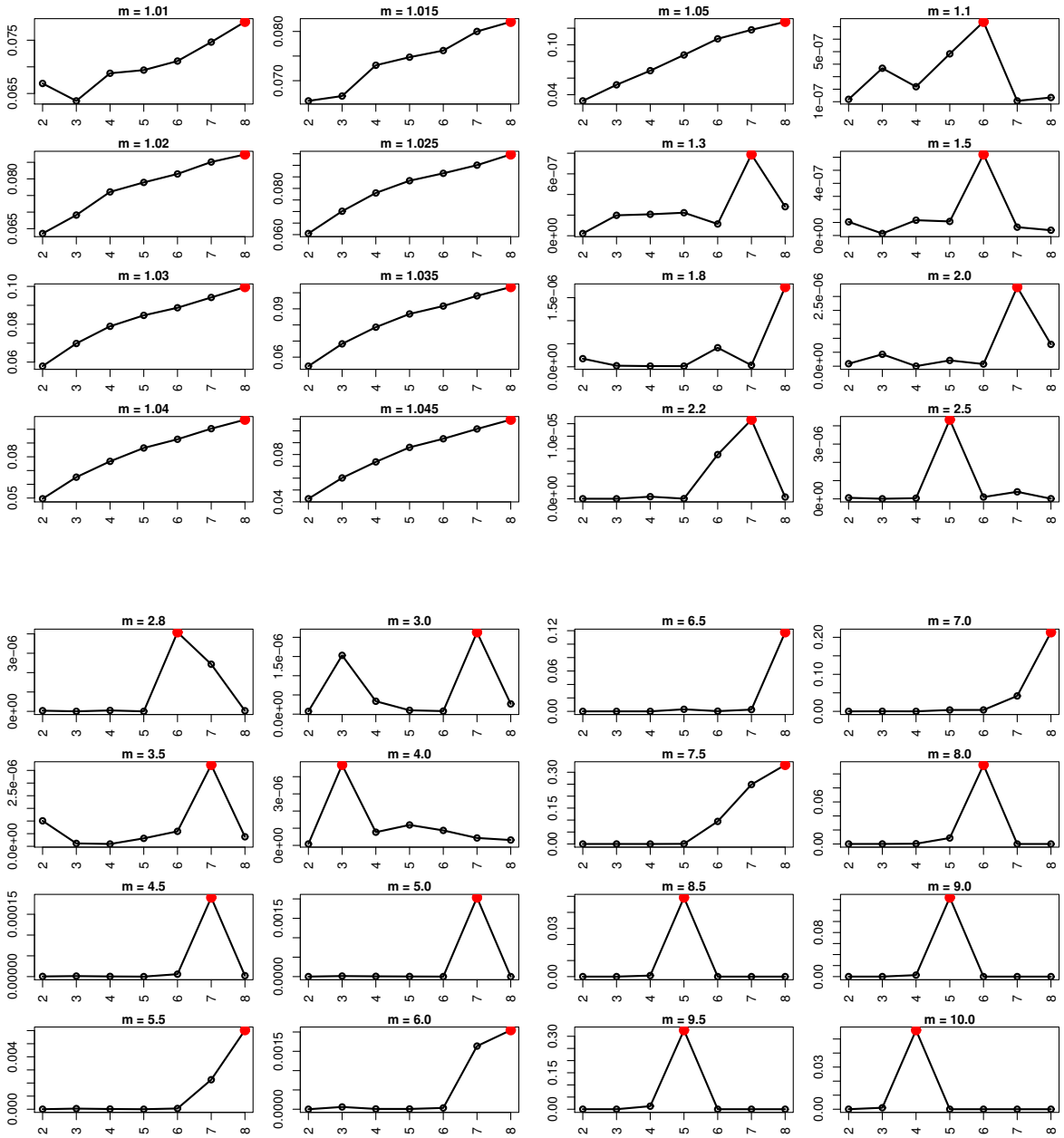


Tabela 196 – Reviews