

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



**Modelação de dados do sistema educativo
português com cadeias de Markov**

Daniel Oliveira Silvestre

Mestrado em Engenharia Informática
Especialização em Interação e Conhecimento

Dissertação orientada por:
Prof. Doutor Luís Correia
Prof. Doutor Luís Cavique

Resumo

A educação abrange vários setores que uma pessoa percorre desde o jardim de infância, ensino básico e secundário, até ao ensino superior. Neste trabalho modelamos o ensino básico e secundário português num período de 12 anos letivos, desde 2008 a 2019.

A exploração de cadeias de Markov tem sido uma abordagem promissora para a modelação deste fluxo pela obtenção de dados relevantes. Neste trabalho estudamos a aplicação de vários tipos de cadeias de Markov para a modelação do fluxo de alunos no ensino português. A previsão e entendimento de dados fornecidos e resultados obtidos é favorável para minimizar problemas que possam surgir no planeamento educacional, mas também melhorar aspetos já presentes como recursos humanos e infraestruturas.

Palavras-chave: fluxo de alunos, cadeias de Markov, aprendizagem automática.

Abstract

The education includes various sectors that a person progresses through from kindergarten, primary and secondary education, and finally higher education. In this work, we model the Portuguese primary and secondary education in a period of 12 school years from 2008 to 2019.

A promising approach for the modeling of this flow has been the exploration of Markov chains by the obtaining of relevant. In this work, we will study the application of different types of Markov chains to model the flow of students in the Portuguese education. The prediction and understanding of the provided and obtained results are propitious for the minimization of problems that might occur in educational planning but also to improve existing aspects such as infrastructures and human resources.

Keywords: student flow, Markov chains, machine learning.

Índice

Capítulo 1	Introdução.....	1
1.1	Motivação.....	1
1.2	Contribuições	2
1.3	Estrutura do documento	2
Capítulo 2	Trabalho relacionado.....	3
2.1	O Problema do Fluxo de Estudantes	3
2.1.1	Ensino português	5
2.2	Processos Estocásticos	7
2.2.1	Processo Markoviano	7
2.2.2	Cadeia de Markov	8
Capítulo 3	Análise de Dados.....	10
3.1	Valores absolutos dos dados	10
3.2	Análise de tendências	12
3.3	Sobreposição de curvas	13
Capítulo 4	Modelação das Cadeias de Markov.....	17
4.1	Modelo proposto	17
4.2	Modelo base	18
4.2	Evolução do modelo	22
4.2.1	Probabilidades de transição	22
4.2.2	Ingressos.....	25
Capítulo 5	Conclusões	32
Referências		34
Apêndices		37

Lista de Figuras

Figura 1: Fluxo de alunos para ano curricular N.....	6
Figura 2: Diagrama exemplo de cadeia de Markov	9
Figura 3: Curvas das probabilidades de abandono do terceiro ciclo	13
Figura 4: Exemplo de como duas curvas mostram semelhanças quando sobrepostas	13
Figura 5: Exemplo ilustrativo de 2 curvas semelhantes: curva A a vermelho e curva B a verde	14
Figura 6: Cálculo de diferenças e diferenças após aplicação de offset	14
Figura 7: Alteração do offset mediano obtém piores resultados	15
Figura 8: Diagrama da cadeia de Markov pretendida	18
Figura 9: Matriz de transição genérica	19
Figura 10: Representação gráfica da diferença dos erros de previsão entre o 1º, 2º e 3º ano curricular (retirado do primeiro gráfico do apêndice F).....	20
Figura 11: Exemplo de modelação de uma regressão quadrática (dados até ano letivo 2017/2018)	24
Figura 12: Exemplo de overfitting	29

Lista de Tabelas

Tabela 1: Resultados do modelo teste	22
Tabela 2: Resultados do modelo com probabilidades dinâmicas.....	25
Tabela 3: Média de idades por ano curricular	25
Tabela 4: Coeficientes de determinação das regressões lineares	26
Tabela 5: MAPEs das diferentes regressões polinomiais para a previsão dos ingressos	28
Tabela 6: MAPEs das diferentes regressões polinomiais para a previsão dos ingressos com um desfasamento de 1 ano.....	30
Tabela 7: Resultados do modelo com probabilidades e ingressos dinâmicos (saldo migratório)	31
Tabela 8: Resultados do modelo com probabilidades e ingressos dinâmicos (imigrações)	31

Capítulo 1

Introdução

1.1 Motivação

O projeto ModEst tem como objetivo o estudo e análise do fluxo dos alunos no ensino português. Para tal, aborda o tópico em diferentes ambientes como a visualização de informação, modelação e outros métodos de análise (Caviqúe et al, 2020). Este projeto é possível com a colaboração da Direção-Geral de Estatísticas da Educação e Ciência (DGEEC) que nos disponibiliza dados dos alunos compreendendo os anos letivos desde 2008/09 até o ano 2019/20 (12 anos letivos).

A previsão do fluxo de alunos pode trazer imensos benefícios para o ensino. Com uma melhor compreensão deste fluxo naturalmente se obtém um melhor ensino e seu desenvolvimento através de uma melhor organização e utilização de recursos.

As cadeias de Markov propostas neste trabalho irão ser uma abordagem relevante a este problema devido à natural evolução de um sistema educacional que evolui e esse altera-se sequencialmente ao longo dos anos. O sucesso deste trabalho poderá também, revelar aplicabilidade dos modelos propostos em outros casos de estudo.

Objetivos

Os objetivos deste trabalho, sendo um trabalho do projeto ModEst, são a modelação do fluxo de alunos no ensino básico e secundário português para os anos letivos fornecidos pela DGEEC. Estes dados incluem várias informações necessárias para a modelação das quais as mais importantes para este trabalho são as relativas a cada aluno:

- Ano letivo;
- Ano curricular;
- Idade;
- Modalidade de ensino (diferentes tipos de cursos);
- Se é um novo aluno no ensino português (ingresso);
- Resultado do ano letivo (transição, retenção ou abandono do ensino).

Alguns dos resultados que se pretendem obter são a previsão de número de inscritos, transições, retenções e abandonos por ano curricular. Para tal, também com um objetivo, pretende-se modelar cadeias de Markov dinâmicas com o propósito de obter resultados com menor erro aos já obtidos com cadeias de Markov tradicionais.

1.2 Contribuições

Os estudos realizados neste trabalho, revelam que existem possíveis explorações e adaptações às cadeias de Markov para modelar um sistema aberto, que neste caso será o ensino em que alunos entram e saem do ensino.

Com a ambição de obter resultados promissores, modificou-se a matriz de transição das cadeias de Markov, que tipicamente utilizam valores constantes, para utilizarem valores dinâmicos que evoluem ao longo do tempo, procurando assim uma evolução no modelo ao longo do tempo tal como acontece com o próprio ensino.

1.3 Estrutura do documento

Este relatório encontra-se dividido em cinco capítulos em que o atual, capítulo 1 descreve o contexto do trabalho realizado. Os restantes capítulos estão organizados da seguinte forma:

- **Capítulo 2** – Trabalho Relacionado
- **Capítulo 3** – Análise de Dados
- **Capítulo 4** – Modelação com cadeias de Markov
- **Capítulo 5** – Conclusões

Capítulo 2

Trabalho relacionado

No trabalho relacionado vamos descrever o problema do fluxo de estudantes no ensino e estudar abordagens possíveis e realizadas focando nas que recorrem a processos estocásticos.

2.1 O Problema do Fluxo de Estudantes

A educação abrange vários setores que uma pessoa percorre na sua infância e juventude, desde o jardim de infância, ensino fundamental e médio, até ao ensino superior. O setor educacional pode ser visto como uma série de componentes onde cada aluno segue um caminho desejado que vai ao encontro às suas próprias aspirações (Tavares, 1995).

Muitos exemplos de modelos matemáticos aplicados ao planeamento da educação existem desde os finais dos anos de 1960 (Lovell 1971). Neste problema do fluxo de estudantes identificámos várias abordagens: KPI (Key performance indicators), visualização, cadeias de Markov e simulações hipotéticas. Estas abordagens estão ordenadas pela sequência usual da análise de dados.

No estudo de análise de sobrevivência aplicado ao fluxo de estudantes, dois indicadores-chave de desempenho são destacados: abandono e a retenção (Junior et al, 2012). O conhecimento e possível previsão destes indicadores permitem-nos avaliar o desempenho de um sistema de ensino, destacando-se assim como relevantes no estudo do fluxo de alunos.

A representação visual do caminho de um aluno até à sua formatura é idêntica a outros tipos de fluxo já explorados desde finais do século XIX (Heileman et al, 2015). As vantagens da análise visual permitem-nos explorar uma grande quantidade de dados rapidamente facilitando a deteção de anomalias e reconhecimento de tendências. São também uma forma de obter provas visuais do fluxo do aluno, em vez de especulações de como o aluno comum percorre o seu caminho de aprendizagem.

É apetecível a visualização do sistema educacional como um processo de Markov quando apresentados com a simplicidade de definir os estados de um aluno, no ensino ou fora (Bessent & Bessent, 1980). A observação de propriedades estocásticas promove-nos ao cálculo das probabilidades de transição utilizando os dados disponíveis para a modelação do problema e potencialmente obter previsões do fluxo de alunos.

Com simulações hipotéticas, obtemos a possibilidade de observar e testar o impacto de alterações políticas curriculares, estruturas e/ou recursos humanos no comportamento de um sistema (Fiallos & Ochoa, 2017; Saltzman & Roeder, 2012).

Vários estudos realizados revelam que os progressos dos alunos nas suas educações mostram diversas características estocásticas o que nos incentiva e compele à utilização de cadeias de Markov para a modelação do fluxo dos estudantes e várias abordagens já foram realizadas.

Em 2015, Brezavšček, A. e Baggia, A. desenvolveram um modelo do tipo cadeia de Markov já foram utilizados para a análise do fluxo de alunos de um curso na Universidade de Maribor. Definiram vários estados possíveis para cada aluno como o ano em que se encontrava, se se encontrava no momento inativo do curso, se desistiu do mesmo e se já se graduou do curso. Uma característica de notar destes estados são os estados de graduação e de desistência que não transitam para outro estado a não ser o próprio, ou seja, são estados absorventes (Brezavšček & Baggia, 2015).

Após esta definição criaram uma matriz de transição e obtiveram que transições entre estados eram possíveis, isto é, que a probabilidade de transição não fosse igual a 0.

Utilizaram como *dataset* um intervalo de 6 anos letivos consecutivos (2008/09 a 2014/15). Com este *dataset* realizaram os seguintes passos para cada ano letivo:

- Para cada estado do sistema, fizeram a contagem de quantos alunos faziam cada transição definida anteriormente;
- Dividiram a contagem da cada transição pelo somatório de contagens de transições realizadas por esse estado nesse ano letivo, obtendo assim valores que representam as probabilidades de um aluno naquele estado transitar para outros naquele mesmo ano letivo.

Tendo as várias probabilidades de transição de vários anos letivos, calcularam a probabilidade média de cada transição criando assim uma matriz de transição com os valores obtidos.

Em 2017 (Brezavšček, Pejić Bach & Baggia, 2017) continuaram o trabalho já realizado em 2015, agora com um *dataset* de um intervalo de 8 anos letivos consecutivos do mesmo curso utilizado anteriormente.

Para conseguirem prever o número de alunos em anos futuros calcularam a distribuição dos alunos pelos estados definidos no último ano do seu *dataset* e também a distribuição prevista para os anos futuros ao aplicar a matriz de transição à distribuição do último ano. No caso do primeiro ano do curso, em que a maioria dos alunos que estão neste estado são ingressos fora do sistema, assumiram, que por cada ano académico, entram no curso cerca de 120 alunos novos (Brezavšček et al, 2017). Este caso é bastante importante para o projeto que se quer desenvolver porque é bastante semelhante ao caso dos alunos que estão inscritos no primeiro ano curricular do ensino português.

Na Universidade do Kuwait (Alawadhi & Konsowa, 2010), com um *dataset* de 9 anos académicos (1996-1997 até 2004-2005), tal como no estudo realizado na Universidade de Maribor, fizeram um estudo ao fluxo de estudantes na universidade onde foram definidos estados, incluindo estados absorventes para a desistência e graduação do curso e posteriormente o cálculo de frequências para os estados definidos.

A utilização de estados absorventes tem sido uma abordagem comum nestes estudos e que já é utilizada há bastantes anos (Bessent & Bessent, 1980). No estudo realizado na Universidade do Texas, já era utilizado um estado absorvente para a representação da graduação do curso. Neste estudo ponderaram também um outro estado absorvente para a desistência forçada, mas que não foi incluído pela probabilidade ser quase nula. Portanto e também para outros estados transitórios possíveis, foram excluídos para a simplificação do modelo.

Num estudo ao fluxo de estudantes de alunos na UUM (Universiti Utara Malaysia) (Rahim, Ibrahim, Kasim & Adnan, 2013) fizeram uma divisão dos dados em coortes, em que utilizaram como critério de distinção a data de início dos estudos do aluno. Existe também, a preocupação de medir a precisão do modelo obtido. Para tal, utilizaram o MAPE (*mean absolute percentage error*) como métrica. Compararam os resultados obtidos das coortes com os dados do primeiro coorte (coorte que começou os estudos há mais tempo) e obtiveram valores para o MAPE abaixo dos 10%.

2.1.1 Ensino português

Em Portugal, o sistema de educação está dividido em várias categorias: educação pré-escolar, ensino básico, ensino secundário, ensino superior e ainda educação e formação de jovens e de adultos (DGEEC, 2021).

Desde finais dos anos 50 que os primeiros quatro anos de ensino são obrigatórios, mas esta obrigatoriedade escolar foi alargada para todos os jovens com idades até os 18 anos nos últimos anos (Decreto-Lei no 176/2012; Lei no 85/2009). Um estudo no Canadá (Oreopoulos, 2005) para observar se o ensino obrigatório teve impacto no sucesso dos alunos mostra que esta obrigatoriedade teve impactos significativos e positivos. O número de alunos a obter graduação aumentou foi estimado a obter um aumento de 13%. Esta obrigatoriedade escolar poderá ter impactos significantes notáveis nos nossos dados e é necessária uma análise dos mesmos tendo em conta uma possível tendência devido a essa obrigatoriedade.

A educação pré-escolar, também conhecido como educação de infância, é dedicada para crianças com menos de 6 anos.

O ensino básico é o começo da escolaridade obrigatória e pode ser dividido em três ciclos:

- Primeiro ciclo - 1º, 2º, 3º e 4º ano curricular;
- Segundo ciclo - 5º e 6º ano curricular;
- Terceiro ciclo - 7º, 8º e 9º ano curricular.

Com o terceiro ciclo completado, os alunos entram no ensino secundário, a última etapa do ensino obrigatório, onde os alunos podem escolher de diversos cursos o que desejam de aprofundar o seu conhecimento em. Este ensino final, da obrigatoriedade escolar, tem uma duração de 3 anos. No entanto, para esta última fase existe uma enorme diversidade de tipos de ensino que serão compreendidas para um melhor entendimento dos nossos dados:

- Cursos Científico-Humanísticos – O percurso mais comum que os alunos percorrem após a aprovação do 9º ano curricular. Este tem duração de três anos letivos que correspondem ao 10º, 11º e 12º ano curricular e é conhecido pelos quatro possíveis cursos: Ciências e Tecnologias, Ciências Socioeconómicas, Línguas e Humanidades e Artes Visuais;

- Cursos com Planos Próprios – Também conhecidos como Cursos Científico-Tecnológicos, são cursos que permitem ao aluno obter uma formação científica e uma formação tecnológica;

- Cursos Artísticos Especializados – Para jovens que mostram aptidão e/ou talento para uma ou mais das diversas áreas artísticas;

- Cursos Profissionais – Que facilitam a inserção do aluno no mercado de trabalho através de uma preparação profissional para o mesmo.

Após o ensino secundário, os que assim desejarem, podem entrar no ensino superior para obterem um grau de ensino superior como é o caso de licenciaturas, mestrados e doutoramentos.

Existem ainda outros cursos para diversas outras situações de jovens e de adultos: Cursos de Educação e Formação (CEF), Cursos de Especialização Tecnológica (CET), cursos de Educação e Formação de Adultos (EFA), Cursos vocacionais, ensino recorrente, Formações Modulares Certificadas (FMC), Percursos Curriculares Alternativos (PCA), processos de reconhecimento, validação e certificação de competências (RVCC) e programas integrados de educação e formação (PIEF). A grande diferença entre estes últimos mencionados são a sua menor frequência em relação aos cursos que representam o ensino secundário e de serem mais apropriados para pessoas em casos específicos de formações incompletas.

Para um determinado ano letivo, os alunos inscritos num determinado ano curricular n podem ser observados de três fontes diferentes (Cavique et al, 2020):

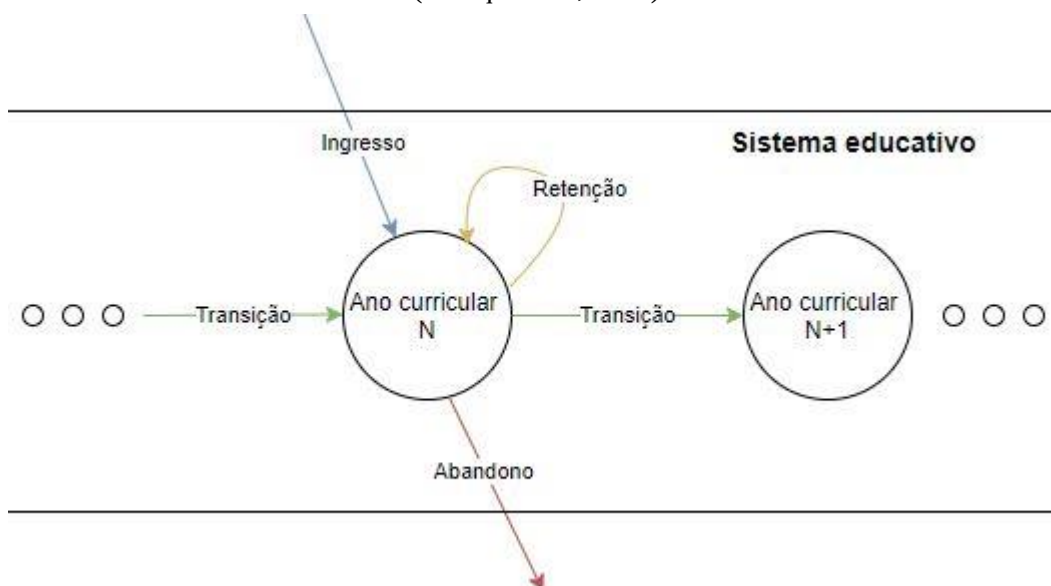


Figura 1: Fluxo de alunos para ano curricular N

- Ingressos de jovens fora do sistema educativo português. No ano letivo anterior, o jovem não está inscrito em nenhum ano curricular e pretende ingressar no sistema de educação. Temos como exemplos deste caso: as crianças que se inscrevem para o primeiro ano curricular, jovens imigrantes, etc.
- Transições de jovens que tenham completado o ano curricular em que se encontravam no ano letivo anterior. Quase todos os casos deste tipo são completações do ano curricular $n - 1$.
- Retenções de jovens que não tenham conseguido completar o ano curricular com sucesso ou alunos que tenham trocado de curso.

No final de um ano letivo, um aluno, se conseguir obter aprovação do ano curricular corrente n , transitará para o ano curricular $n + 1$. Se não conseguir obter aprovação, este irá se manter no mesmo ano curricular. Existe também a possibilidade de este descer de ano curricular em certos casos específicos tal como por exemplo a mudança de curso profissional.

2.2 Processos Estocásticos

Neste capítulo serão aprofundadas definições e conceitos do contexto do trabalho realizado, metodologias abordadas e seus desenvolvimentos ao longo do período de trabalho. A primeira metodologia abordada foi a cadeia de Markov que é um caso particular dos processos estocásticos e, portanto, será primeiramente aprofundado o que são estes processos.

Um processo estocástico é um conjunto de variáveis aleatórias X_t , em que t pode ter um valor de um conjunto de índices que pode ser contínuo ou discreto (Dehling, Gottschalk & Hoffmann, 2007). Este parâmetro pode representar diversos valores como por exemplo distâncias ou número de vezes que uma ação foi realizada, mas este parâmetro é na maior parte das vezes utilizado como um parâmetro temporal que será também o significado dele neste trabalho.

Processos estocásticos são também definidos pelo espaço de estados que o conjunto de variáveis pode tomar. Mas ao contrário de um processo determinístico em que é possível determinar como o conjunto de variáveis se irá apresentar, ou seja, quais os valores que as variáveis irão tomar num determinado instante t , num processo estocástico não se terá garantias de quais os valores que o conjunto irá tomar, mas sim probabilidades de diversas apresentações/distribuições do conjunto (Dehling et al, 2007).

Existem diversos tipos de processos estocásticos e para os classificar são utilizadas as cardinalidades do conjunto de estados e do parâmetro t . O conjunto de estados ou o parâmetro t podem tomar valores discretos (valores finitos ou enumeráveis) ou valores contínuos (Ching, & Ng, 2006). No caso deste trabalho, pretende-se que o conjunto de variáveis represente os alunos no ensino português e, portanto, os valores que estes podem tomar irá ser o ano curricular em que estão matriculados (valor discreto) para um ano letivo n (valor discreto) que será utilizado como parâmetro t do nosso processo estocástico.

Decidiu-se também explorar, em mais detalhe, os modelos com a propriedade de Markov. A propriedade de Markov é uma propriedade de condicionalidade das probabilidades que pertence a um processo estocástico se os valores/estados futuros das variáveis dependerem apenas do estado atual e não de nenhum valor/estado passado da variável. Um processo estocástico é classificado como um processo Markoviano se o processo contém esta característica da propriedade de Markov (Bryc, 1996; Ching, & Ng, 2006; Häggström, 2002; Meyn, & Tweedie, 2012).

2.2.1 Processo Markoviano

Este tipo de processos é também denominado de processo sem memória (Konstantopoulos, 2009). Esta denominação vem da propriedade Markov onde os acontecimentos/valores anteriores ao estado em questão são “ignorados” (desprezados).

Portanto as probabilidades deste tipo de processo podem ser representadas formalmente como:

$$P\{X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1, X_0 = x_0\} = P\{X_{t+1} = x_{t+1} | X_t = x_t\}$$

A expressão traduzida para uma forma menos informal: A probabilidade condicional de um evento futuro de uma variável depende só da probabilidade anterior (presente se for o caso do futuro imediato, isto é, do próximo valor de t) e não dos valores que a variável tomou no passado.

Existem então probabilidades de transição entre qualquer valor do espaço de estados para qualquer outro valor. Se o conjunto de estados for finito e discreto iremos ter n^2 probabilidades

sendo n o número de estados. Para melhor observação e utilização destas probabilidades pode-se dispor estas probabilidades numa matriz M de dimensão $n \times n$, denominada de matriz de transição.

Sendo $P\{X_{t+1} = j | X_t = i\} = p_{ij}$, probabilidade de a variável X transitar para o estado j na iteração $t + 1$, quando está no estado i na iteração t , a matriz de transição irá ter a seguinte forma:

$$M = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

Uma observação que deve ser feita destas probabilidades de transição é que a soma das probabilidades que estão situadas na mesma linha da matriz é igual a 1 (Häggström, 2002). Isto acontece, pois, apesar da transição não ser determinada, sabemos que uma variável irá se transitar para um dos estados definidos e, portanto, podemos afirmar como verdadeira a seguinte equação para qualquer valor de i (linha da matriz de transição):

$$\sum_{j=1}^n p_{ij} = 1$$

Sendo k a dimensionalidade do nosso espaço de estados, i.e., cada estado é representado por k variáveis, podemos criar uma distribuição das nossas variáveis pelos valores possíveis obtendo assim uma distribuição S no seguinte formato:

$$S_t = [X_t(1) \quad X_t(2) \quad \cdots \quad X_t(k-1) \quad X_t(k)]$$

Podemos então utilizar a matriz de transição à matriz das variáveis para obter a distribuição dos valores para o próximo valor de t da seguinte forma:

$$S_{t+1} = S_t \cdot M$$

2.2.2 Cadeia de Markov

A um processo markoviano dá-se o nome de cadeia de Markov tipicamente quando o espaço de estados é finito tal como o parâmetro temporal t (Dehling et al, 2007).

Com estados discretos é possível criar um diagrama para representar este processo markoviano. Temos de seguida um exemplo de um diagrama para uma cadeia de Markov com 3 estados possíveis (1,2 e 3):

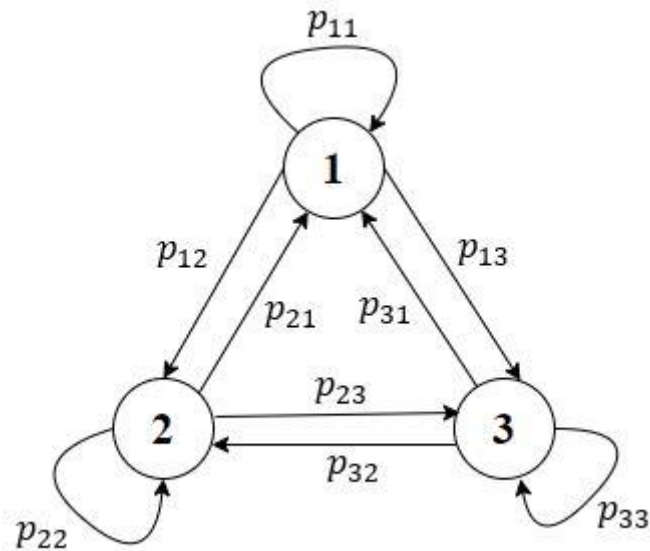


Figura 2: Diagrama exemplo de cadeia de Markov

Dentro das cadeias de Markov, existem diversos tipos de cadeias e propriedades que as definam.

Quando temos uma distribuição S que ao longo do tempo não varia, isto é, que ao multiplicar a distribuição pela matriz de transição resulta na mesma distribuição então está-se perante uma distribuição estacionária (Meyn, & Tweedie, 2012).

Uma propriedade que as probabilidades podem ter que também é importante referir para este trabalho é o caso de quando uma probabilidade p_{ii} (probabilidade de transição para o próprio estado) é igual a 1, o que causa as restantes probabilidades p_{ij} com $i \neq j$, serem iguais a 0. Quando isto acontece, estamos perante a um estado absorvente. Um estado absorvente é um estado do qual quando uma variável toma este valor, a variável irá tomar sempre esse valor pois não existem nenhuma transição possível a não ser para o próprio estado atual.

Capítulo 3

Análise de Dados

Para a criação do modelo recolhemos dados dos alunos inscritos no ensino português entre o ano letivo 2008/2009 e 2018/2019. Estes dados, fornecidos pela DGEEC ao projeto ModEst, foram retirados da sua base de dados e colocados em ficheiros do tipo .xlsx e .csv para futura análise utilizando diversas ferramentas de bibliotecas Python. Destes dados é importante observar várias características em particular para se poder compreender a tendência destes valores e conseguir obter valores concretos para as probabilidades de transição entre estados do modelo pretendido.

O nosso fornecedor de dados, DGEEC, tem uma vasta experiência na coleção e análise de dados de alunos portugueses ao recolher informações de mais de um milhão de alunos por ano letivo. Estes dados respeitam obrigatoriamente o regulamento geral de proteção de dados (RGPD) e estão anonimizados fornecendo-nos zero informações pessoais de identificação (Magalhães & Pereira, 2018).

Este trabalho foca-se nos dados dos alunos que se encontram do primeiro ao décimo segundo ano curricular, deixando dados do ensino superior para possível continuação deste trabalho. Mesmo filtrando alguns alunos do *dataset*, este fica com uma média de um milhão e meio de alunos por ano letivo.

3.1 Valores absolutos dos dados

O número de alunos inscritos no sistema educativo português pelos diversos anos curriculares é uma importante informação para qualquer tipo de probabilidade que se pretenda a ser analisada e calculada.

Para as transições pretendidas entre os estados do nosso modelo de Markov é necessário observar valores absolutos e relativos (em relação ao número total de inscritos) dos resultados que um aluno pode obter após um ano letivo. Estes valores são o número de alunos que obtiveram aprovação no ano curricular que estão inscritos, o número de alunos que reprovaram o ano curricular em que encontram inscritos e os que abandonaram o sistema educativo.

Um outro dado importante a analisar são os ingressos dos alunos para o sistema educativo. Esta informação irá alimentar o nosso sistema de forma que o número de alunos nos primeiros anos curriculares previstos pelo modelo seja o mais correto possível pois o número de alunos no primeiro ano curricular irá ser, em grande maioria, ingressos de crianças que não se encontravam no sistema educativo (crianças que se encontravam em educação pré-escolar).

Para conseguir calcular as probabilidades/percentagens das transições entre estados, ou mesmo para observar possíveis tendências do número total de alunos, é necessário observar o número de alunos que está inscrito em cada ano letivo e por cada ano curricular. Para tal, um primeiro passo é o cálculo das frequências de cada ocorrência para cada estado (Brezavšček & Baggia, 2015; Alawadhi & Konsowa, 2010; Shah & Burke, 1999). Criamos tabelas de frequência para cada tipo de transição entre estados do nosso modelo, (transição, retenção e abandono), e concebemos gráficos para uma visualização dos dados. Estas tabelas e gráficos encontram-se no apêndice A e numa breve análise, observa-se que existem claras diferenças entre anos curriculares.

A primeira observação que fazemos do gráfico de inscritos do apêndice A, é de que existe um enorme número de inscritos nos primeiros anos (2008-2010) do *dataset* para o nono e décimo segundo ano curricular. Este número de inscritos diminui drasticamente durante os primeiros 4 anos, mas sugere-nos questões sobre a qualidade destes dados. Observamos um número enorme de inscritos para estes dois anos curriculares, mas o que acontece a estes alunos? Para os que se encontravam no décimo segundo ano curricular poder-se-ia argumentar que a maioria deles se graduaram e completaram o ensino. No entanto, o que acontece aos alunos que se encontravam no nono ano curricular? Neste mesmo gráfico, não se observa que um grande número de inscritos no próprio ano curricular nem no próximo ano curricular nos próximos anos letivos. É devido a irregularidades como esta que se deu uma grande atenção à análise e compreensão dos dados e identificação de dados ruidosos pois queremos ter a certeza de que os nossos dados nos são úteis para o nosso objetivo. Analisando os outros gráficos, notamos que existe algo notável para o nono ano curricular. Esta diferença de valores observada no gráfico de inscritos, também se pode observar nos gráficos correspondentes aos ingressos, retenções e abandonos deste apêndice, mas curiosamente não de forma significativa no gráfico de transições.

Devido à enorme diferença entre percursos dos alunos, criamos dois *subdatasets* divididos pela modalidade em que o aluno se encontra. Um *dataset* que contém os percursos mais comuns das crianças que ingressam no ensino português no primeiro ano curricular. Este *dataset* define-se por conter as modalidades:

- Cursos gerais;
- Cursos artísticos especializados;
- Cursos científico-humanísticos;
- Cursos com planos próprios.

O outro *dataset* irá conter as restantes modalidades não mencionadas em que não representam o curso mais comum de uma criança que se tenha ingressado no ensino português no primeiro ano curricular e/ou os dados não são tão consistentes como no primeiro *dataset*. Casos como um aluno realizar um curso de 3 anos e nos dados esse aluno aparecer sempre no mesmo ano curricular durante 3 anos e no fim graduar obriga-nos a realizar esta separação para que seja possível obter um modelo com resultados significantes e reais. No entanto será benéfico analisar estes dados de qualquer forma e concebemos os mesmos tipos de tabelas e gráficos para este *dataset* que podem ser encontrados no apêndice B.

Devido às diversas irregularidades deste *dataset*, não é possível retirar muitas conclusões, no entanto, observa-se também a descida de inscritos até o ano letivo de 2012/2013 e uma “estabilização” desse número nos seguintes anos letivos.

Para o *dataset* com os percursos mais comuns, geramos, do mesmo modo, tabelas e gráficos que se encontram no apêndice C. Observa-se claramente que existe uma descida do número de

alunos inscritos do primeiro ao sétimo ano curricular ao longo deste intervalo de doze anos letivos. Finalmente nos anos curriculares correspondentes ao ensino secundário observam-se para cada um destes anos, um número relativamente constante de alunos inscritos ao longo do tempo. Nos ingressos, nota-se a falta de muitos ingressos nos anos curriculares do nono e décimo segundo curricular devido à separação do *dataset* o que é de esperado porque são nas outras modalidades que existem ingressos em cursos que equivalem ao nono e décimo segundo ano curricular.

3.2 Análise de tendências

Como as probabilidades de cada acontecimento serão um aspeto importante do modelo de Markov, realizámos também análises às evoluções das probabilidades ao longo do tempo. Para cada acontecimento possível e para cada ano curricular criámos gráficos com os valores relativos. Estes gráficos encontram-se no apêndice D.

Observando os valores relativos nos gráficos construídos para cada tipo de possível acontecimento (transição, retenção e abandono) podemos retirar algumas conclusões que nos permitirão orientar alguns passos de desenvolvimento e evolução do nosso modelo de predição.

Analisando os gráficos relativos às probabilidades de transição, os gráficos anos curriculares correspondentes ao primeiro ciclo são ligeiramente ruidosos e sem tendência que seja detetada numa primeira observação. O primeiro ano curricular mostra valores relativamente constantes e os restantes deste ciclo, valores finais superiores aos iniciais. Os próximos 4 anos curriculares (5º ao 8º) e 10º ano curricular mostram resultados interessantes. Valores com uma tendência positiva, isto é, atenuação do decréscimo por ano letivo para um crescimento desta probabilidade com concavidade para cima e valores mínimos nos anos letivos entre 2011-2013. Estas tendências podem muito provavelmente ser reflexões das legislações mencionadas (Decreto-Lei no 176/2012; Lei no 85/2009). Nos restantes anos curriculares (9º, 11º e 12º), as probabilidades crescem linearmente. De seguida, a análise dos gráficos correspondentes às probabilidades de retenção. A grande parte destes gráficos mostram uma tendência oposta à observada nos gráficos anteriores, isto é, com uma concavidade voltada para baixo e com valores máximos também no intervalo de anos letivos 2011-13. No caso dos abandonos, observa-se um ligeiro declínio nos primeiros anos letivos dos nossos dados e uma estabilização destes valores nos últimos anos letivos.

Como a probabilidade de transição entre estados é um aspeto importante das cadeias de Markov, estas análises levantam questões sobre estes valores como, por exemplo, que tipo de curvas estes valores mostram, para que valores tendem e quais os fatores que influenciam a variação dos mesmos.

3.3 Sobreposição de curvas

Como no ensino português existem definições como disciplinas obrigatórias e carga horária semanal mínima para cada ciclo (Decreto-Lei no 55/2018), uma investigação que poderia ser potencialmente revelador sobre o fluxo seriam seria se as tendências estudadas anteriormente de um ano curricular arbitrário mostram semelhanças visuais com as curvas do mesmo ciclo (intra-ciclo) e se existem distinções em relação às tendências dos anos curriculares dos outros ciclos (inter-ciclo).

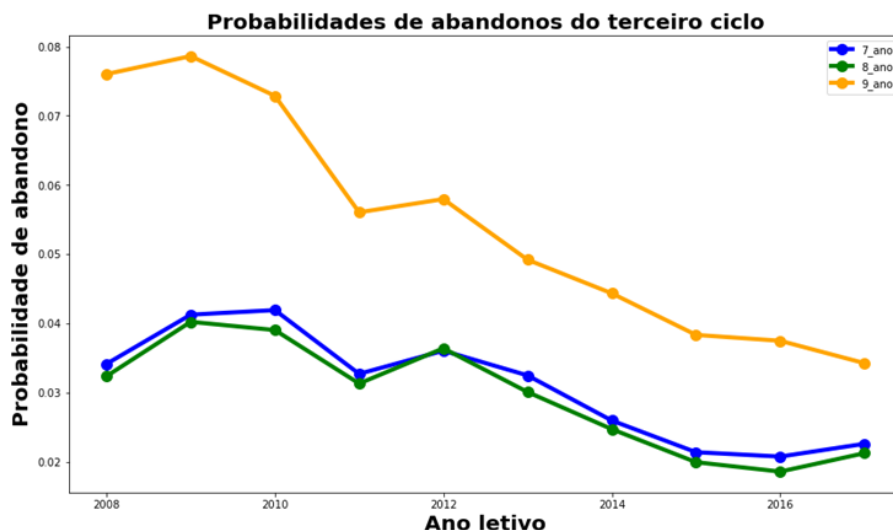


Figura 3: Curvas das probabilidades de abandono do terceiro ciclo

Mas para tal é necessário primeiro utilizar um algoritmo para sobreposição de curvas. Se queremos avaliar se duas curvas são semelhantes, necessitamos de ter as curvas sobrepostas e avaliar as diferenças entre elas. Temos, como por exemplo, na figura 3, curvas de probabilidades de abandono do terceiro ciclo, em que apesar da curva do nono ano (curva amarela) ter valores superiores em relação aos outros dois anos letivos, mostra algumas semelhanças com as restantes curvas quando sobreposta como podemos observar na figura 4.

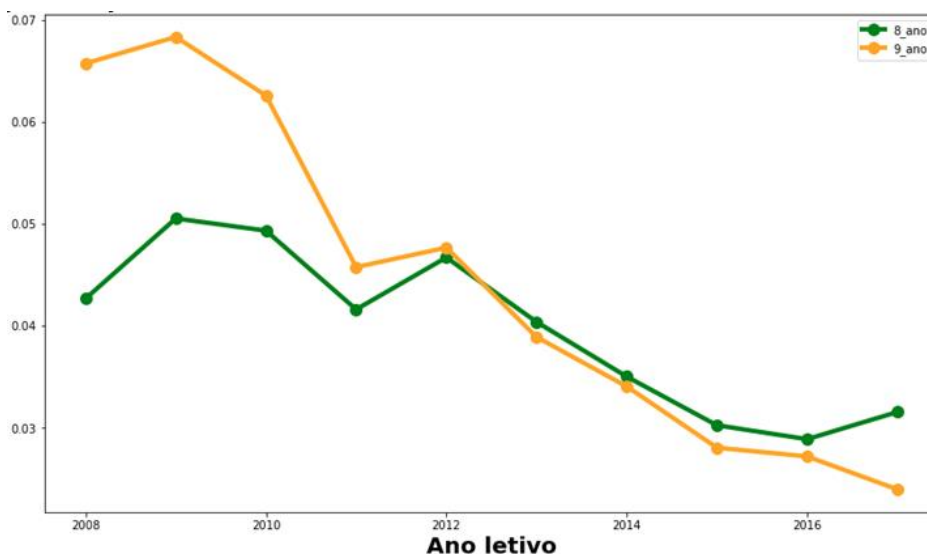


Figura 4: Exemplo de como duas curvas mostram semelhanças quando sobrepostas

Desenvolvemos um algoritmo para cálculo de um offset a aplicar sobre 2 curvas para a minimização de erro/distância entre pontos com o mesmo ano letivo (variável independente).

Começamos por calcular a diferença do valor (probabilidade/variável dependente) de cada ponto da curva A menos o valor do ponto correspondente da curva B e guardar esse valor numa lista. Obtemos assim os valores que teríamos de retirar a cada ponto da curva A para que o mesmo ficasse sobre o ponto da curva B ou, no caso contrário, adicionar aos pontos da curva B para sobrepor a curva A (exemplo para o primeiro ponto: a diferença será igual a $8 - 9 = -1$. Podemos retirar esta diferença ao valor da curva A para obter o valor da curva B: $8 - [-1] = 9$ ou adicionar esta diferença ao valor da curva B para obter o valor da curva A: $9 + [-1] = 8$).

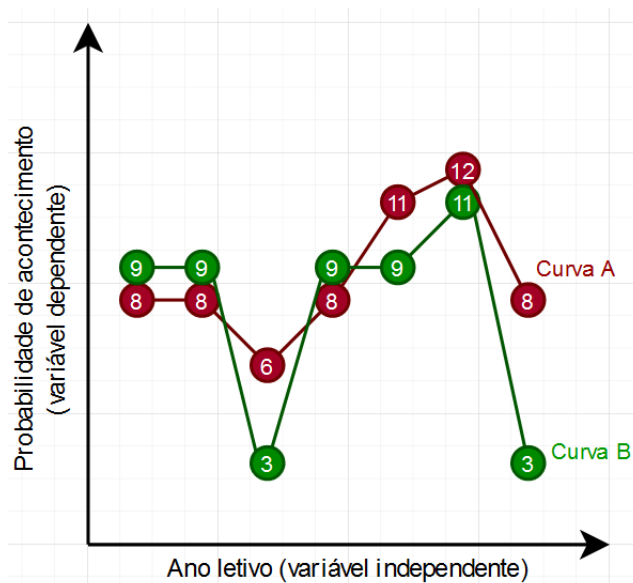


Figura 5: Exemplo ilustrativo de 2 curvas semelhantes: curva A a vermelho e curva B a verde

Com a lista das diferenças entre cada ponto das curvas, ordenamos essa lista por ordem crescente. O valor para ser utilizado como offset será o valor mediano desta lista que iremos denominar como m . Após cálculo das diferenças entre as curvas com o offset aplicado, qualquer outro valor utilizado como offset irá obter uma soma das diferenças maior em valor absoluto. Para provar esta afirmação apresentamos a seguinte prova:

Diferenças	Diferenças ordenadas	Diferenças após offset	Diferenças ordenadas após offset
$8 - 9 = -1$	-1	$8 - 1 - 9 = -2$	-2
$8 - 9 = -1$	-1	$8 - 1 - 9 = -2$	-2
$6 - 3 = 3$	-1	$6 - 1 - 3 = 2$	-2
$8 - 9 = -1$	1	$8 - 1 - 9 = -2$	0
$11 - 9 = 2$	2	$11 - 1 - 9 = 1$	1
$12 - 11 = 1$	3	$12 - 1 - 11 = 0$	2
$8 - 3 = 5$	5	$8 - 1 - 3 = 4$	4
	(14)		(13)

Legenda:

Valores da Curva A

Valores da Curva B

Valor mediano/offset

(Soma das diferenças absolutas)

Figura 6: Cálculo de diferenças e diferenças após aplicação de offset

Se não escolhermos por utilizar o valor mediano como offset, temos dois tipos de opções:

- Escolher um valor para offset superior ao valor mediano;
- Escolher um valor para offset inferior ao valor mediano.

Se considerarmos n o número de diferenças antes ou depois do valor mediano (o número de diferenças será então $2n + 1$): Ao escolher um valor superior a m , podemos observar um agravamento das diferenças. Isto acontece porque as diferenças antes da mediana e a própria mediana, $n + 1$ casos, irão obter resultados piores pela natureza de estes valores serem ≤ 0 enquanto nos restantes n casos, em que os valores são de natureza ≥ 0 , poderão obter, no melhor dos casos em que nenhum valor desce dos 0 como acontece na figura 7, valores inferiores.

No entanto, mesmo no melhor dos casos mencionado, o número de diferenças afetadas

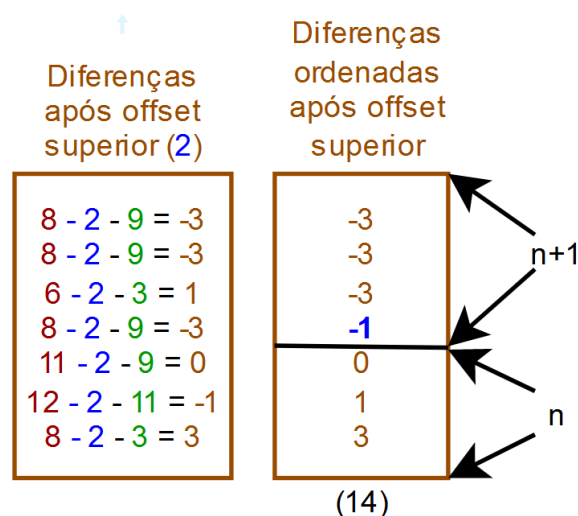


Figura 7: Alteração do offset mediano obtém piores resultados

negativamente é superior ao de afetadas positivamente resultando assim numa soma de diferenças dos pontos superior à obtida com o valor mediano concluindo assim de que não será benéfico utilizarmos um valor superior ao valor mediano das diferenças iniciais. Utilizando esta lógica, também se prova de que a utilização de um valor inferior irá causar degradação da soma, concluindo assim de que o valor mediano será o valor ótimo para minimizar a soma da diferença entre pontos de variável independente igual de duas curvas discretas.

Para podermos definir o comportamento típico para cada ciclo, foram criadas curvas médias para cada um dos ciclos utilizando as curvas dos anos curriculares correspondentes. Após a criação dessas curvas, foram sobrepostas as curvas de cada ano curricular para avaliação de semelhança das curvas.

Como métrica destas semelhanças decidiu-se utilizar uma variante das distâncias de Fréchet (Eiter, & Mannila, 1994; Wylie, 2013). As distâncias de Fréchet discretas são uma alternativa à métrica tradicional que se adapta facilmente a aproximações poligonais de curvas como temos no nosso estudo para as curvas de probabilidades de transições entre estados. Esta métrica é utilizada para avaliar a semelhança entre curvas, no entanto, esta é igual à distância mínima necessária entre dois pontos que percorram as duas curvas simultaneamente, mas sem restrições de velocidade e/ou distância percorrida na própria curva. Infelizmente, o algoritmo

desenvolvido anteriormente, não é eficiente a minimizar a distância de Fréchet entre as duas curvas, pois foi desenvolvido apenas com a minimização de erros em mente. Apesar deste erro, é um algoritmo que consegue sobrepor duas curvas de maneira benevolente e, portanto, a partir de resultados com valores similares não será possível retirar conclusões concretas.

Produzimos os gráficos apresentados no apêndice E, onde podemos comparar as distâncias de Fréchet de cada ano curricular com as curvas médias de cada ciclo para cada tipo de transição entre estados. Nestes gráficos, colocámos as barras referentes às distâncias de Fréchet às curvas dos próprios ciclos de cada ano curricular para uma interpretação facilitada. Algo que se observa em quase todos estes gráficos, com poucas exceções é a que as distâncias intra-ciclos são menores que as inter-ciclos. Este resultado era esperado visto que os anos curriculares em cada ciclo partilham bastantes disciplinas e carga horária (Decreto-Lei no 55/2018).

Visto que os anos curriculares de cada ciclo partilham várias características, uma possível abordagem à modelação do fluxo de estudantes no ensino português seria agrupar os anos curriculares de cada ciclo num único estado que representaria o ciclo a que esse ano curricular pertence.

Capítulo 4

Modelação das Cadeias de Markov

4.1 Modelo proposto

No âmbito do que é pretendido deste trabalho, utilizaremos como modelo uma adaptação das Cadeias de Markov. Como queremos modelar o fluxo de alunos no ensino português, será bastante natural a criação de doze estados, um para cada ano curricular. Uma outra abordagem mais genérica seria criar um estado para cada ciclo, no entanto iremos explorar a abordagem mais detalhada. Iremos também considerar dois estados adicionais, neste caso ambos absorventes, que irão também conter dados relevantes. Tal como já foi visto nos estudos anteriores (Brezavšček & Baggia, 2015; Brezavšček et al, 2017; Alawadhi & Konsowa, 2010; Bessent & Bessent, 1980), é frequente a utilização de um estado que irá representar a conclusão do ensino que na maioria dos casos dos alunos que frequentam o ensino obrigatório serão os que obtêm aprovação no décimo segundo ano curricular. Também iremos considerar um estado para as desistências do ensino. Foram também consideradas várias alternativas para o número e representação de estados como por exemplo:

- Um estado para a representação de desistência temporária para cada ano curricular em que teria uma probabilidade de transição para o estado do ano curricular correspondente;
- Um ou vários estados para a representação do estado de fora do ensino que iria utilizar os valores dos ingressos para o cálculo das probabilidades de transição para os estados pertencentes ao ensino português.

Para o cálculo de probabilidades necessitamos do número de casos favoráveis e do número total de casos (Correa, 2003). O nosso *dataset* contém os ingressos nos diversos anos curriculares, mas defrontamo-nos com um problema: Qual é o número total de casos para o cálculo desta probabilidade? Todos os jovens fora do ensino? Os jovens que desistiram do ensino? Não existe uma resposta certa que nos permite obter um estado que represente “fora do ensino escolar” e, portanto, os ingressos irão serem trabalhados como uma adição ao nosso sistema. Estes ingressos irão ter uma especial atenção para que seja possível obter um modelo com resultados relevantes.

Assim, decidimos criar o nosso modelo com 14 estados possíveis (12 para os anos curriculares, 1 para a conclusão do ensino e 1 para a desistência/abandono) (ver fig. 8). No entanto, é necessário de imediato decidir como o nosso modelo irá lidar com os valores de ingresso. Como será já averiguado no próximo capítulo, uma grande percentagem dos alunos inscritos no primeiro ano curricular são alunos que acabaram de ingressar no ensino. A estratégia utilizada para nos resolver este problema dos ingressos foi adaptar o nosso modelo para que em vez de

termos uma distribuição S que nos representa de como os alunos estão distribuídos pelos estados, utilizar valores absolutos para representar quantos alunos estão em cada estado.

Esta abordagem não só nos permite obter os dados mais relevantes sem tratamento (multiplicar pelo número total de alunos no modelo) mas também nos facilita a incorporar os ingressos. Ao trabalhar com valores absolutos, no final de cada ano letivo, podemos observar como os alunos transitam entre anos curriculares e adicionar o número de ingressos para obter o número de alunos no próximo ano letivo. Finalmente, sendo n_x o número de alunos no x -ésimo ano curricular, n_c o número de alunos que completam o ensino e n_a o número de alunos que abandonam, iremos representar o estado de cada ano letivo S da seguinte forma:

$$S = [n_1 \ n_{12} \dots \ n_{11} \ n_{12} \ n_c \ n_a]$$

Assim ao aplicar a matriz de transição à distribuição, obtemos previsões de quantos alunos estão em cada estado na próxima iteração (ano letivo). Sendo I_k , o número de ingressos no ano letivo k e com uma distribuição S de valores absolutos para o ano letivo k , podemos assim obter a distribuição do ano letivo $k + 1$ através da seguinte fórmula:

$$S_{k+1} = S_k \cdot MT_k + I_{k+1},$$

sendo MT_k a matriz de transição para o ano k .

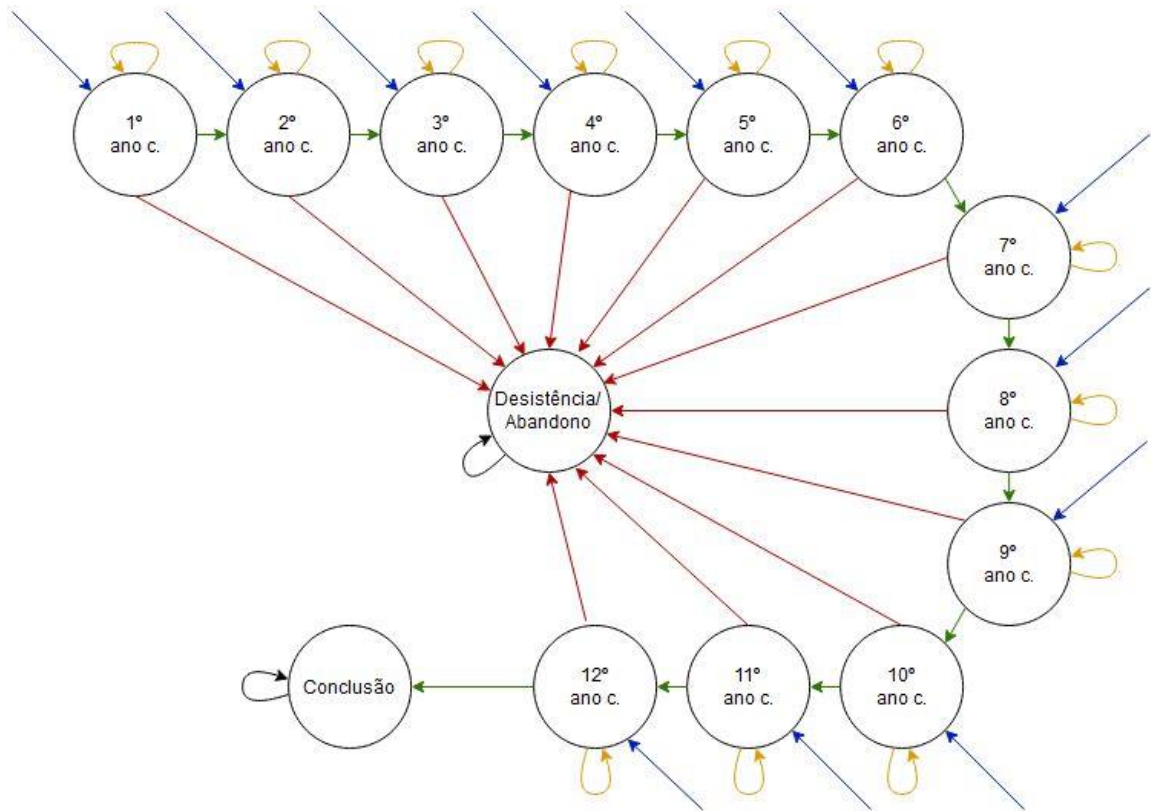


Figura 8: Diagrama da cadeia de Markov pretendida

4.2 Modelo base

Com a análise e compreensão de dados realizada podemos então começar a criar e desenvolver um modelo para a previsão de dados. Antes de desenvolver um modelo mais complexo do que

os estudados e realizados pelos trabalhos estudados no capítulo 2, será benéfico para este estudo a criação de um modelo como os já criados por vários motivos:

- Confirmação de que modelos de Markov têm bons resultados no estudo do fluxo de estudantes como já foi estudado em diversos trabalhos;
- Verificação de que os dados obtidos são também adequados para estudo pretendido. Ao modelar um modelo de Markov mais simples do que pretendido podemos rapidamente, sem grandes custos, averiguar se temos um conjunto de dados úteis para o objetivo pretendido;
- Utilização dos resultados do modelo de probabilidade estática como comparação para os resultados do modelo pretendido posterior. Ao tornar o nosso modelo mais complexo, queremos também que os resultados melhorem de forma significativa.

O primeiro passo será então definir o modelo. Tal como foi mencionado no capítulo 4.1, queremos um modelo com 12 estados para representar os anos curriculares, um estado que represente um estado de conclusão do percurso escolar e um de desistência.

Com os estados do modelo definidos, podemos então ter em mente de como será a nossa matriz de transição MT_k (matriz quadrada de ordem 14) tal como demonstrada na imagem seguinte:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{112} & p_{1C} & p_{1D} \\ p_{21} & p_{22} & \cdots & p_{212} & p_{2C} & p_{2D} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{121} & p_{122} & \cdots & p_{1212} & p_{12C} & p_{12D} \\ p_{C1} & p_{C2} & \cdots & p_{C12} & p_{CC} & p_{CD} \\ p_{D1} & p_{D2} & \cdots & p_{D12} & p_{DC} & p_{DD} \end{bmatrix}$$

Figura 9: Matriz de transição genérica

em que uma entrada p_{ij} irá representar a probabilidade de um aluno ir do estado i para o estado j . Os estados correspondentes aos anos curriculares estão representados pelo número do mesmo, o estado de conclusão do ensino está representado pela letra C e o estado de desistência pela letra D .

Nesta primeira exploração dos modelos de Markov, a matriz de transição irá manter os valores das probabilidades ao longo do tempo, isto é, em cada interação temporal que efetuarmos ao nosso modelo, as probabilidades irão manter-se constantes, tal como foi explorado em diversos estudos (Brezavšek & Baggia, 2015; Alawadhi & Konsowa, 2010; Shah & Burke, 1999), iremos utilizar as tabelas de frequências absolutas que foram já criadas no capítulo passado.

Utilizando as tabelas de frequências, calculamos a proporção de alunos para cada ano curricular para cada ano letivo que transitou para o ano curricular seguinte, que ficou retido e que desistiu do sistema educativo. Estas proporções serão utilizadas como probabilidades e iremos definir nestes primeiros modelos, na matriz de transição a probabilidade de um tipo de transição entre estados como a média das proporções calculadas de todos os anos letivos utilizados para modelar a nossa cadeia de Markov.

Com uma matriz de transição definida, resta só definir como serão os ingressos calculados para podermos testar o modelo. Nesta primeira fase, tal como as probabilidades, os ingressos irão ser calculados utilizando médias dos valores dos dados utilizados para a modelação. Calculamos quantos alunos ingressaram em cada ano letivo e de seguida a média desses valores. Para cada

ano letivo, calculamos qual a proporção de ingressos de que cada ano curricular representava em relação ao número total de ingressos nesse ano letivo. Com as proporções calculadas, calcularemos as médias dessas proporções para cada ano curricular, obtendo assim uma proporção média de quantos alunos, dos que ingressam, estão a ingressar para cada ano curricular. Ao multiplicar estas proporções, pela média do número de ingressos total, obtém-se assim uma estimativa de quantos alunos ingressam em cada ano curricular.

Para confirmação de que os resultados são consistentes, irá criar-se um modelo para a previsão o número de alunos no ano letivo de 2019/20, de seguida um modelo com menos um ano letivo utilizado nos dados para o modelo e utilizado para prever o número de alunos no ano letivo 2018/19 e o ano 2019/20 e finalmente um terceiro modelo com menos um ano letivo nos dados utilizados para o modelo que o anterior e este será para prever os três anos letivos seguintes: 2017/18, 2018/19 e 2019/20.

Para cada modelo, tal como explicado no subcapítulo anterior, multiplicamos o número de alunos do último ano utilizado para a modelação pela matriz de transição e finalmente adicionados os ingressos a esse resultado. Este processo é realizado recursivamente para assim obter várias previsões dos números de alunos nos anos letivos seguintes. Estas previsões estão apresentadas no apêndice F com os resultados numéricos na primeira tabela do apêndice I.

Nesta primeira iteração dos modelos pretendidos obtêm-se resultados curiosos que devemos de imediato ter em nota para o desenvolvimento do modelo. Observa-se uma diferença significativa nos valores previstos e reais para o número de alunos no primeiro ano curricular que se apresenta claramente nas visualizações gráficas tal como por exemplo é mostrado na figura 10. Olhando de volta para o nosso modelo, acredita-se que esta discrepância se deve em grande parte a este valor de previsão não ser calculado pela matriz de transição pois nesta, a probabilidade de um aluno do primeiro ano curricular transitar para o segundo ano é bastante alta. O valor de previsão do primeiro ano curricular é maioritariamente e quase totalmente calculado pelo número de ingressos no sistema educacional. Com esta análise, estabelecemos que se deverá de dar especial atenção aos ingressos quando estivermos a desenvolver o modelo.

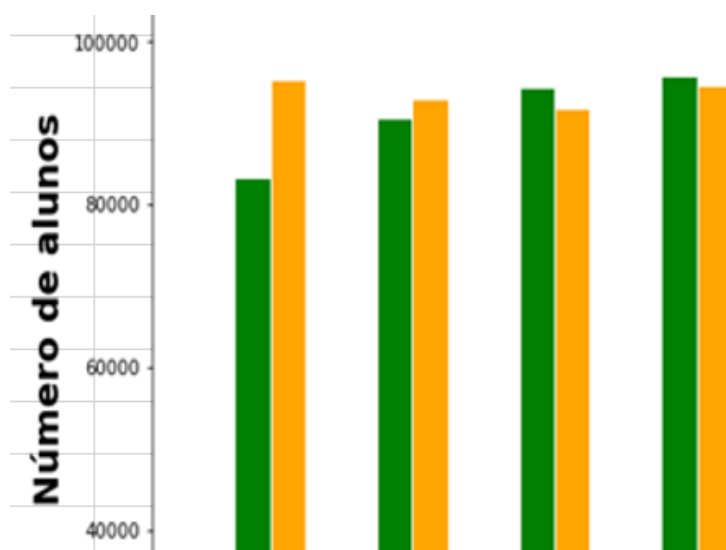


Figura 10: Representação gráfica da diferença dos erros de previsão entre o 1º, 2º e 3º ano curricular (retirado do primeiro gráfico do apêndice F)

Uma outra observação é que apesar dos valores de previsão dos anos correspondentes ao secundário serem ligeiramente melhores em relação aos do primeiro ano curricular, ainda existe uma diferença significativa nos valores do MAPEs em relação aos restantes anos curriculares (segundo ao nono ano curricular). Isto acontece devido ao percurso ser diferente entre alunos e assim gerar heterogeneidade nos percursos.

De seguida, criaremos modelos com os mesmos dados para a modelação e feitas previsões para os mesmos anos utilizando apenas dados que representam os alunos que frequentam os percursos mais comuns. Os resultados destes modelos encontram-se no apêndice G e a tabela dos MAPEs corresponde à segunda tabela do apêndice I.

Comparando os resultados numéricos nas duas tabelas mencionadas do apêndice I, observa-se, em geral, um aumento nos MAPE. Este é um resultado surpreendente, pois estes últimos modelos, contêm percursos de alunos que são mais semelhantes entre si. Prosseguindo a análise dos resultados, observando agora os gráficos concebidos, continua-se a observar uma grande diferença no primeiro ano curricular como seria de esperar pois nenhum dos cursos filtrados correspondem a este ano. Observa-se também agora uma grande diferença entre a previsão e o valor real para as previsões do décimo ano curricular. Com esta diferença é natural questionar os dados utilizados para o modelo e uma hipótese ponderada foi se os cursos que estamos a utilizar nestes últimos modelos são de facto semelhantes entre si.

Uma grande porção dos dados, que pela dimensão poderia ter impacto nos modelos e que poderia ser a razão desta diferença são os cursos profissionais. Para testar, se este grupo era o fator que poderia estar a causar algum desequilíbrio com os restantes dados, foram realizados de novo, modelos dos percursos mais comuns, mas nesta iteração foram também filtrados os cursos profissionais. Para este tipo de modelos foi necessária uma especial atenção aos dados utilizados na modelação devido a casos particulares de percursos de alunos. Um aluno que esteja, por exemplo, atualmente no nono ano curricular pode obter aprovação desse mesmo ano curricular e, no entanto, pretender continuar o seu percurso escolar num curso profissional e neste caso, deverá ser contabilizado como um aluno que “sai” do sistema e, portanto, os alunos que troquem para um curso não correspondente aos mais comuns irão ser contabilizados como se abandonassem o sistema. Estes resultados podem ser consultados no apêndice H, com os resultados numéricos dos MAPEs na terceira tabela do apêndice I. Comparando os resultados, observa-se resultados semelhantes ao primeiro grupo de modelos (apêndice F e primeira tabela do apêndice I) que utilizaram todos os tipos de cursos para a modelação. Observa-se, tal como em todos os outros modelos, o mesmo problema para o primeiro ano curricular em que a discrepância significativa entre o valor real e previsão afeta previsões de anos letivos seguintes.

Analisando as tabelas do apêndice I, obtemos diversas conclusões:

- Os cursos profissionais não têm um fluxo de alunos semelhante aos restantes percursos mais comuns e se quisermos obter um modelo dos cursos mais comuns será conveniente fazer esta separação para obter uma análise mais real do fluxo mais comum de alunos;
- Os ingressos são uma parte fundamental, destes modelos e que devem ser analisados com mais detalhe. Uma grande falha destes primeiros modelos é a incapacidade de obter previsões do primeiro ano curricular em relação aos outros anos curriculares e esta imprecisão afeta os anos curriculares seguintes quando o nosso modelo tenta prever os anos letivos subsequentes;
- Finalmente, existe claramente uma distinção nos resultados dos últimos anos curriculares e dos primeiros (com a exceção do primeiro ano curricular). Este resultado, sugere uma especial atenção aos dados a utilizar pois o percurso inicial dos alunos é praticamente o

mesmo e quando começam a existir opções para os alunos obterem percursos diversos entre si, a precisão do modelo desce ligeiramente.

Com estas observações, decidimos utilizar os dados referentes ao ensino comum para análise e comparação da evolução do modelo. Estes dados não só revelaram os melhores resultados nas previsões, mas também são os que contêm alunos com percursos mais semelhantes entre si, apresentado assim um fluxo mais uniforme. Os resultados destes modelos podem ser consultados na tabela 1. Estes irão funcionar como base de comparação para os próximos modelos apresentados neste trabalho.

Tabela 1: Resultados do modelo teste

Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18	2019/20	13,47	2,67	2,93	1,53	0,98	0,82	1,57	5,97	3,31	3,07	4,06	3,37	3,65	3,41
2016/17	2018/19	11,92	0,95	1,40	0,75	0,45	0,38	1,88	4,57	0,94	4,25	2,44	1,97	2,66	3,22
	2019/20	14,82	13,27	2,20	2,85	0,58	0,41	1,81	8,21	7,52	4,66	7,89	1,92	5,51	4,86
2015/16	2017/18	7,60	0,44	0,07	0,36	0,97	0,82	2,75	3,06	2,25	4,09	1,50	4,47	2,36	2,20
	2018/19	12,88	7,77	0,72	0,56	0,66	1,45	1,74	7,38	3,96	6,90	5,94	1,44	4,28	3,90
	2019/20	15,75	14,84	4,15	1,97	1,21	0,00	1,14	8,57	10,32	7,82	10,26	1,13	6,43	5,59

4.2 Evolução do modelo

Apesar dos resultados relativamente modestos dos modelos iniciais, existem várias oportunidades para desenvolvimento dos mesmos. Existe o caso dos ingressos que poderia ser melhorado e dar alguma especial atenção aos ingressos do primeiro ano curricular porque não só é o ano curricular com mais ingressos, mas também quase todos os alunos deste ano curricular estão a ingressar no mesmo ano letivo no sistema educativo português. Outra especial atenção será dada às probabilidades de transição entre estados. O sistema educativo é um sistema em constante evolução que sofre alterações frequentemente. Portanto é natural que as probabilidades de transição entres estados do nosso modelo que representam a transição, retenção e abandonos também se alterem com o decorrer dos anos.

4.2.1 Probabilidades de transição

O ensino tem diversos aspetos que causam e motivam a evolução do mesmo. Uma observação que já foi referida na análise dos dados é o caso das legislações mencionadas anteriormente (Decreto-Lei no 176/2012; Lei no 85/2009) que invertem as tendências dos valores relativos das transições, retenções e abandonos por volta do ano letivo de 2012/13. Nesta próxima fase, iremos tentar compreender como estes acontecimentos e suas probabilidades evoluem ao longo do tempo, isto é, tentar observar se estas probabilidade têm alguma relação com a unidade de tempo.

Regressões polinomiais são um tipo de modelação comum e eficiente considerada para este problema (Ostertagová, 2012). As flexibilidades deste tipo de modelos permitem-nos uma exploração da utilização dos mesmos para a previsão de como estas probabilidades variam para valores futuros.

Com regressões polinomiais, defronta-se também um problema de *overfitting*. É evidente que quanto maior o grau da nossa regressão polinomial, maior será o nosso coeficiente de determinação, mais conhecido como R^2 , mas é importante reconhecer que queremos que o nosso modelo consiga prever valores corretos, neste caso, para anos letivos futuros. Para nos ajudar a detetar casos de *overfitting* foi também estudado outros tipos de métricas que nos permitem avaliar as nossas regressões como é o caso do *predicted R^2* (Babyak, 2004; Ye & Zhao, 2010). A ideia principal desta métrica consiste em remover uma pequena porção do nosso *dataset*, usar esse restante para a modelação e com o modelo obtido fazer predição da porção retirada. Assim, podemos avaliar se o nosso modelo realmente é capaz de fazer predições de valores não utilizados na modelação ou se o modelo obtido está a fazer um *overfit* aos dados fornecidos.

Iniciamos o estudo deste novo passo ao tentar ajustar um modelo de regressão polinomial aos pontos visualizados no apêndice D. Iremos também analisar as regressões quando se retira anos ao modelo pois se quisermos implementar estas regressões no nosso modelo de Markov o nosso objetivo será obter previsões futuras e é importante confirmar se estes modelos conseguem fazer previsões satisfatórias. Para cada modelo calculamos os R^2 e os *predicted R^2* .

Os valores de R^2 servirão para confirmar se estes modelos são uma boa adaptação ao problema em questão. Com as médias destes coeficientes calculados e com médias superiores a 0,7 para R^2 s para graus polinomiais ≥ 2 , estes modelos serão uma investigação interessante para o projeto. Mas estes valores dos R^2 s não serão suficientes para a decisão de qual o melhor modelo a utilizar e para tal, calcularemos e analisaremos agora os *predicted R^2* s.

Mas antes de analisar os valores dos *predicted R^2* , é importante relembrar de que esta métrica não tem um intervalo de valores possíveis como o R^2 $([0;1])$ e de que a utilização desta métrica é para ser comparada com outros valores para fins comparativos entre modelos de regressão.

Ao analisar as tabelas desta métrica, existe uma correspondência com a análise previamente realizada no capítulo 3.2. Neste capítulo, notou-se de que as tendências para as probabilidades de transição e retenção apresentavam características visuais côncavas. Se analisarmos os resultados dos *predicted R^2* , podemos observar, para ambos estes tipos de acontecimento, que as médias dos *predicted R^2* são favoráveis no caso das regressões polinomiais de grau 2. Para cada acontecimento possível, foram criadas regressões quadráticas para cada ano curricular tal como ilustrado na figura 11.

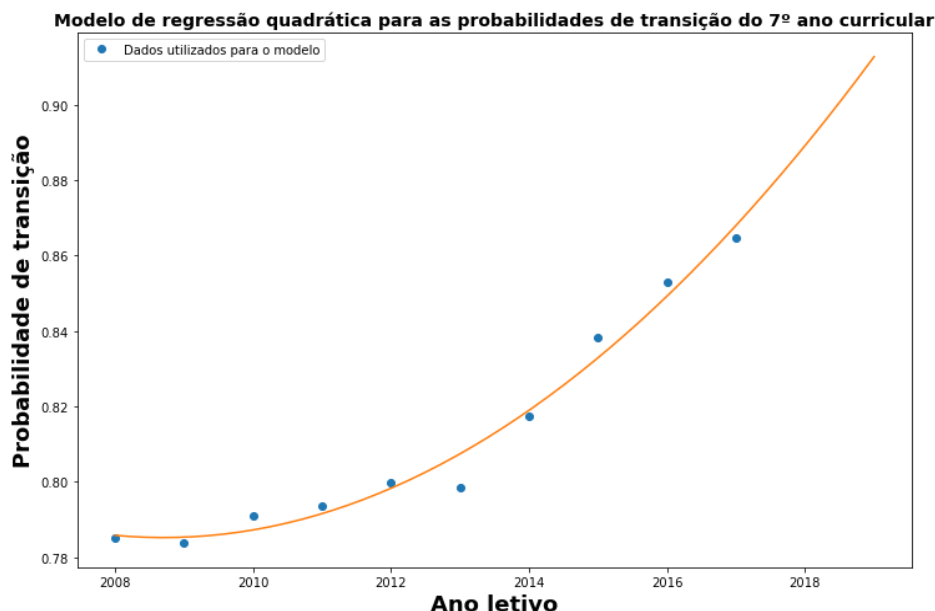


Figura 11: Exemplo de modelação de uma regressão quadrática (dados até ano letivo 2017/2018)

Com esta observação foi decidido a utilização de regressões polinomiais de grau 2 para o cálculo das probabilidades de transição da matriz de transição do nosso modelo de Markov. Para tal são necessárias algumas ponderações:

- Ao utilizar regressões independentes entre si para cada tipo de probabilidade surge a possibilidade de para um determinado ano curricular e para um determinado ano letivo de a soma da probabilidade de transição, de retenção e abandono (os 3 acontecimentos possíveis para cada estado do nosso modelo de Markov) não ser igual a 1. Isto acontece devido à independência das regressões entre si;
- Regressões polinomiais, desde que não sejam constantes, irão tender para valor infinitos. É necessário ter esta tendência em atenção pois não nos faz sentido obter previsões de valores probabilísticos fora do intervalo $[0;1]$.

Com estes dois problemas em reparo foi obtida uma solução para resolver ambos. O modelo irá para cada ano seguinte, calcular novas probabilidades de acontecimento para cada tipo de transição entre estados utilizando as tais regressões polinomiais. No entanto, antes de colocar os valores numa matriz de transição e aplicá-la no algoritmo para obtenção dos números de alunos do ano seguinte, irá ser realizada uma normalização dos valores. Irão ser somados os valores de probabilidade de transição, retenção e abandonos para cada ano curricular e calculados os valores relativos de cada um desses acontecimentos do valor total. Assim garantimos, que a soma dos nossos valores é igual 1 mantendo assim a lógica do fluxo de alunos pelos nossos estados e também que o valor das nossas probabilidades é sempre um valor no intervalo $[0;1]$.

Foram então criados modelos com os mesmos parâmetros dos últimos modelos criados no capítulo anterior (dados dos ensinos comuns). Após a modelação foi aplicado o mesmo algoritmo à distribuição numérica dos alunos que se encontram no último ano letivo utilizado para a modelação, obtendo assim previsões dos alunos dos seguintes anos letivos seguintes. As visualizações gráficas dos resultados de previsão destes modelos estão disponíveis no anexo K e os resultados numéricos encontram-se na tabela 2.

Tabela 2: Resultados do modelo com probabilidades dinâmicas

Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18	2019/20	13,24	1,14	2,26	1,46	4,43	2,88	2,49	2,49	2,56	3,00	4,44	3,10	3,63	3,18
2016/17	2018/19	11,75	2,08	2,38	2,68	6,63	1,81	3,20	1,84	0,87	2,43	3,07	0,84	3,30	3,05
	2019/20	14,64	9,63	5,91	5,58	8,49	9,02	1,81	1,03	3,25	4,19	2,59	0,33	5,54	4,24
2015/16	2017/18	7,55	2,22	3,62	5,74	6,26	2,03	1,26	0,99	0,70	2,22	1,29	5,42	3,28	2,37
	2018/19	12,82	4,63	6,63	9,14	11,34	7,96	0,14	1,67	0,71	1,17	0,05	2,39	4,89	4,56
	2019/20	15,69	11,48	2,31	12,47	13,79	13,72	5,23	4,65	3,76	1,83	2,61	4,86	7,70	5,25

Observando esta tabela e comparando-a com os resultados do tipo de modelo anterior, nota-se numa ligeira descida na performance do modelo. Apesar de se observar aumentos no valor do MAPE, esta subida é no pior dos casos um aumento de 1,27% (6,43% para 7,70% na previsão do ano letivo 2019/20 com o modelo que utiliza dados até o ano letivo 2015/16 para sua modelagem) estes resultados continuam a ser resultados propícios (<10%) (Rahim, Ibrahim, Kasim & Adnan, 2013).

4.2.2 Ingressos

Um outro aspeto de enorme importância que deve ser analisado e tratado em particular são os ingressos. Apesar de estes não terem grande impacto na maioria dos anos curriculares, têm um impacto enorme quando se trata no primeiro ano curricular e como foi observado nos resultados dos modelos anteriores, este é um grande ponto fraco do nosso modelo atual.

Ao contrário das probabilidades de transição entre estados do nosso modelo, os ingressos são uma variável mais previsível. Com o sistema educacional obrigatório em Portugal, as crianças por volta dos 5-6 anos tendem a ingressar para o primeiro ano curricular. Podemos então em vez da utilização de uma variável temporal, utilizar o número de nados-vivos como variável independente para a modelação de ingressos no sistema educacional português.

Foram então realizadas regressões lineares na tentativa de verificação de correlação dos ingressos com os nascimentos. Para os ingressos de cada ano curricular foi modelada uma regressão linear com, como variável independente, o número de nascimentos há x anos, sendo x o número médio das idades para esse ano curricular.

Tabela 3: Média de idades por ano curricular

Ano curricular	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º
Média de idade	6,1	7,18	8,21	9,25	10,38	11,44	12,48	13,44	14,45	15,8	16,79	17,8

Foram então calculadas as médias de idade para cada ano curricular para obter conhecimento do desfaseamento entre os nascimentos e os ingressos nos diversos anos curriculares, obtendo assim

os resultados apresentados na tabela 3. Para simplificação desse valor nos cálculos seguintes, x será sempre igual a 5 + ano curricular.

Para determinar se existe de facto correlação, utilizamos o coeficiente de determinação (R^2) para cada ano curricular. Após cálculo dos mesmos obtemos a tabela 4 para rápida visualização.

Tabela 4: Coeficientes de determinação das regressões lineares

Ano curricular	Coef. de determinação
1º	0,87
2º	0,11
3º	0,19
4º	0,11
5º	0,37
6º	0,03
7º	0,06
8º	0,37
9º	0,43
10º	0,62
11º	0,26
12º	0,09

A conclusão destes resultados que obtemos é de que existe realmente uma correlação entre os nascimentos e os ingressos no primeiro ano curricular com a obtenção do valor 0,87 para o R^2 . O mesmo não se pode afirmar para os restantes anos curriculares em que o melhor resultado foi do décimo ano curricular com o valor 0,62 que apesar de ser um valor razoável, é significativamente menos representativo do que o caso do primeiro ano curricular. Estes resultados são os resultados esperados pois é natural que uma criança que nasça em Portugal se inscreva no primeiro ano curricular e não que salte o(s) primeiro(s) ano(s) curricular(es). No entanto, existem crianças e jovens que se inscrevem pela primeira vez no ensino português em anos curriculares que não são o primeiro. Existem várias possibilidades para este acontecimento, uma das consideradas foi o ingresso de jovens imigrantes. Uma criança/jovem pode obter equivalência de habilitações de ensino obtidas num outro país, conseguindo assim a possibilidade de não ter de repetir anos de escolaridade já realizados noutros países e ingressando no ensino português num ano curricular mais adequado (AMA, 2019).

Continuando esta exploração, foram obtidos dados sobre migrações do Instituto Nacional de Estatística (INE, 2020). O objetivo deste segmento será testar qual será a melhor variável independente para a modelação e previsão dos ingressos face à fraca correlação com os nascimentos (exceto o primeiro ano curricular). Foram então escolhidas como possíveis variáveis independentes: as imigrações e os saldos migratórios (imigrações - emigrações).

Para cada possível variável independente foram utilizados os mesmos intervalos de anos que foram utilizados para a modelação dos modelos de Markov para a obtenção das regressões

polinomiais. Outro detalhe desta modelação foi a utilização do número de imigrantes/saldo migratório relacionado com a idade média de cada ano curricular, isto é, para cada ano curricular foram utilizados dados referentes às crianças/jovens da idade correspondente a esse ano curricular como foi calculado na tabela 3. Forem também criados gráficos para analisar a tendências dos valores de ingressos em relação às possíveis variáveis independentes. No entanto, sem nenhuma observação destacável e diferenças significativas entre anos curriculares, foram realizados modelos polinomiais de 4 graus diferentes (primeiro a quarto grau) e calculados os erros de previsão de cada um. Com estes erros foram calculados os MAPEs e respetivos desvio-padrões e colocados na tabela 5 para comparações. Estes valores irão nos permitir compreender quais são os modelos que estão a fazer um overfit aos nossos dados de treino pois esses irão obter grandes diferenças entre os seus valores de previsão e os valores reais para os anos de teste.

Tabela 5: MAPEs das diferentes regressões polinomiais para a previsão dos ingressos

				MAPEs (Mean Absolute Percentage Error)			Desvio padrão		
Desfasamento	Variável independente	Dados até ano	Grau polinomial	2017	2018	2019	2017	2018	2019
Sem desfasamento	Imigrações	2016	1	34	30	60	32	22	57
			2	28	38	209	27	36	179
			3	42	67	1095	40	105	1396
			4	62	174	10721	68	261	8857
		2017	1		25	50		16	47
			2		33	208		29	163
			3		61	980		107	1565
			4		174	10239		313	10836
		2018	1			40			38
			2			153			107
			3			508			329
			4			4360			2831
	Saldo Migratório	2016	1	26	22	30	28	12	21
			2	31	29	91	31	21	66
			3	30	34	116	37	23	113
			4	57	52	637	76	58	797
		2017	1		19	26		10	17
			2		23	79		17	60
			3		26	116		14	96
			4		30	371		33	433
		2018	1			22			14
			2			75			51
			3			105			69
			4			324			424

Ao observar e analisar a tabela anterior podemos tirar várias conclusões. Ao colocar o grau da regressão polinomial a 4, ocorre um overfit extremo o que causa o modelo a obter previsões dispares dos valores reais como pode ser observado na figura 12.

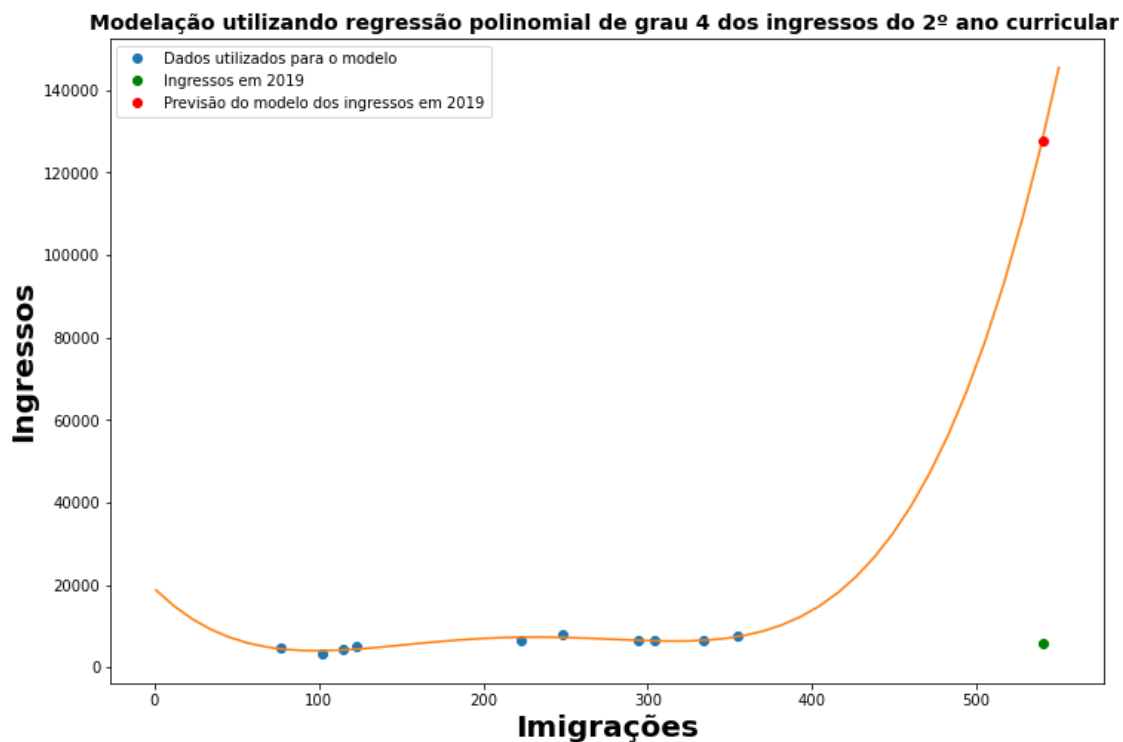


Figura 12: Exemplo de overfitting

O modelo que mostrou melhores resultados nas previsões foi o que utilizou valores de saldo migratório como variável independente e com grau polinomial igual a 1 (regressão linear). No entanto, mesmo este melhor modelo, obteve resultados acima dos esperados (19% vs 2-8% nos MAPEs das cadeias de Markov). Com a ambição de obter melhores resultados e de valores mais próximos dos obtidos pelos modelos de Markov, ponderou-se também se fazia sentido utilizar um desfasamento das variáveis independentes para estes, caso fossem utilizados para prever os ingressos do ano letivo seguintes e não os do ano letivo corrente. Então, para cada caso calculado na tabela anterior, foram calculados os MAPEs e desvio-padrões para modelos de regressão com desfasamento de 1 ano letivo entre a variável independente e a dependente. Estes resultados estão apresentados na tabela 6.

Tabela 6: MAPEs das diferentes regressões polinomiais para a previsão dos ingressos com um desfasamento de 1 ano

				MAPEs (Mean Absolute Percentage Error)			Desvio padrão		
Desfasamento	Variável independente	Dados até ano	Grau polinomial	2017	2018	2019	2017	2018	2019
Desfasamento de 1 ano (Variável independente do ano anterior)	Imigrações	2016	1	24	21	27	23	14	20
			2	19	15	26	17	7	22
			3	28	20	53	23	16	75
			4	20	20	99	14	16	264
		2017	1		20	25		11	17
			2		13	25		7	24
			3		16	46		10	52
			4		18	89		14	225
		2018	1			22			16
			2			26			24
			3			47			52
			4			90			222
	Saldo Migratório	2016	1	20	21	23	19	10	8
			2	15	31	38	14	28	37
			3	26	45	58	22	61	102
			4	22	46	68	15	75	131
		2017	1		19	21		8	7
			2		29	36		23	32
			3		41	56		60	102
			4		46	71		82	150
		2018	1			18			6
			2			27			19
			3			26			30
			4			24			35

Numa primeira análise, observa-se que estes últimos modelos obtiveram resultados melhores do que os modelos sem desfasamento com o melhor valor dos MAPEs a 12,68%.

Na eventual escolha de qual o tipo de modelo regressivo para ser integrado no nosso modelo de Markov, foi escolhido a utilização do desfasamento de 1 ano. Dentro desta escolha ainda existem 8 combinações de parâmetros (2 variáveis independentes vezes 4 graus polinomiais). No entanto, destacam-se, pelos seus resultados, dois modelos: Regressão quadrática (grau 2) com imigrações como variável independente por obter o melhor resultado (13%), mas também a regressão linear com saldo migratório como variável independente por obter os resultados mais consistentes em todas as combinações para este tipo de modelo (18% - 23%).

Foram então realizados modelos que integram estas regressão polinomiais para a obtenção de uma previsão dos valores dos ingressos para cada ano curricular. Foram realizados dois tipos de modelos com estas novas adaptações aos ingressos. Ambos contêm uma regressão linear que

obtem uma previsão dos ingressos do primeiro ano curricular utilizando os nascimentos em Portugal, mas para os restantes anos curriculares, para a obtenção da previsão de ingressos, o primeiro dos modelos utiliza regressões lineares com o saldo migratório como variável independente enquanto o segundo utiliza regressões quadráticas com imigrações como variável independente. Os resultados destes modelos estão representados nas tabelas 7 e 8.

Tabela 7: Resultados do modelo com probabilidades e ingressos dinâmicos (saldo migratório)

Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18	2019/20	0,98	0,02	1,49	0,83	3,70	2,27	3,05	2,14	2,31	3,40	4,28	3,37	2,32	1,30
2016/17	2018/19	1,60	1,06	1,67	2,10	6,08	1,41	3,62	1,62	1,11	2,73	2,95	0,98	2,25	1,46
	2019/20	4,16	2,15	4,09	4,22	7,21	7,86	2,80	0,28	2,77	4,84	2,15	0,08	3,55	2,38
2015/16	2017/18	1,24	1,59	3,27	5,54	6,11	1,90	1,42	0,89	0,59	2,43	1,22	5,42	2,64	1,98
	2018/19	6,16	0,43	5,30	8,23	10,59	7,41	0,70	1,30	0,36	1,60	0,28	2,59	3,75	3,62
	2019/20	8,86	6,77	4,95	10,58	12,20	12,37	4,07	3,76	3,12	2,62	2,02	5,46	6,40	3,72

Tabela 8: Resultados do modelo com probabilidades e ingressos dinâmicos (imigrações)

Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18	2019/20	0,98	7,55	7,48	7,24	11,08	7,03	1,83	6,22	6,21	2,05	7,94	0,13	5,48	3,39
2016/17	2018/19	1,60	8,04	7,61	8,53	13,02	5,97	1,03	5,49	2,88	2,39	6,67	2,43	5,47	3,56
	2019/20	4,16	5,89	16,53	16,09	20,17	19,03	6,59	8,60	10,33	4,55	10,47	7,17	10,80	5,72
2015/16	2017/18	1,24	8,00	8,92	11,43	12,37	6,15	2,87	4,70	4,39	2,69	4,97	2,03	5,81	3,65
	2018/19	6,16	7,03	17,08	19,62	22,42	17,80	8,15	9,18	8,07	7,33	8,28	4,60	11,31	6,09
	2019/20	8,86	1,42	17,85	27,34	29,29	28,05	19,02	15,91	14,48	10,41	14,01	6,74	16,12	8,79

Nestes últimos modelos observamos até agora o nosso melhor resultado. Apesar do modelo que utilizou valores de imigrações para um modelo de regressão quadrática não obter os melhores resultados, o modelo que utilizou o saldo migratório como variável independente numa regressão linear obteve resultados que rivalizam os do modelo base (Em 6 modelos, obteve um menor MAPE em 4 dos casos, 1 aproximadamente igual e 1 pior). No entanto, em todos os casos, este novo modelo obteve um melhor resultado em todos os casos para o desvio-padrão.

Capítulo 5

Conclusões

Ao aplicar modelos de Markov ao fluxo de estudantes do ensino português, neste trabalho foi possível obter diversos resultados promissores na área de modelação de fluxos.

Numa primeira fase, foi necessária uma forte compreensão não só de como os modelos de Markov atuam, mas também dos dados fornecidos e do problema em questão. Perante ao problema de o ensino português não ser um sistema fechado, isto é, existe a entrada de variáveis (neste caso alunos) de fora para dentro do sistema foi necessário realizar uma adaptação do método que se pretendia utilizar ao problema. A típica representação do sistema numa distribuição do total pelos valores possíveis de cada variável não se enquadrava ao nosso problema devido ao ingresso de novos alunos no sistema. Um dos objetivos criados durante a execução deste trabalho foi a avaliação se este tipo de modelo com valores absolutos em vez de relativos e com um elemento aditivo iria obter resultados significativos. Após a criação dos primeiros modelos de controlo, não só se confirmou que um modelo baseado na propriedade de Markov é adequado ao problema em questão ao obter MAPEs entre os 2 e 7%, mas também de que a adaptação realizada para estimar os ingressos de alunos é um caminho a ser mais explorado.

A partir dos modelos de controlo, observou-se de imediato que a fraqueza do modelo inicial era de facto a má previsão dos valores do primeiro ano curricular pois estes estariam a ser calculados apenas através de médias de valores anteriores ao ano em questão e o erro de previsão de qualquer modelo inicial para o primeiro ano curricular era sempre o maior de todos anos curriculares. Após uma análise de várias possibilidades para a previsão dos ingressos, concluímos que podemos obter um bom progresso na adaptação do modelo ao problema. Ao obter uma correlação forte dos nascimentos em Portugal com os ingressos do primeiro ano curricular, melhoram-se significativamente as previsões do primeiro ano curricular que refletiu-se numa melhoria nos MAPEs. Quanto aos restantes anos curriculares, conclui-se que a utilização das imigrações como variável independente não trouxe benefícios ao desenvolvimento do modelo. No entanto, com a utilização do saldo migratório, observaram-se resultados mistos. Uma possível continuação deste segmento seria uma análise separada de cada ano curricular e para cada um obter qual o melhor tipo de regressão polinomial/modelo.

Trabalho futuro

Com resultados de controlo satisfatórios e ambição de aperfeiçoar os modelos, foram então identificados dois possíveis aspetos a melhorar: as probabilidades da matriz de transição e forma de como os ingressos são calculados. Nesta fase de aperfeiçoamento tentou-se aperfeiçoar a maneira de como as probabilidades seriam calculadas através de uma matriz dinâmica que seria

dependente do ano letivo. Porém, os resultados após esta modificação não melhoraram, visto que o intervalo dos MAPEs aumentou, embora apenas 1% (de 2-7% para 3-8%). Este resultado e trabalho não devem ser descartados apesar de resultados negativos, mas sim motivar em entender, obter e otimizar modelos para predição das probabilidades de transição entre estados e incorporá-los no modelo de Markov.

O algoritmo de cálculo de offset desenvolvido no capítulo 3.3 seria também um ponto de continuação. Apesar de o algoritmo obter o offset que dá a mínima soma de erros, este não irá obter um valor ótimo na métrica de distância de Fréchet. Mas com alterações para ter em conta o quadrado de cada erro poderia tornar o algoritmo ótimo em obtenção do melhor offset para métricas como a RMSD (root-mean-squared deviation) que também é utilizado no estudo de sobreposição de curvas (Wylie, 2013).

Referências

- AMA. (2019). Pedir equivalência de habilitações estrangeiras do ensino básico e secundário. Acedido a 9 agosto 2022. Disponível em eportugal.gov.pt.
- Alawadhi, S., & Konsowa, M. (2010). Markov chain analysis and student academic progress: An empirical comparative study. *Journal of Modern Applied Statistical Methods*, 9(2), 26.
- Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3), 411-421.
- Bessent, E. W., & Bessent, A. M. (1980). Student flow in a university department: Results of a Markov analysis. *Interfaces*, 10(2), 52-59.
- Brezavšček, A., & Baggia, A. (2015). Analysis of student's flow in higher education study programmes using discrete homogeneous Markov chains. In *SOR'15 proceedings, 13th International Symposium on Operational Research in Slovenia* (pp. 473-478).
- Brezavšček, A., Pejić Bach, M., & Baggia, A. (2017). Markov analysis of students' performance and academic progress in higher education. *Organizacija*, 50(2).
- Bryc, W. (1996). *Applied Probability and Stochastic Processes*. University of Cincinnati, Cincinnati, OH, USA.
- Cavique, L., Pombinho, P., Tallón-Ballesteros, A. J., & Correia, L. (2020, November). Data Pre-processing and Data Generation in the Student Flow Case Study. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 35-43). Springer, Cham.
- Ching, W. K., & Ng, M. K. (2006). *Markov chains. Models, algorithms and applications*.
- Correa, S. M. B. B. (2003). *Probabilidade e estatística*.
- Decreto-Lei no 176/2012 de 2 de agosto do Ministério da Educação e Ciência. Diário da República: I série, No 149 (2012). Acedido a 5 abril 2022. Disponível em www.dre.pt.
- Decreto-Lei no 55/2018 de 6 de julho da Presidência do Conselho de Ministros. Diário da República: I série, No 129 (2018). Acedido a 22 abril 2023. Disponível em www.dre.pt.
- Dehling, H. G., Gottschalk, T., & Hoffmann, A. C. (2007). *Stochastic modelling in process technology*. Elsevier.
- DGEEC. (2021), Ministério da Educação & Ministério da Ciência, Tecnologia e Ensino Superior. *Educação e Formação em Portugal*.
- Duarte, A. (2012). A introdução dos cursos profissionais nas escolas públicas: perfil socioeducativo dos alunos. CIES e-Working Paper.
- Eiter, T., & Mannila, H. (1994). Computing discrete Fréchet distance.

- FFMS. (2022). Nados-vivos de mães residentes em Portugal: total e fora do casamento. PORDATA. Acedido a de 8 agosto 2022. Disponível em www.pordata.pt.
- Fiallos, A., & Ochoa, X. (2017, October). Discrete event simulation for student flow in academic study periods. In 2017 Twelfth Latin American Conference on Learning Technologies (LACLO) (pp. 1-7). IEEE.
- Häggström, O. (2002). Finite Markov chains and algorithmic applications (Vol. 52). Cambridge University Press.
- Heileman, G. L., Babbitt, T. H., & Abdallah, C. T. (2015). Visualizing student flows: Busting myths about student movement and success. *Change: The Magazine of Higher Learning*, 47(3), 30-39.
- INE. (2020). Emigrantes permanentes (N.º) por Idade; Anual. Dados extraídos em 21 de maio de 2021. Disponível em www.ine.pt.
- Konstantopoulos, T. (2009). Markov chains and random walks. Lecture notes.
- Lei no 85/2009 de 27 de agosto da Assembleia da República. Diário da República: I série, No 166 (2009). Acedido a 5 abril 2022. Disponível em www.dre.pt.
- Lima Junior, P., Silveira, F. L. D., & Ostermann, F. (2012). Survival analysis applied to student flow in undergraduate Physics courses: an example from a Brazilian university. *Revista Brasileira de Ensino de Física*, 34.
- Lovell, C. C. (1971). Student Flow Models: A Review and Conceptualization.
- Magalhães, F. M., & Pereira, M. L. (2018). Regulamento Geral de Proteção de Dados. Manual Prático, 2.
- Meyn, S. P., & Tweedie, R. L. (2012). Markov chains and stochastic stability. Springer Science & Business Media.
- Oreopoulos, P. (2005). Canadian compulsory school laws and their impact on educational. Analytical Studies Branch Research Paper Series, Statistics Canada, catalogue, (11F0019MIE).
- Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, 48, 500-506.
- Pombinho, P., Cavique, L., & Correia, L. (2022, October). Errors of Identifiers in Anonymous Databases: Impact on Data Quality. In 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022) Salamanca, Spain, September 5–7, 2022, Proceedings (pp. 547-556). Cham: Springer Nature Switzerland.
- Rahim, R., Ibrahim, H., Kasim, M. M., & Adnan, F. A. (2013). Projection model of postgraduate student flow. *Applied Mathematics and Information Sciences*, 7(2), 383-387.
- Saltzman, R. M., & Roeder, T. M. (2012). Simulating student flow through a college of business for policy and structural change analysis. *Journal of the Operational Research Society*, 63(4), 511-523.
- Shah, C., & Burke, G. (1999). An undergraduate student flow model: Australian higher education. *Higher Education*, 37(4), 359-375.
- Tavares, L.V. (1995), On the development of educational policies, *European Journal of Operational Research*, vol. 82(3), pp. 409–421.

UNESCO Institute for Statistics. (2012). International standard classification of education: ISCED 2011. Comparative Social Research, 30.

Wylie, T. R. (2013). The discrete Fréchet distance with applications (Doctoral dissertation, Montana State University-Bozeman, College of Engineering).

Ye, F., & Zhao, A. (2010, April). What You See May Not be What You Get—A Brief Introduction to Overfitting. In Cancer Biostatistic Workshop, Vanderbilt-Ingram Cancer Center.

Apêndices

Apêndice A – Dados de todo ensino

Apêndice B – Dados de ensino não comum

Apêndice C – Dados do ensino comum

Apêndice D – Visualização dos valores relativos de cada tipo de transição (para cada ano curricular)

Apêndice E – Resultados das distâncias de Fréchet às curvas médias de cada ciclo

Apêndice F – Resultados dos modelos de probabilidades estáticas (todo ensino português)

Apêndice G – Resultados dos modelos de probabilidades estáticas (ensino comum com cursos profissionais)

Apêndice H – Resultados dos modelos de probabilidades estáticas (ensino comum sem cursos profissionais)

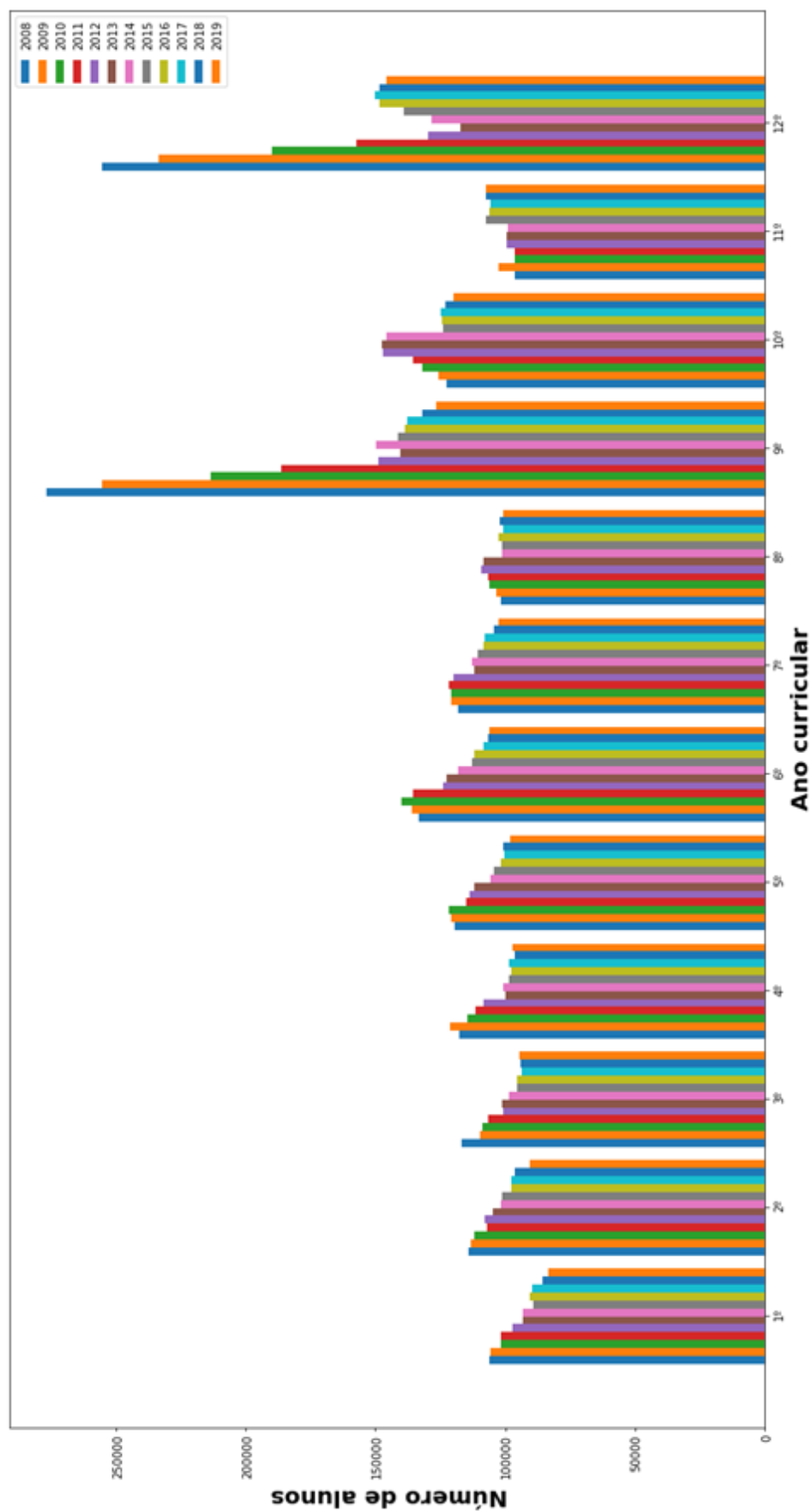
Apêndice I – MAPEs dos modelos com probabilidades estáticas

Apêndice J – Predicted R^2 para os acontecimentos possíveis (Ensino comum)

Apêndice K – Resultados dos modelos de probabilidades dinâmicas utilizando regressões polinomiais (ensino comum)

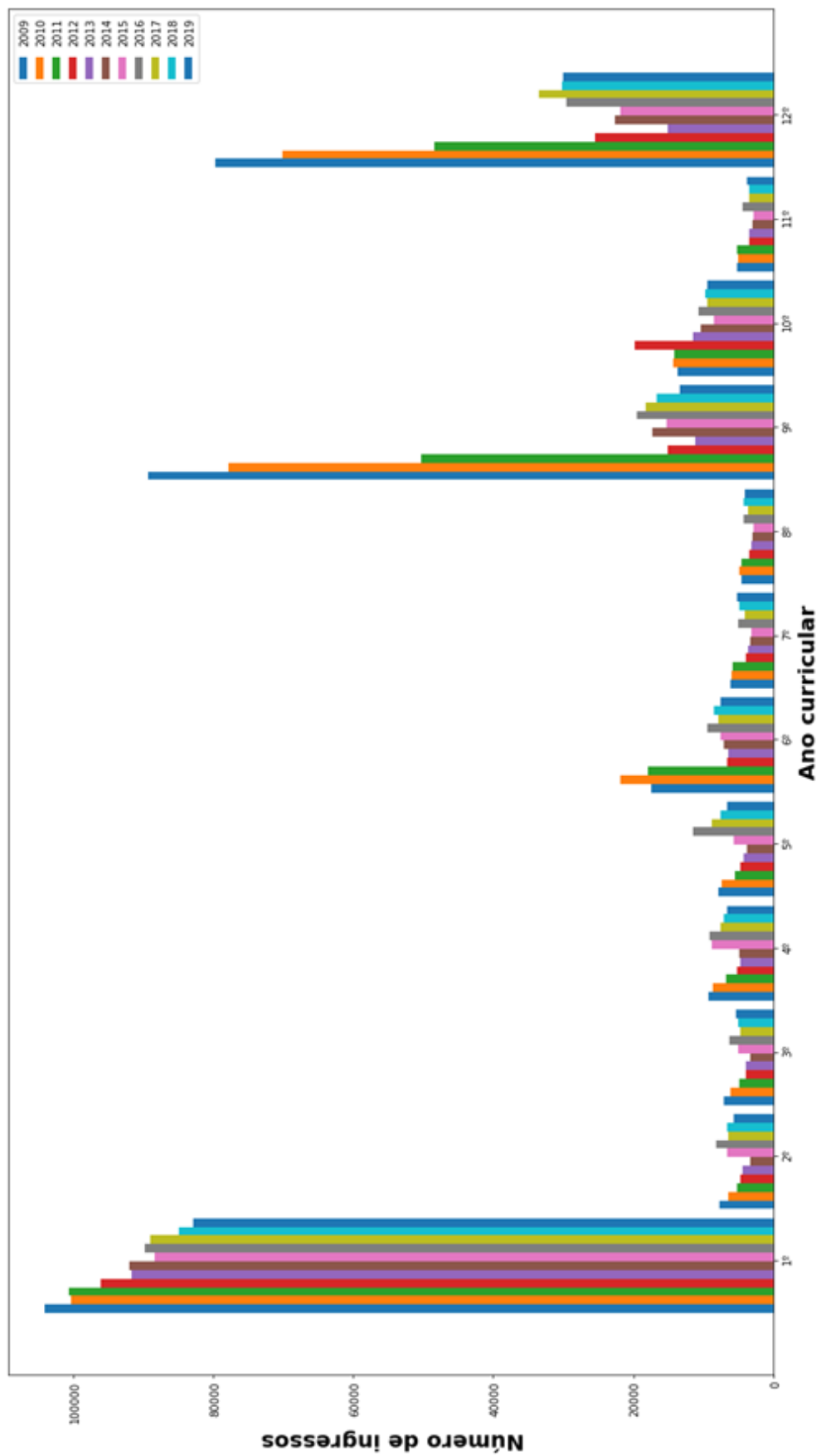
Apêndice A. Dados de todo ensino

Inscritos



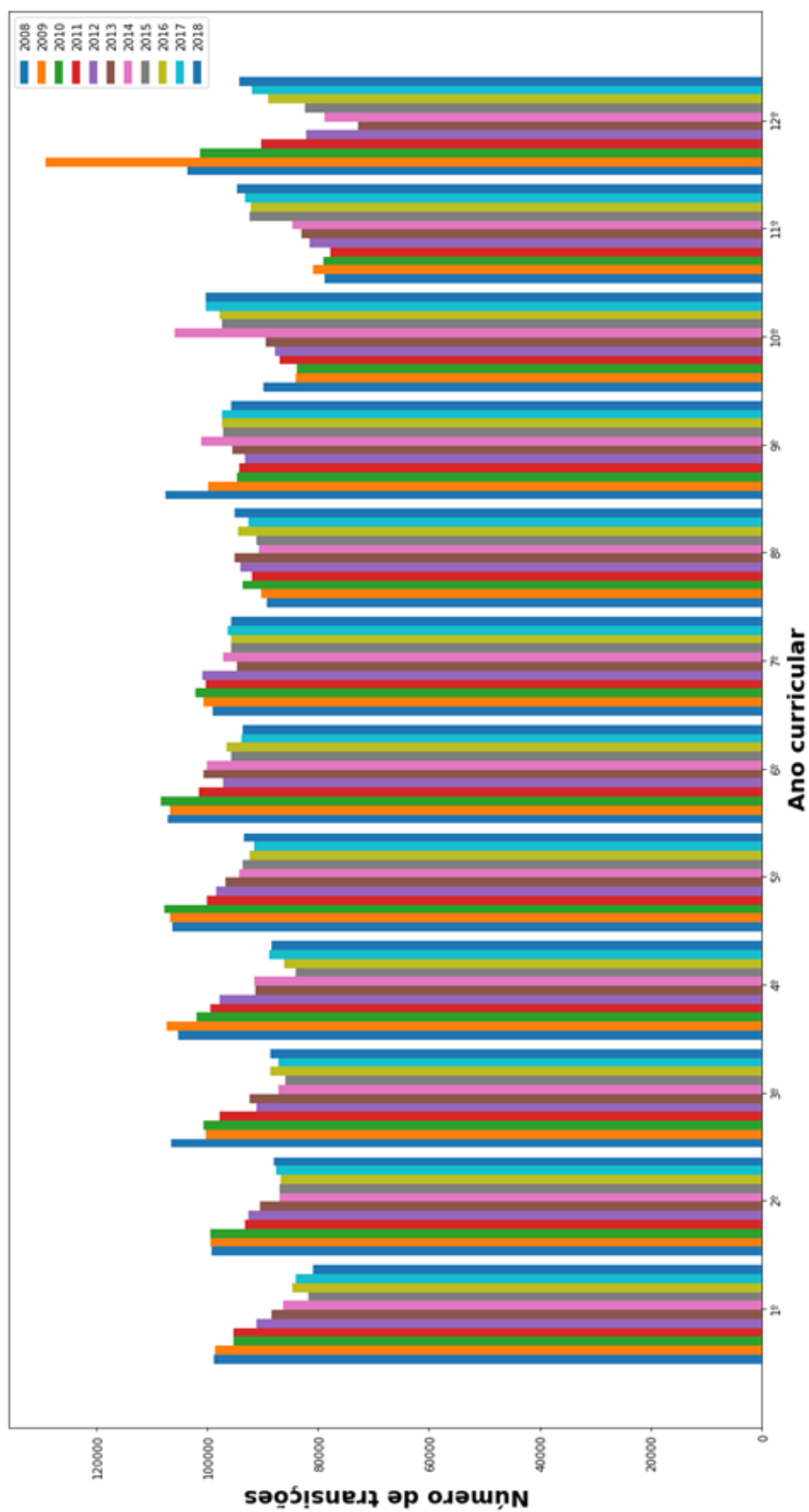
	Ano letivo (X/X+1)												
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
1º	106204	105418	101584	101720	97177	92909	92930	89194	90507	89640	85507	83436	
2º	113900	113234	111987	106844	107855	104837	101476	101264	97676	97587	96173	90666	
3º	116772	109721	108785	106715	100751	101174	98312	95470	95553	93356	94211	94471	
4º	117671	121067	114328	111549	108158	99910	100828	98592	97686	98596	96465	97247	
5º	119585	120624	121662	114875	113553	111826	105769	104489	101452	100096	100681	98244	
6º	133200	135904	140034	135300	123675	122517	118157	112909	112038	108211	106554	105999	
7º	118296	120802	120851	121758	119678	111929	112734	110720	108286	107674	104158	102574	
8º	101641	103468	106151	106279	109188	108461	101238	100965	102428	100596	101856	100911	
9º	277072	255390	213535	186443	148845	140475	149609	141480	138710	137660	132038	126563	
10º	122720	125630	131907	135268	147143	147573	145767	123998	124265	124825	123224	119806	
11º	96345	102360	96450	96284	99401	99210	98802	107467	106224	105577	107420	107177	
12º	255472	233406	190060	157277	129599	117346	128249	138894	148509	150405	148391	145920	

Ingressos



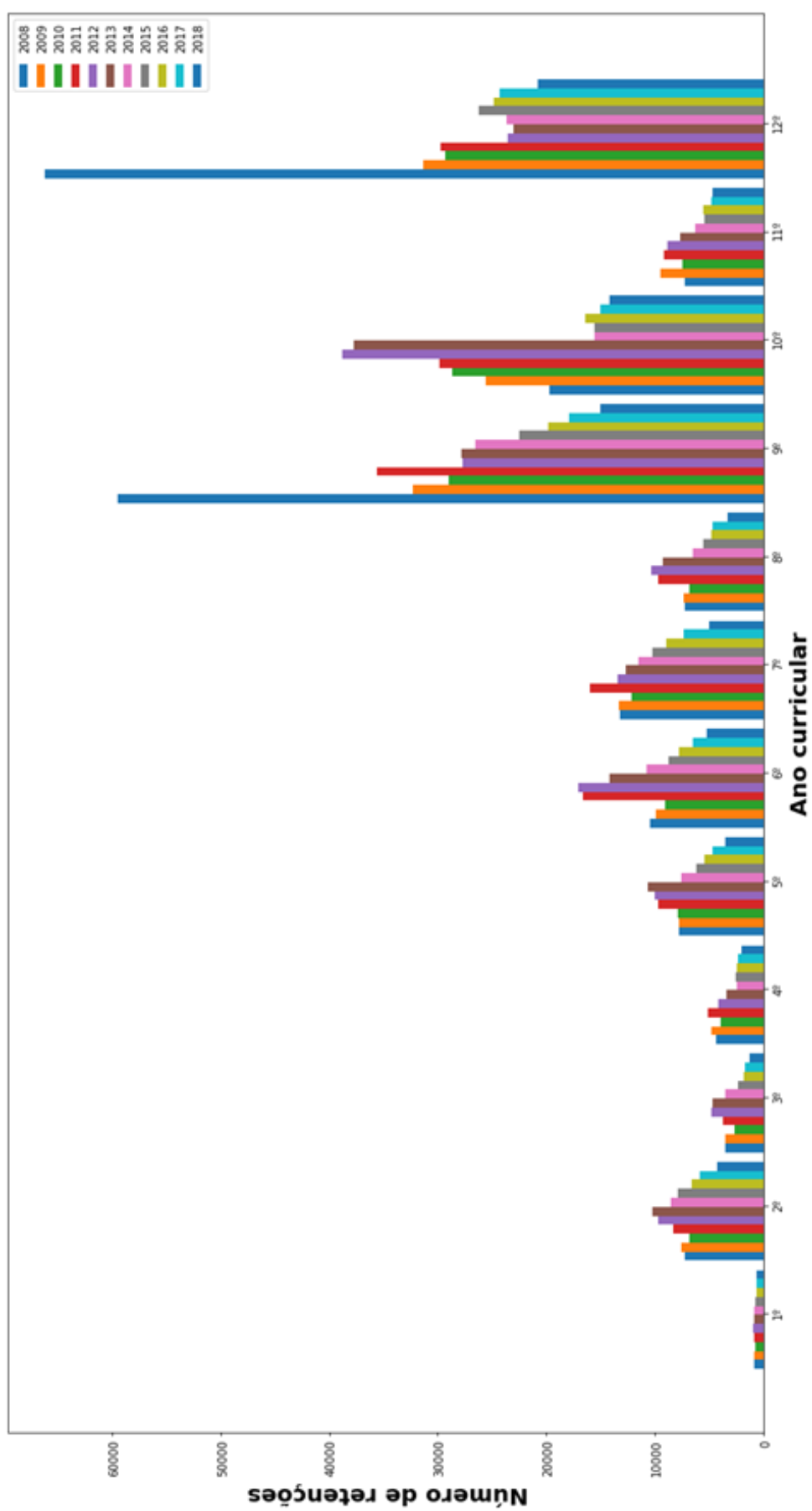
		Ano letivo (X/X+1)											
		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Ano curricular	1º	104013	100269	100565	96119	91643	92029	88374	89784	88974	84913	82805	
	2º	7736	6411	5131	4701	4318	3302	6498	8090	6388	6499	5709	
	3º	7067	6176	4885	3975	3969	3197	5070	6187	4746	4994	5315	
	4º	9203	8658	6776	5225	4651	4837	8843	9162	7457	7051	6601	
	5º	7819	7400	5425	4644	4232	3787	5606	11508	8842	7451	6499	
	6º	17380	21870	17949	6568	6486	7063	7583	9389	7896	8428	7471	
	7º	6102	5957	5747	3979	3539	3349	3123	4949	4136	4821	5233	
	8º	4557	4842	4575	3389	3137	2963	2729	4155	3602	4227	4091	
	9º	89367	77891	50344	15123	11101	17287	15219	19512	18172	16664	13323	
	10º	13660	14295	14078	19708	11384	10312	8403	10706	9366	9773	9399	
	11º	5234	4990	5126	3446	3419	2998	2772	4334	3365	3458	3722	
	12º	79739	70098	48381	25440	15051	22677	21852	29534	33448	30196	29938	

Transições



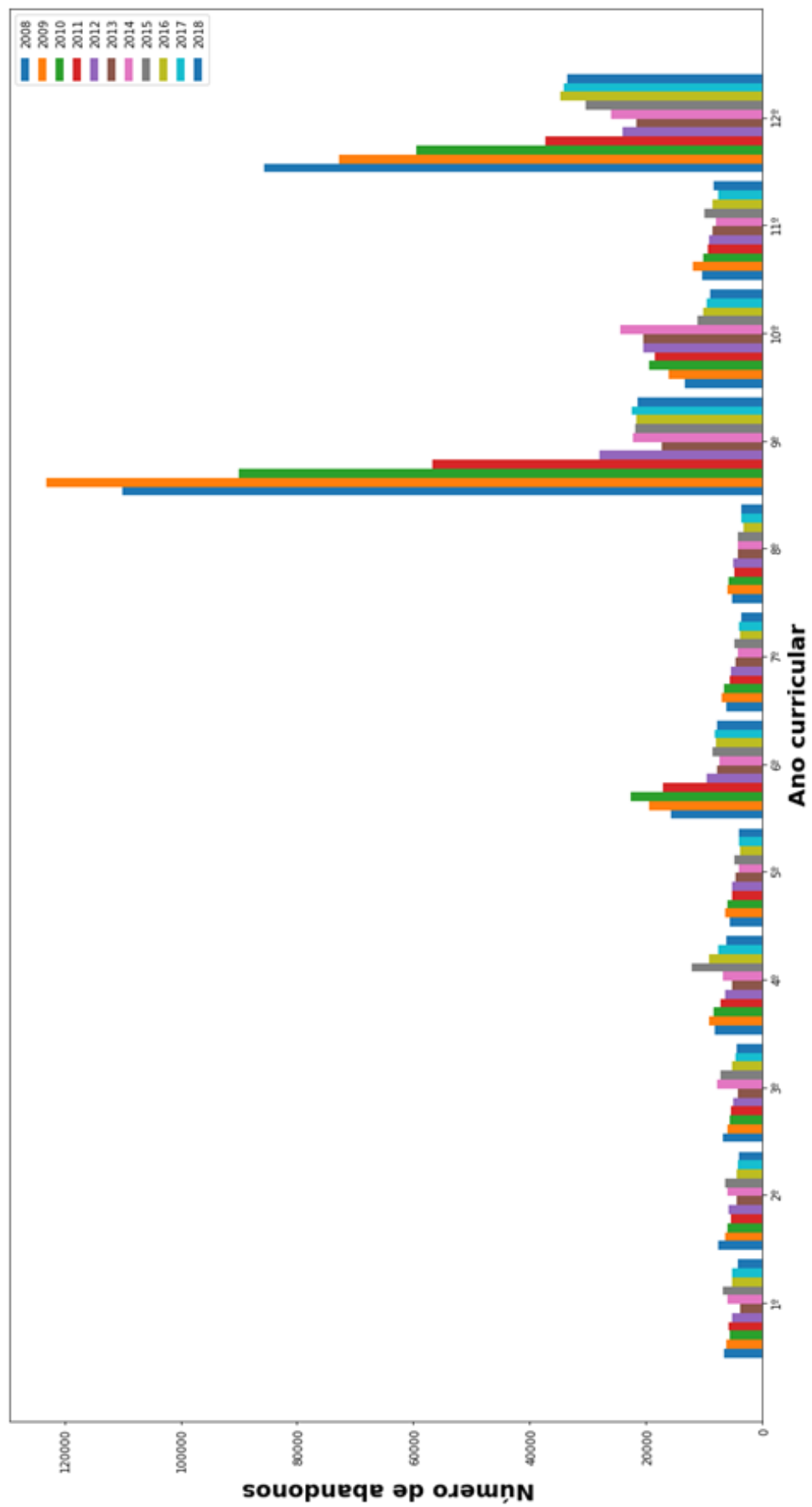
	Ano letivo (X/X+1)											
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
1º	98798	98548	95265	95323	91156	88345	86292	81773	84760	83938	80816	
2º	99213	99337	99360	93200	92504	90385	87002	87002	86841	87505	87961	
3º	106578	100342	100701	97701	91081	92449	87225	86001	88681	87061	88687	
4º	105229	107249	102032	99353	97704	91412	91565	84007	86104	88818	88420	
5º	106203	106625	107825	100162	98446	96733	94303	93678	92378	91459	93344	
6º	107216	106690	108461	101581	97116	100731	100023	95696	96461	93748	93651	
7º	99112	100610	102176	100319	100987	94620	97042	95759	95722	96399	95664	
8º	89291	90273	93652	91876	94013	95004	90745	91214	94433	92481	95143	
9º	107535	99856	94633	94147	93278	95446	101057	97125	97453	97396	95619	
10º	89812	84099	83739	87047	87862	89540	105869	97325	97774	100263	100282	
11º	78756	81011	78943	77761	81573	83044	84575	92309	92251	93200	94595	
12º	103564	129289	101336	90318	82134	72851	78768	82279	89002	91962	94136	

Retenções



	Ano letivo (X/X+1)											
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Ano curricular	1º	835	826	776	784	915	798	792	685	655	602	648
	2º	7228	7508	6793	8249	9644	10216	8540	7863	6547	5859	4219
	3º	3503	3460	2595	3717	4796	4681	3465	2373	1810	1727	1252
	4º	4307	4752	3910	5079	4126	3440	2463	2577	2431	2313	2025
	5º	7787	7742	7843	9638	10032	10650	7496	6116	5390	4711	3487
	6º	10398	9899	9028	16619	17059	14141	10778	8761	7755	6434	5160
	7º	13162	13344	12182	16012	13456	12710	11469	10176	8895	7383	5004
	8º	7184	7371	6813	9712	10297	9277	6458	5548	4781	4683	3238
	9º	59501	32325	28965	35554	27666	27850	26474	22503	19760	17936	14961
	10º	19674	25597	28692	29785	38819	37681	15503	15509	16361	15047	14133
	11º	7278	9431	7412	9146	8832	7636	6304	5378	5492	4805	4622
	12º	66232	31339	29319	29756	23510	22995	23618	26215	24859	24284	20717

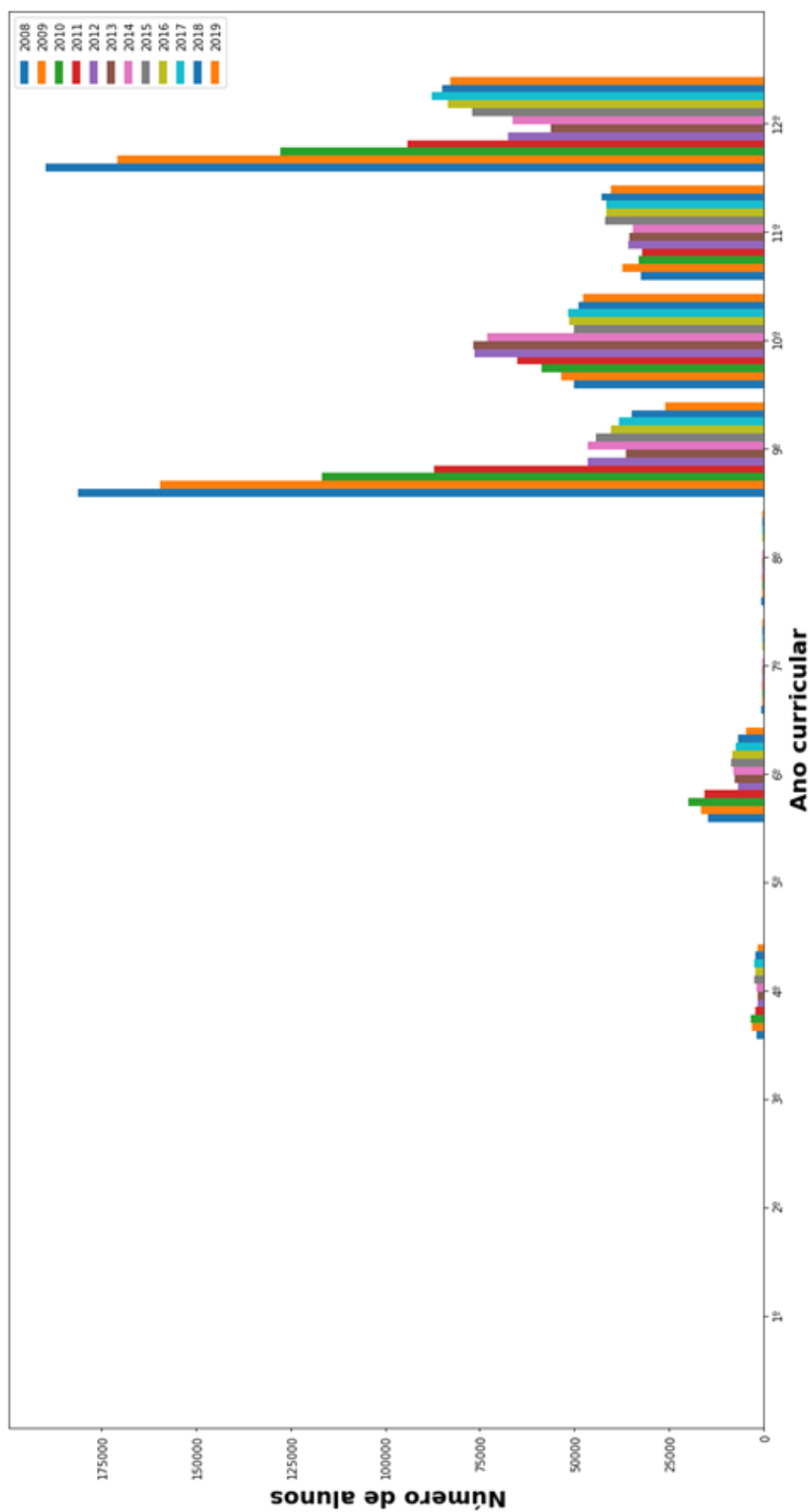
Abandonos



		Ano letivo (X/X+1)											
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Ano curricular	1º	6571	6044	5543	5613	5106	3766	5846	6736	5092	5100	4043	
	2º	7459	6389	5834	5395	5707	4236	5934	6399	4288	4223	3993	
	3º	6691	5919	5489	5297	4874	4044	7622	7096	5062	4568	4272	
	4º	8135	9066	8386	7117	6328	5058	6800	12008	9151	7465	6020	
	5º	5595	6257	5994	5075	5075	4443	3970	4695	3684	3926	3850	
	6º	15586	19315	22545	17100	9500	7645	7356	8452	7822	8029	7743	
	7º	6022	6848	6493	5427	5235	4599	4223	4785	3669	3892	3490	
	8º	5166	5824	5686	4691	4878	4180	4035	4203	3214	3432	3475	
	9º	110036	123209	89937	56742	27901	17179	22078	21852	21497	22328	21458	
	10º	13234	15934	19476	18436	20462	20352	24395	11164	10130	9515	8809	
	11º	10311	11918	10095	9377	8996	8530	7923	9780	8481	7572	8203	
	12º	85676	72778	59405	37203	23955	21500	25863	30400	34648	34159	33538	

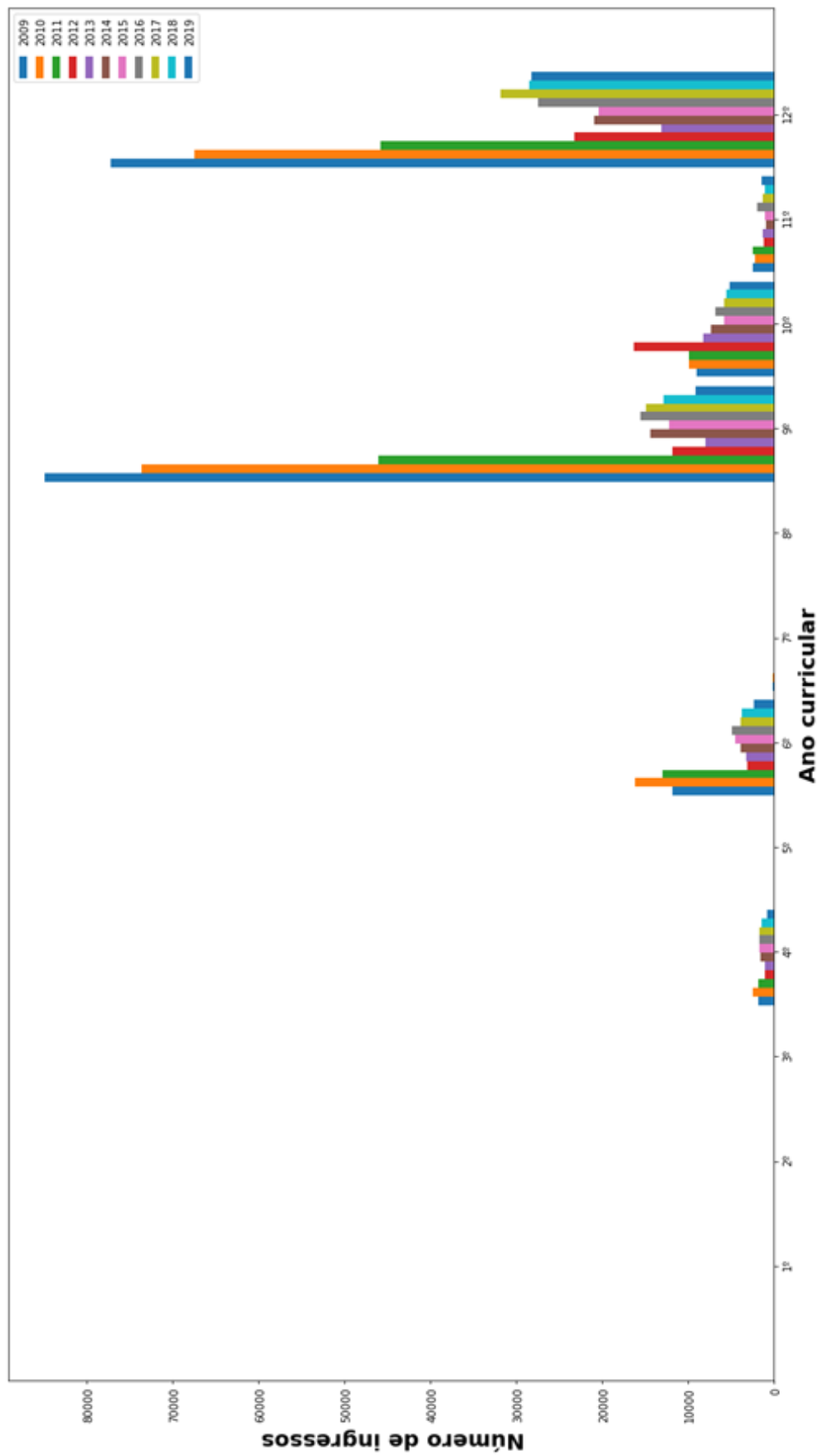
Apêndice B. Dados de ensino não comum

Inscritos



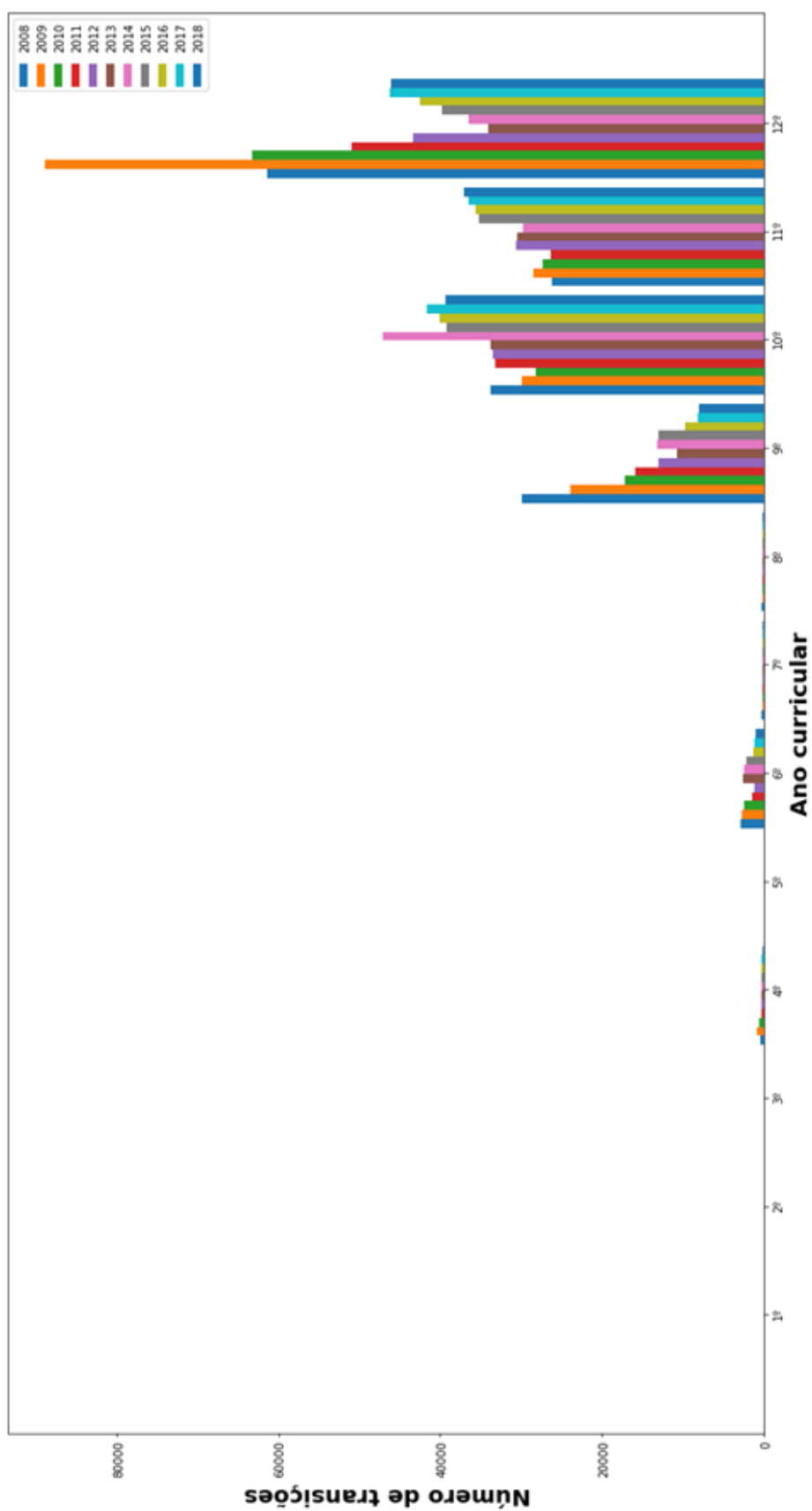
	Ano letivo (X/X+1)											
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1º	0	0	0	0	0	0	0	0	0	0	0	0
2º	0	0	0	0	0	0	0	0	0	0	0	0
3º	0	0	0	0	0	0	0	0	0	0	0	0
4º	1775	2929	3230	2046	1539	1366	1906	2276	2217	2467	2172	1519
5º	0	0	0	0	0	0	0	0	0	0	0	0
6º	14597	16418	19695	15407	6691	7450	8020	8440	8068	7291	6706	4615
7º	487	290	123	141	123	100	111	83	111	117	127	117
8º	517	289	156	121	113	114	94	77	106	115	113	129
9º	181103	159509	116610	87110	46316	36308	46344	44385	40262	38015	34929	25982
10º	49934	53358	58751	65169	76272	76582	72876	50132	51148	51562	48727	47484
11º	32303	37284	33011	32098	35588	35273	34563	41780	41444	41410	42649	40283
12º	189908	170830	127657	94225	67471	56218	66391	76833	83284	87558	84818	82808

Ingressos



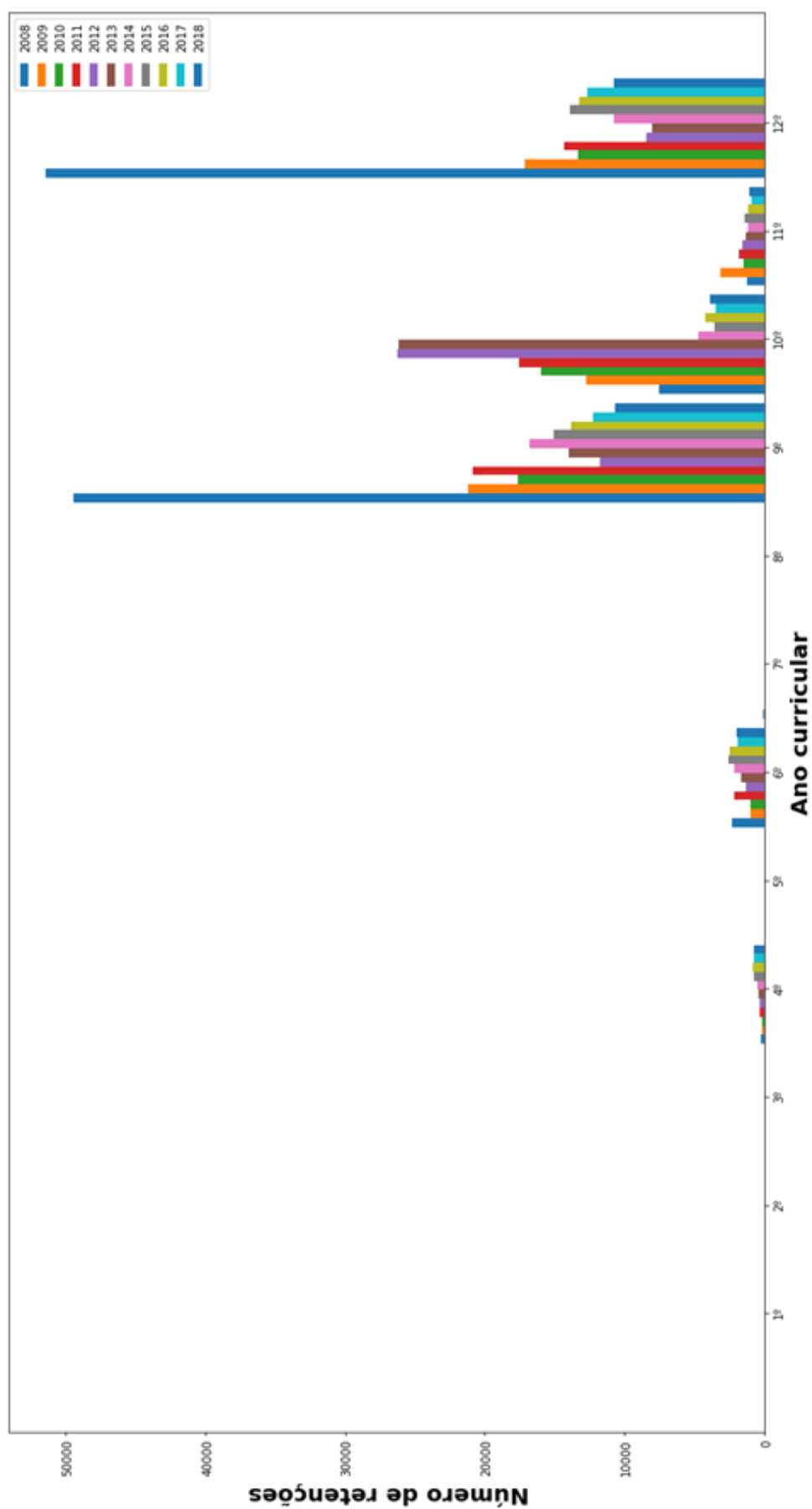
	Ano letivo (X/X+1)										
	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1º	0	0	0	0	0	0	0	0	0	0	0
2º	0	0	0	0	0	0	0	0	0	0	0
3º	0	0	0	0	0	0	0	0	0	0	0
4º	1766	2359	1761	1005	1001	1479	1642	1587	1625	1384	797
5º	0	0	0	0	0	0	0	0	0	0	0
6º	11798	16194	12964	3036	3194	3856	4469	4831	3886	3699	2347
7º	64	47	22	5	10	16	14	10	8	8	7
8º	32	14	9	6	12	5	5	27	2	2	1
9º	84828	73539	45963	11835	7868	14328	12200	15502	14831	12789	9133
10º	8975	9834	9882	16293	8221	7320	5754	6809	5798	5547	5078
11º	2390	2127	2388	1138	1215	936	1046	1945	1243	1044	1348
12º	77167	67499	45728	23244	13065	20883	20415	27441	31784	28462	28252

Transições



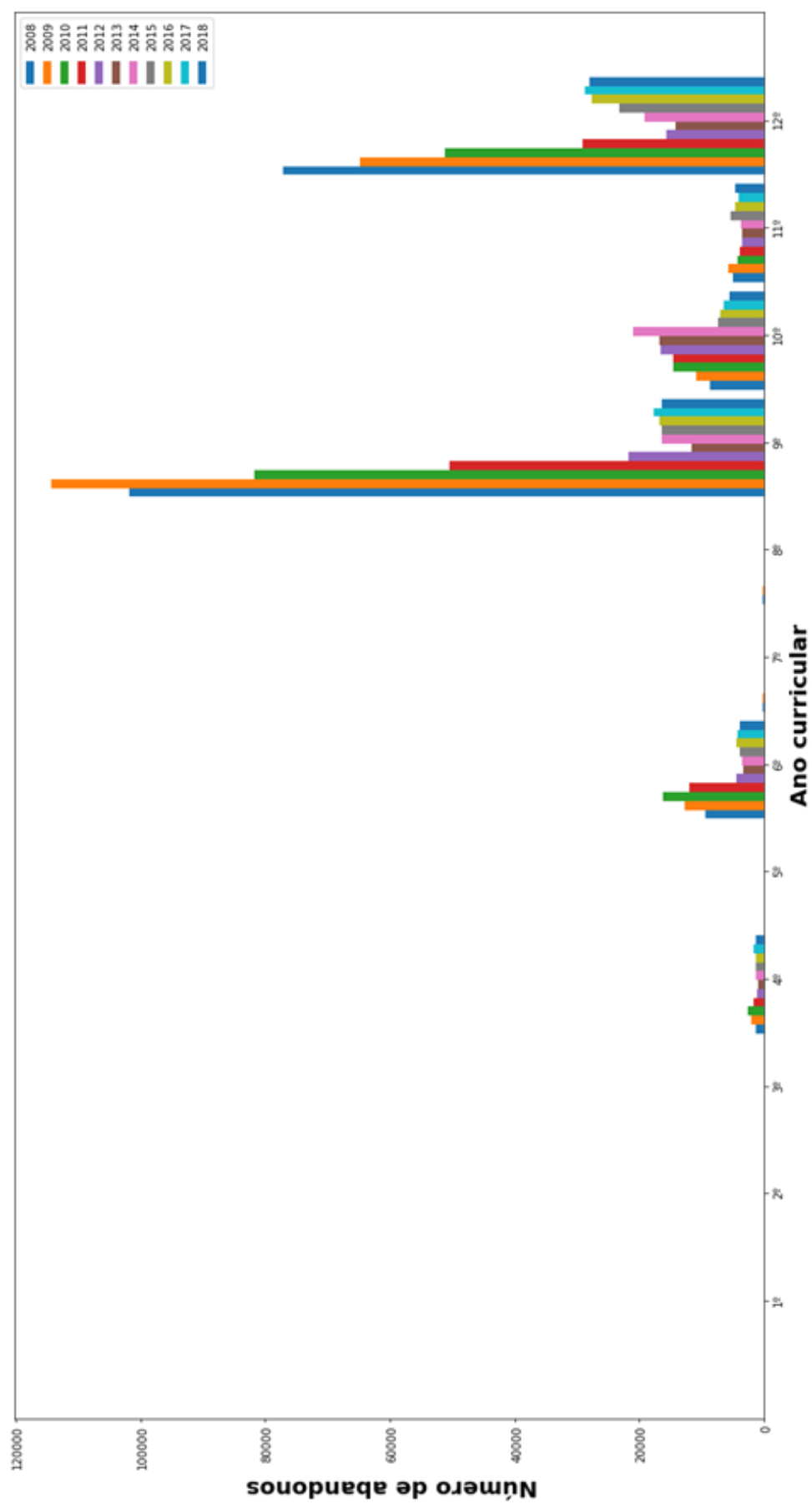
		Ano letivo (X/X+1)											
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Ano curricular	1º	0	0	0	0	0	0	0	0	0	0	0	
	2º	0	0	0	0	0	0	0	0	0	0	0	
	3º	0	0	0	0	0	0	0	0	0	0	0	
	4º	386	830	485	218	233	196	253	296	246	222	183	
	5º	0	0	0	0	0	0	0	0	0	0	0	
	6º	2882	2763	2444	1445	1073	2574	2410	2113	1241	1135	949	
	7º	289	120	100	120	104	92	75	68	109	111	123	
	8º	297	155	131	112	104	109	66	68	104	106	107	
	9º	29899	23917	17160	15826	12954	10762	13132	13028	9727	8127	7975	
	10º	33713	29854	28211	33223	33447	33708	47152	39192	40022	41666	39329	
	11º	26191	28461	27383	26366	30655	30424	29814	35203	35643	36417	36988	
	12º	61377	88958	63204	50882	43414	34122	36468	39736	42520	46160	46019	

Retenções



Ano letivo (X/X+1)												
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Ano curricular	1º	0	0	0	0	0	0	0	0	0	0	
	2º	0	0	0	0	0	0	0	0	0	0	
	3º	0	0	0	0	0	0	0	0	0	0	
	4º	228	168	131	277	328	371	490	716	791	702	
	5º	0	0	0	0	0	0	0	0	0	0	
	6º	2334	942	1014	2109	1309	1674	2133	2522	2462	1876	
	7º	79	0	0	3	3	1	0	1	0	1	
	8º	6	3	3	1	0	0	0	0	0	1	
	9º	49396	21157	17629	20827	11752	14002	16772	15021	13800	12249	
	10º	7534	12756	15945	17506	26242	26185	4686	3517	4215	3429	
	11º	1203	3161	1448	1823	1550	1297	1138	1363	1160	930	
	12º	51458	17153	13345	14348	8411	8005	10791	13868	13247	12665	

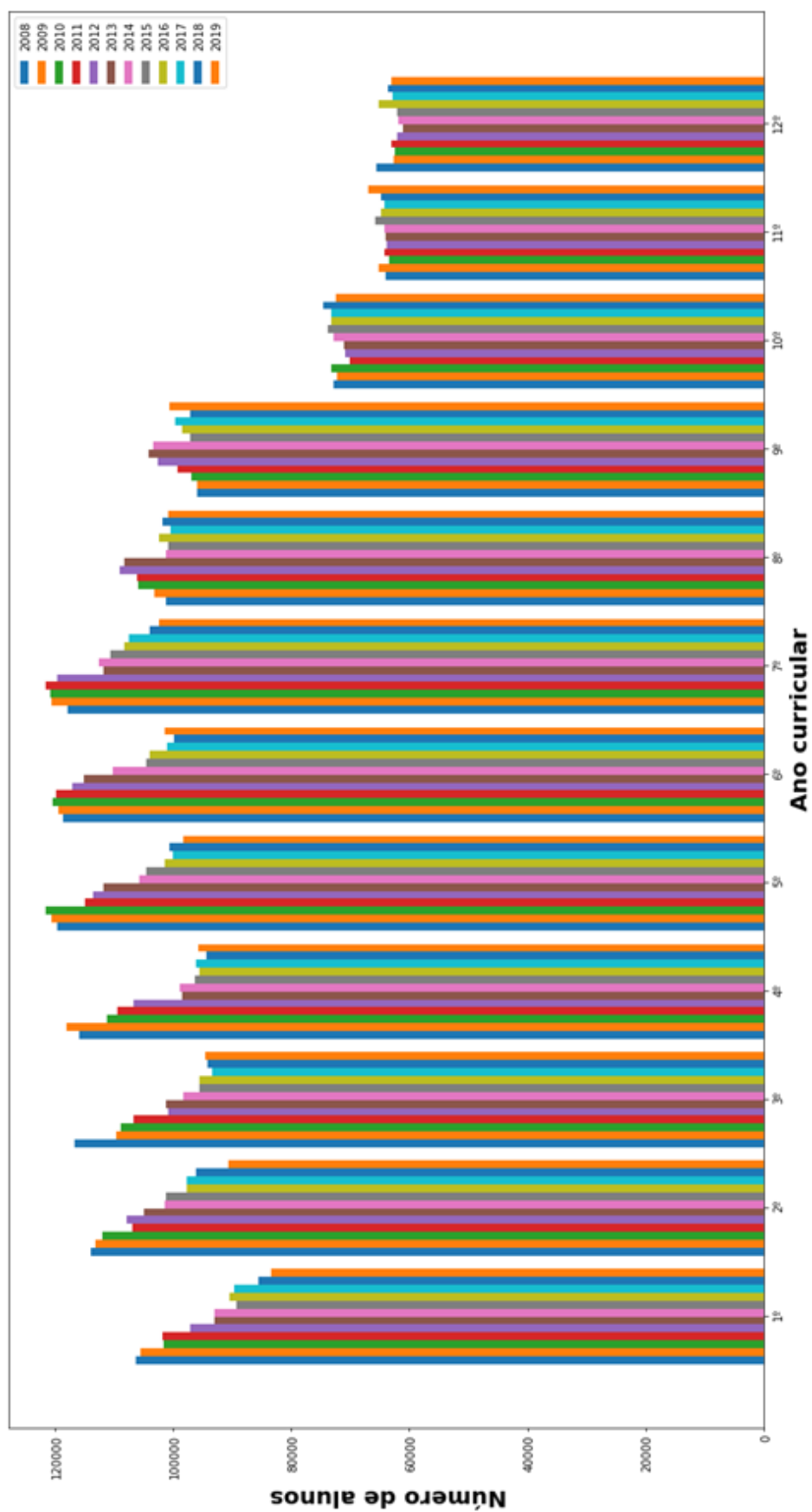
Abandonos



		Ano letivo (X/X+1)										
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Ano curricular	1º	0	0	0	0	0	0	0	0	0	0	0
	2º	0	0	0	0	0	0	0	0	0	0	0
	3º	0	0	0	0	0	0	0	0	0	0	0
	4º	1161	1931	2614	1551	978	799	1163	1264	1180	1532	1287
	5º	0	0	0	0	0	0	0	0	0	0	0
	6º	9381	12713	16237	11853	4309	3202	3477	3805	4365	4280	3821
	7º	119	170	23	18	16	7	36	14	2	5	0
	8º	214	131	22	8	9	5	28	9	2	8	5
	9º	101808	114435	81821	50457	21610	11544	16440	16336	16735	17639	16294
	10º	8687	10748	14595	14440	16583	16689	21038	7423	6911	6467	5496
	11º	4909	5662	4180	3909	3383	3552	3611	5214	4641	4063	4624
	12º	77073	64719	51108	28995	15646	14091	19132	23229	27517	28733	28049

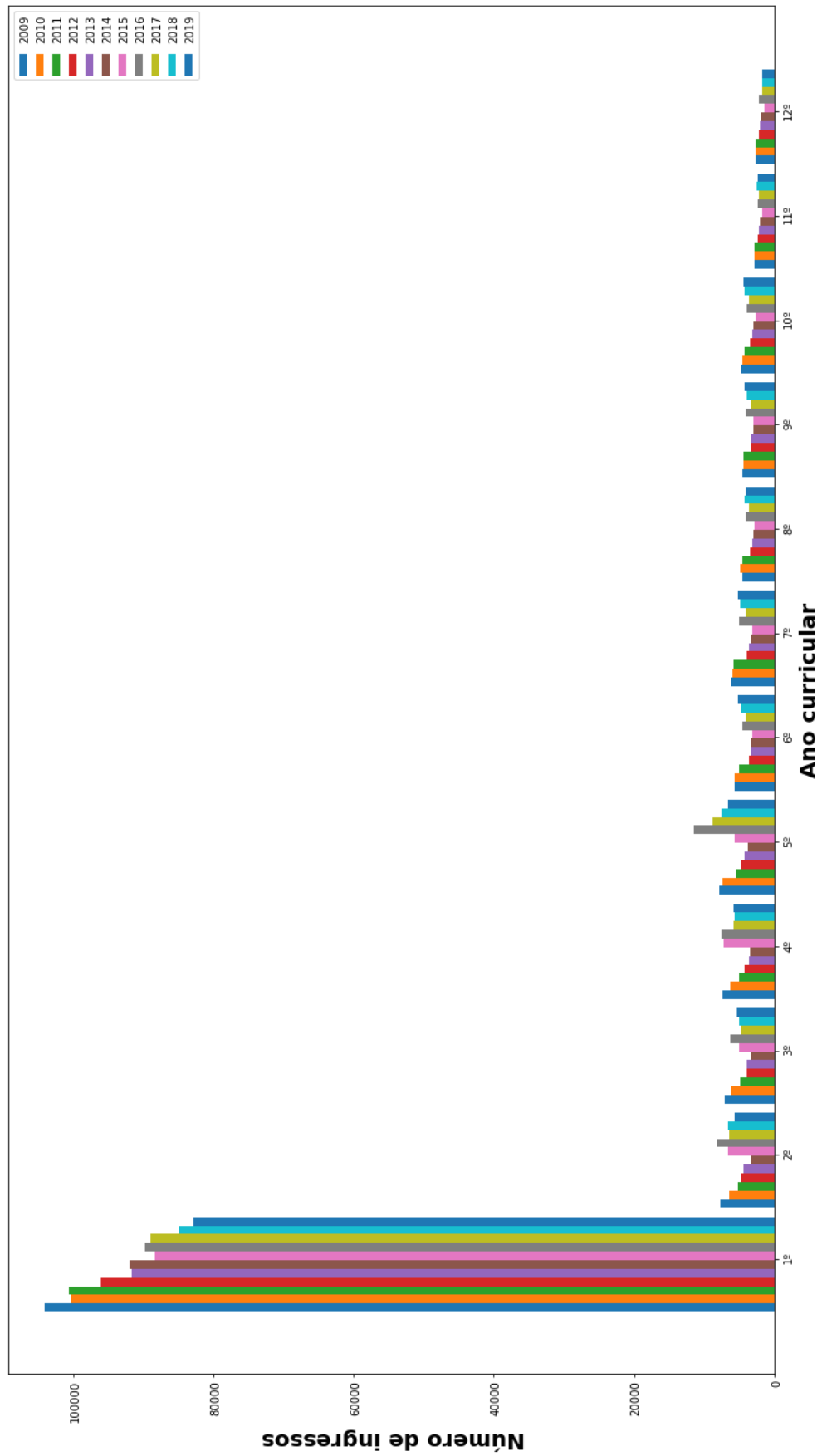
Apêndice C. Dados do ensino comum

Inscritos



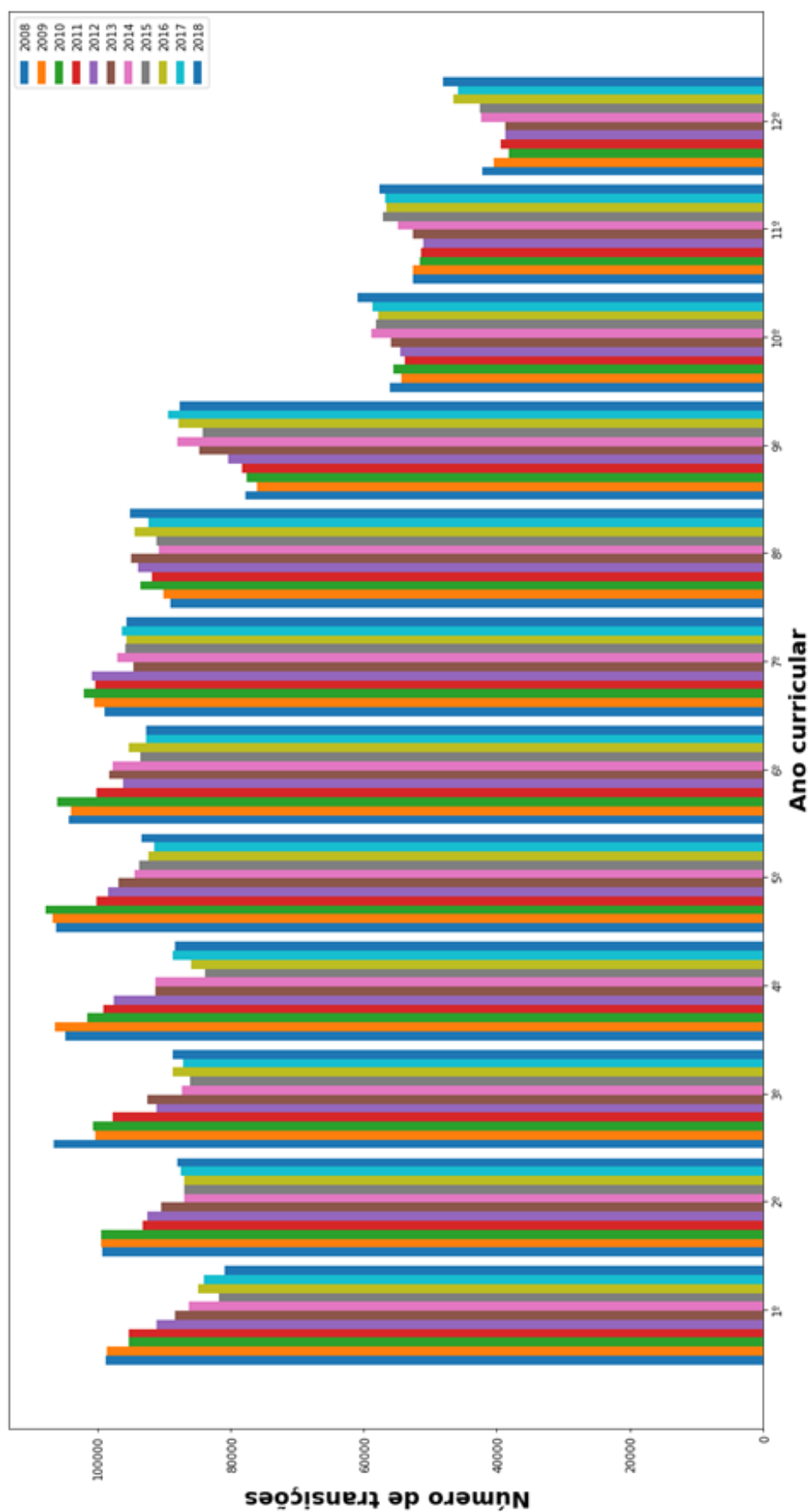
	Ano letivo (X/X+1)											
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1º	106204	105418	101584	101720	97177	92909	92930	89194	90507	89640	85507	83436
2º	113900	113234	111987	106844	107855	104837	101476	101264	97676	97587	96173	90666
3º	116772	109721	108785	106715	100751	101174	98312	95470	95553	93356	94211	94471
4º	115896	118138	111098	109503	106619	98544	98922	96316	95469	96129	94293	95728
5º	119585	120624	121662	114875	113553	111826	105769	104489	101452	100096	100681	98244
6º	118603	119486	120339	119893	116984	115067	110137	104469	103970	100920	99848	101384
7º	117809	120512	120728	121617	119555	111829	112623	110637	108175	107557	104031	102457
8º	101124	103179	105995	106158	109075	108347	101144	100888	102322	100481	101743	100782
9º	95969	95881	96925	99333	102529	104167	103265	97095	98448	99645	97109	100581
10º	72786	72272	73156	70099	70871	70991	72891	73866	73117	73263	74497	72322
11º	64042	65076	63439	64186	63813	63937	64239	65687	64780	64167	64771	66894
12º	65564	62576	62403	63052	62128	61128	61858	62061	65225	62847	63573	63112

Ingressos



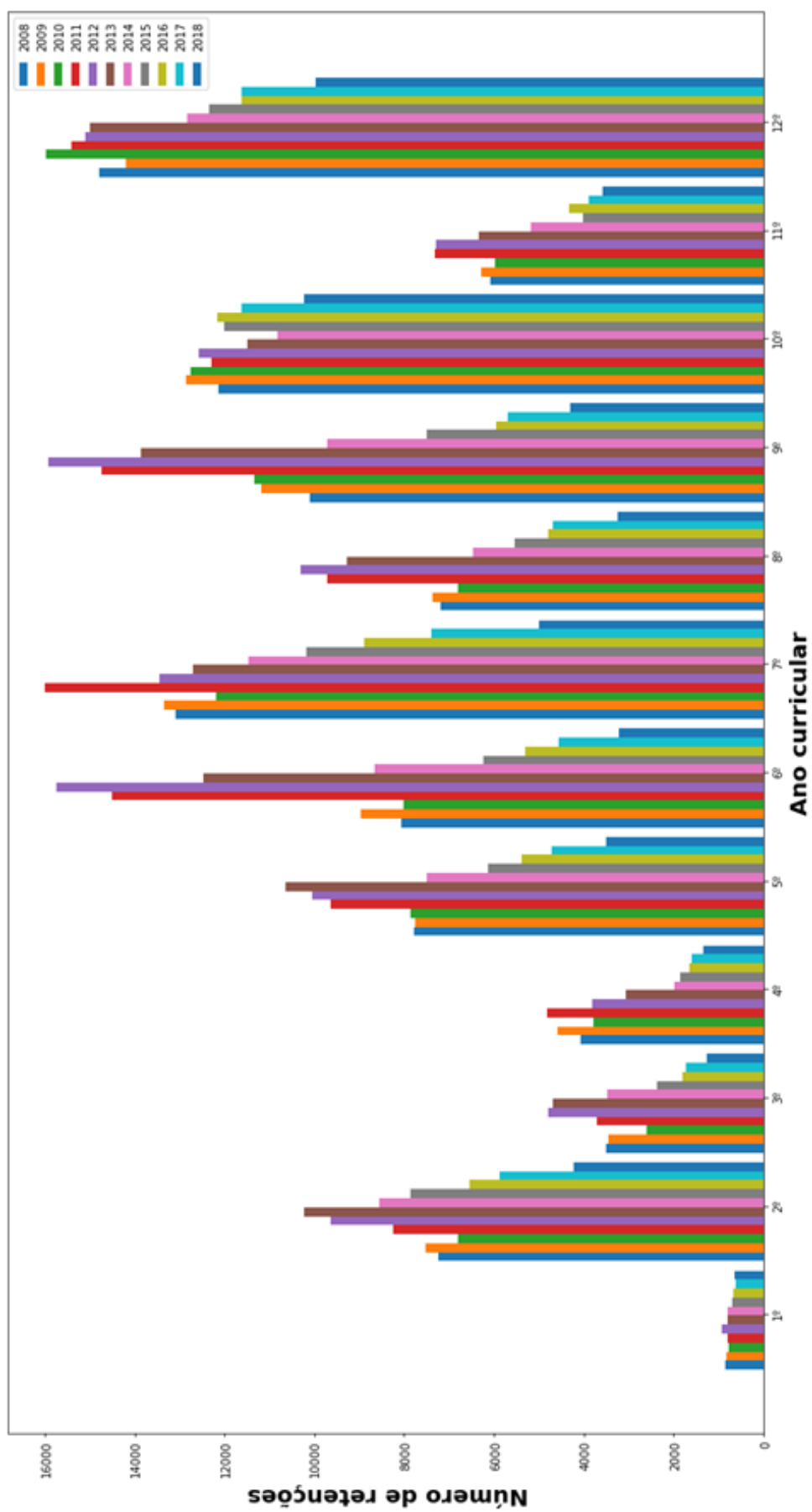
		Ano letivo (X/X+1)										
		2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Ano curricular	1º	104013	100269	100565	96119	91643	92029	88374	89784	88974	84913	82805
	2º	7736	6411	5131	4701	4318	3302	6498	8090	6388	6499	5709
	3º	7067	6176	4885	3975	3969	3197	5070	6187	4746	4994	5315
	4º	7437	6299	5015	4220	3650	3358	7201	7575	5832	5667	5804
	5º	7819	7400	5425	4644	4232	3787	5606	11508	8842	7451	6499
	6º	5582	5676	4985	3532	3292	3207	3114	4558	4010	4729	5124
	7º	6038	5910	5725	3974	3529	3333	3109	4939	4128	4813	5226
	8º	4525	4828	4566	3383	3125	2958	2724	4128	3600	4225	4090
	9º	4539	4352	4381	3288	3233	2959	3019	4010	3341	3875	4190
	10º	4685	4461	4196	3415	3163	2992	2649	3897	3568	4226	4321
	11º	2844	2863	2738	2308	2204	2062	1726	2389	2122	2414	2374
	12º	2572	2599	2653	2196	1986	1794	1437	2093	1664	1734	1686

Transições



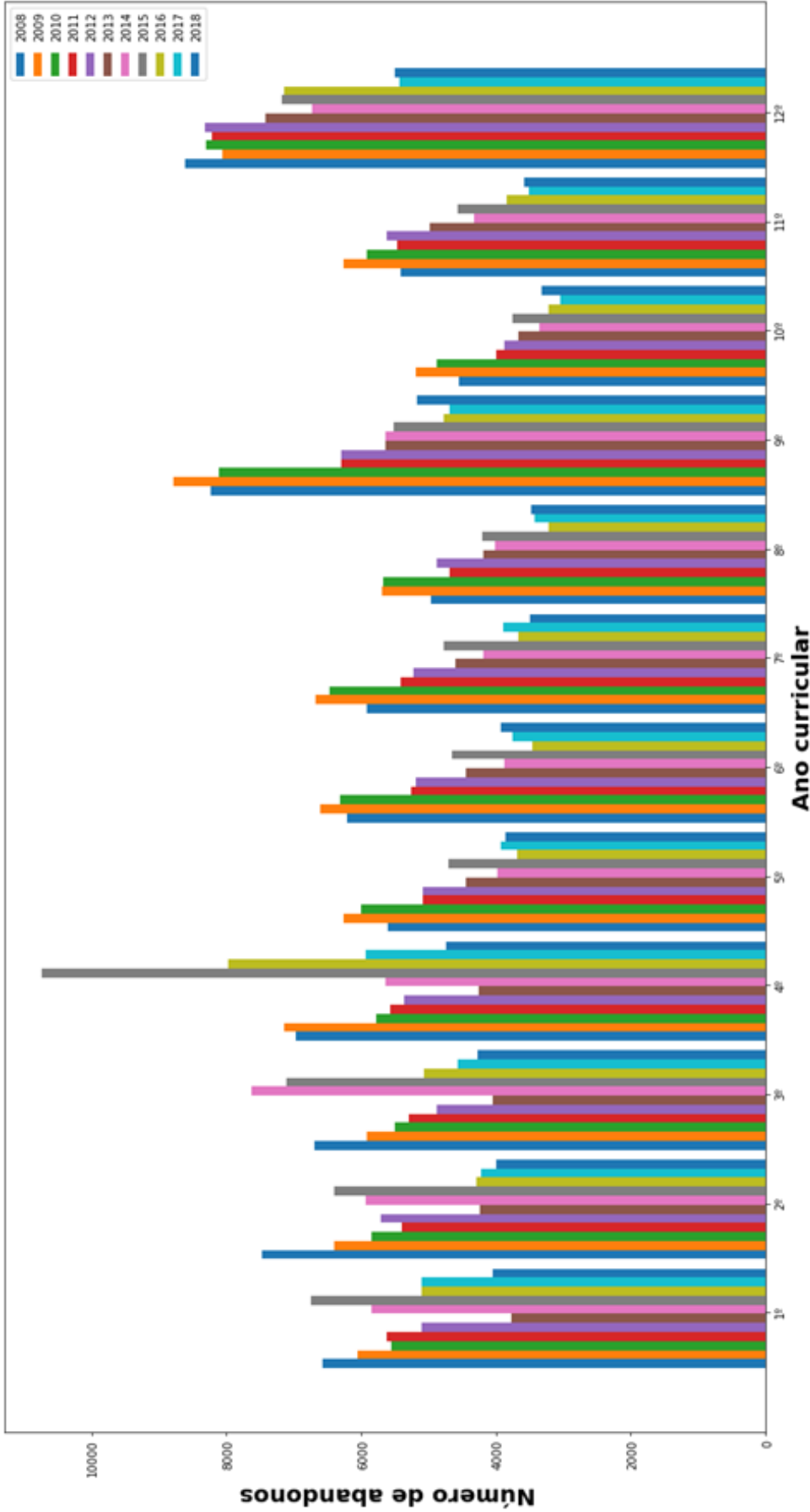
	Ano letivo (X/X+1)											
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
1º	98798	98548	95265	95323	91156	88345	86292	81773	84760	83938	80816	
2º	99213	99337	99360	93200	92504	90385	87002	87002	86841	87505	87961	
3º	106578	100342	100701	97701	91081	92449	87225	86001	88681	87061	88687	
4º	104843	106419	101547	99135	97471	91216	91312	83711	85858	88596	88237	
5º	106203	106625	107825	100162	98446	96733	94303	93678	92378	91459	93344	
6º	104334	103927	106017	100136	96043	98157	97613	93583	95220	92613	92702	
7º	98823	100490	102076	100199	100883	94528	96967	95691	95613	96288	95541	
8º	88994	90118	93521	91764	93909	94895	90679	91146	94329	92375	95036	
9º	77636	75939	77473	78321	80324	84684	87925	84097	87726	89269	87644	
10º	56099	54245	55528	53824	54415	55832	58717	58133	57752	58597	60953	
11º	52565	52550	51560	51395	50918	52620	54761	57106	56608	56783	57607	
12º	42187	40331	38132	39436	38720	38729	42300	42543	46482	45802	48117	

Retenções



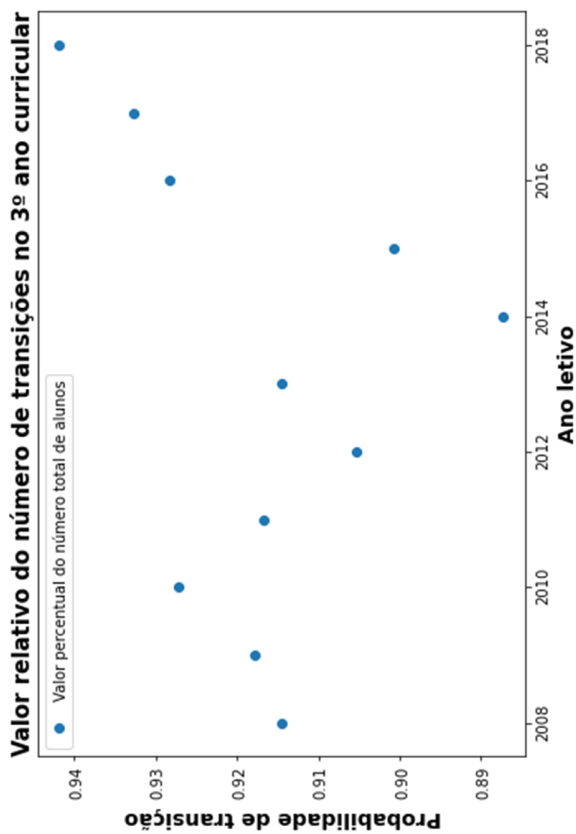
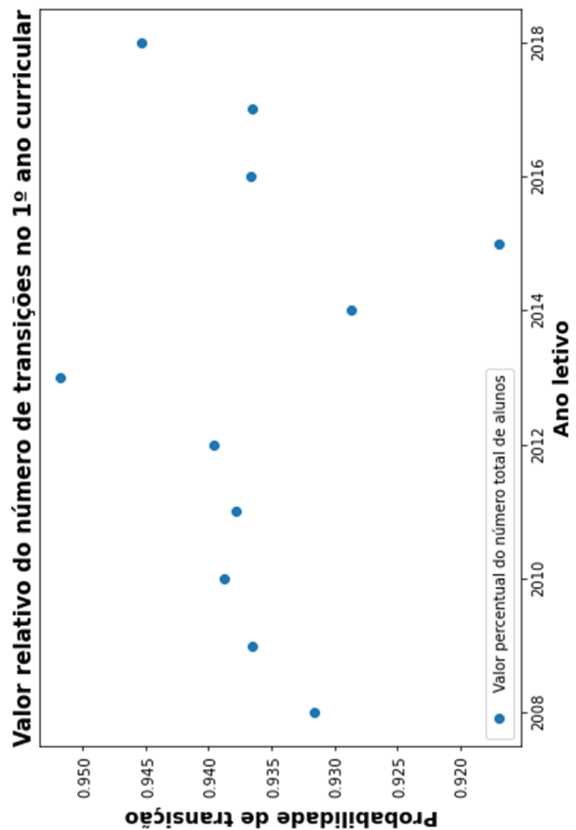
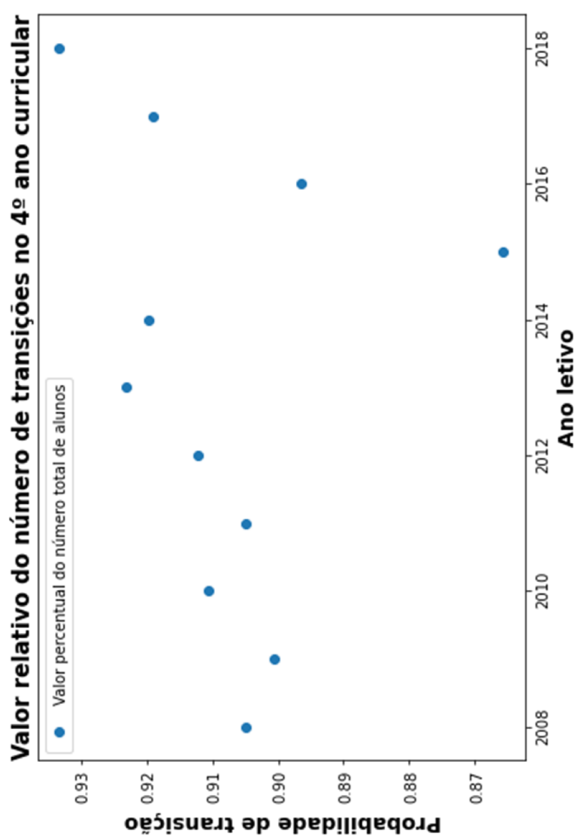
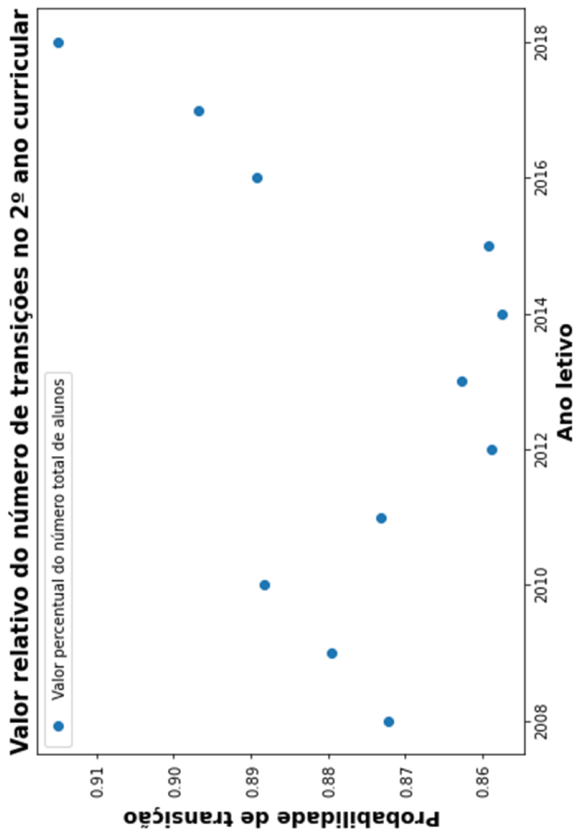
		Ano letivo (X/X+1)										
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Ano curricular	1º	835	826	776	784	915	798	792	685	655	602	648
	2º	7228	7508	6793	8249	9644	10216	8540	7863	6547	5859	4219
	3º	3503	3460	2595	3717	4796	4681	3465	2373	1810	1727	1252
	4º	4079	4584	3779	4802	3798	3069	1973	1861	1640	1600	1323
	5º	7787	7742	7843	9638	10032	10650	7496	6116	5390	4711	3487
	6º	8064	8957	8014	14510	15750	12467	8645	6239	5293	4558	3224
	7º	13083	13344	12182	16009	13453	12709	11469	10175	8895	7382	5000
	8º	7178	7368	6810	9711	10297	9277	6458	5548	4781	4682	3237
	9º	10105	11168	11336	14727	15914	13848	9702	7482	5960	5687	4301
	10º	12140	12841	12747	12279	12577	11496	10817	11992	12146	11618	10231
	11º	6075	6270	5964	7323	7282	6339	5166	4015	4332	3875	3585
	12º	14774	14186	15974	15408	15099	14990	12827	12347	11612	11619	9967

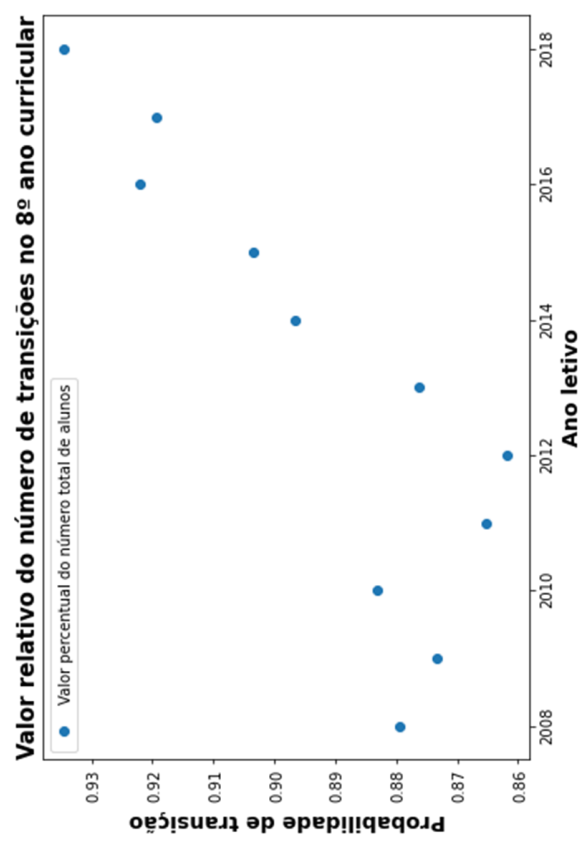
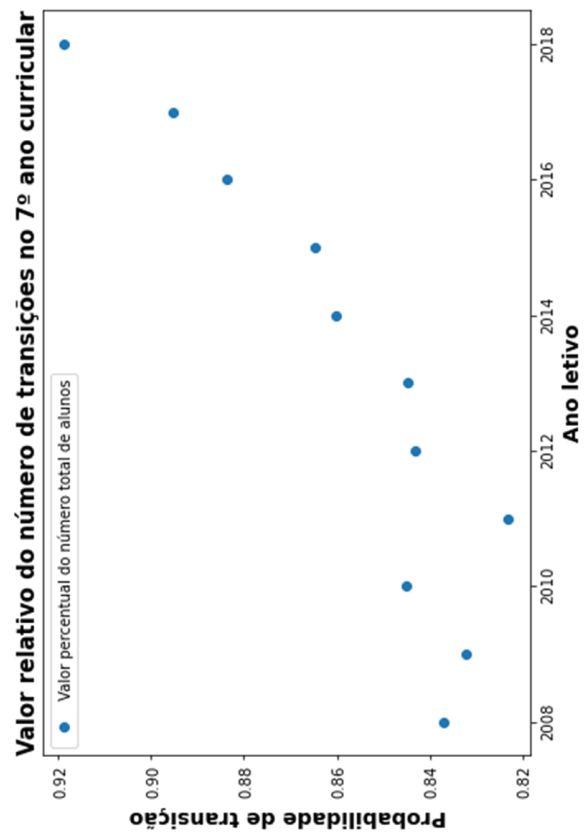
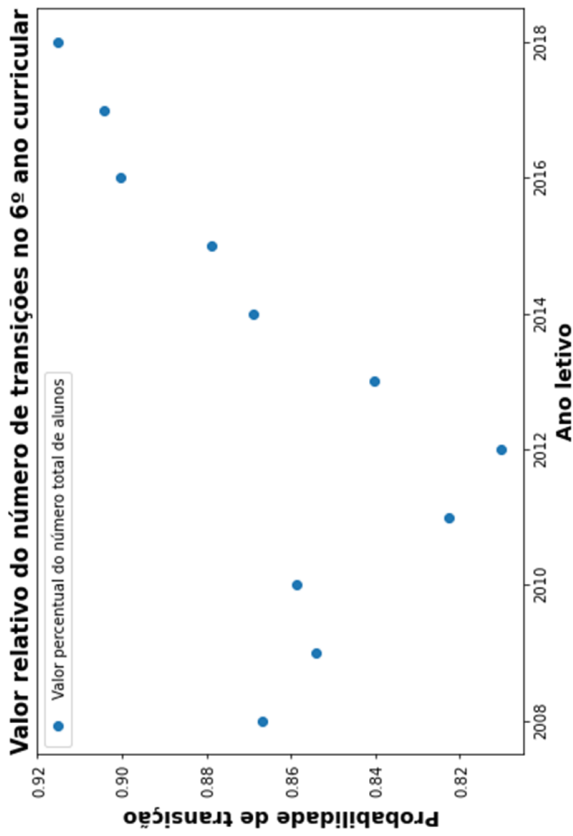
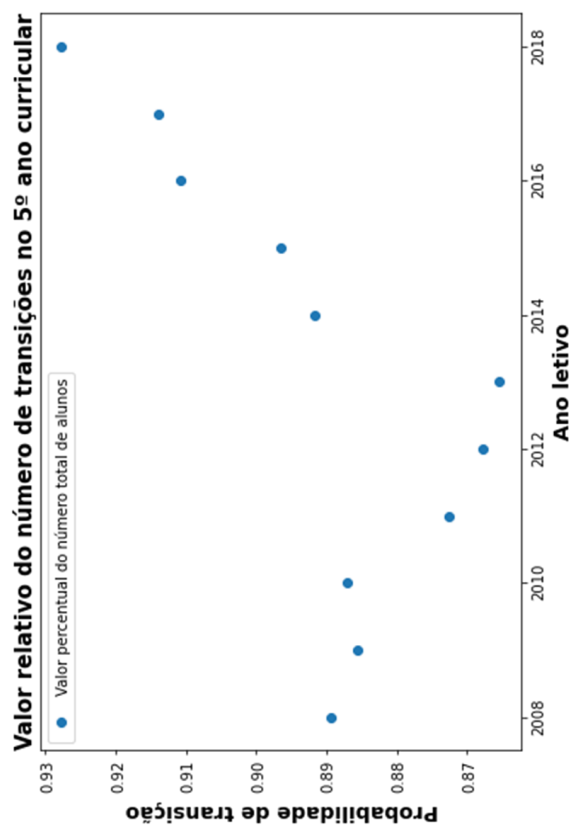
Abandonos



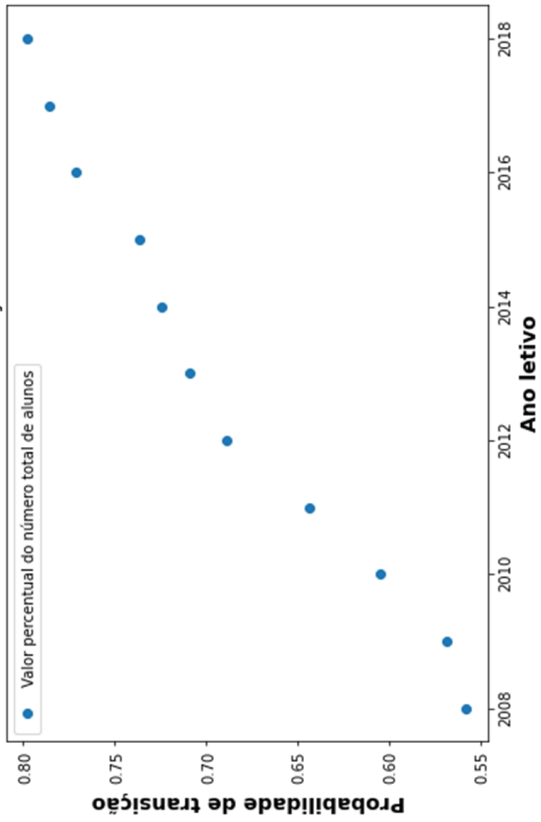
Ano letivo (X/X+1)												
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Ano curricular	1º	6571	6044	5543	5613	5106	3766	5846	6736	5092	5100	4043
	2º	7459	6389	5834	5395	5707	4236	5934	6399	4288	4223	3993
	3º	6691	5919	5489	5297	4874	4044	7622	7096	5062	4568	4272
	4º	6974	7135	5772	5566	5350	4259	5637	10744	7971	5933	4733
	5º	5595	6257	5994	5075	5075	4443	3970	4695	3684	3926	3850
	6º	6205	6602	6308	5247	5191	4443	3879	4647	3457	3749	3922
	7º	5903	6678	6470	5409	5219	4592	4187	4771	3667	3887	3490
	8º	4952	5693	5664	4683	4869	4175	4007	4194	3212	3424	3470
	9º	8228	8774	8116	6285	6291	5635	5638	5516	4762	4689	5164
	10º	4547	5186	4881	3996	3879	3663	3357	3741	3219	3048	3313
	11º	5402	6256	5915	5468	5613	4978	4312	4566	3840	3509	3579
	12º	8603	8059	8297	8208	8309	7409	6731	7171	7131	5426	5489

Apêndice D. Visualização dos valores relativos de cada tipo de resultado escolar para cada ano curricular (ensino comum) Transições

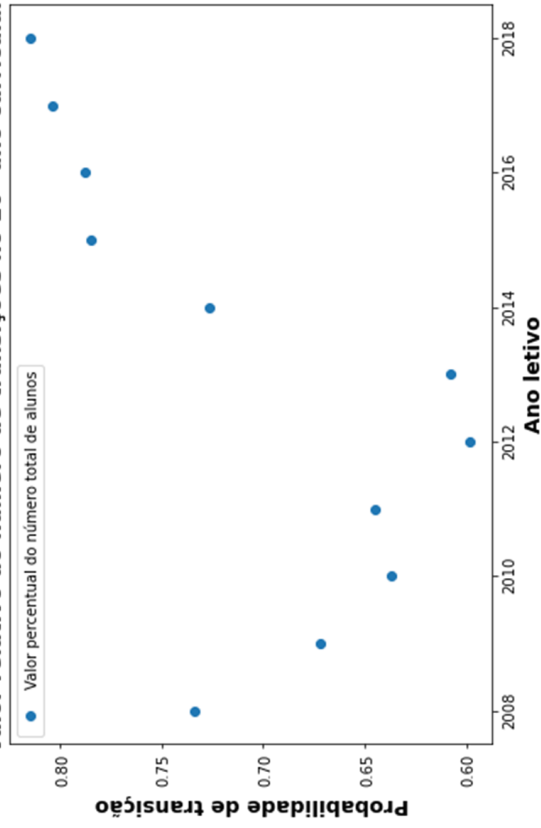




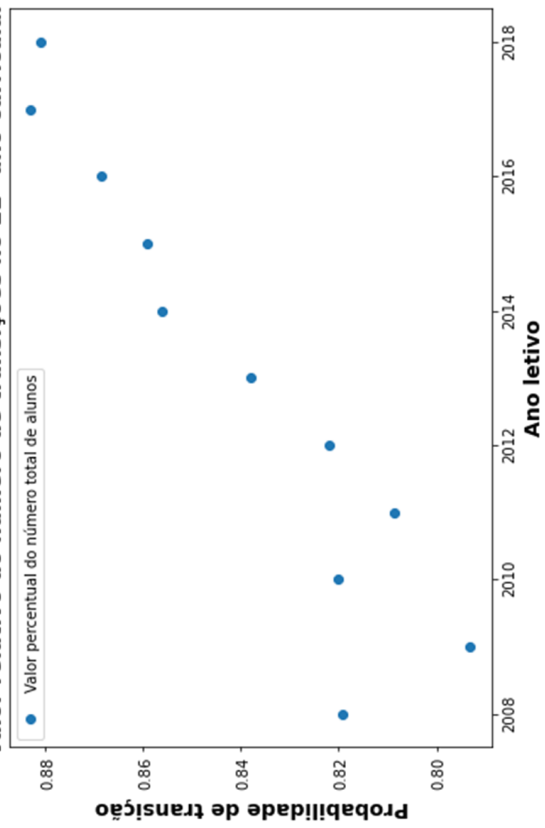
Valor relativo do número de transições no 9º ano curricular



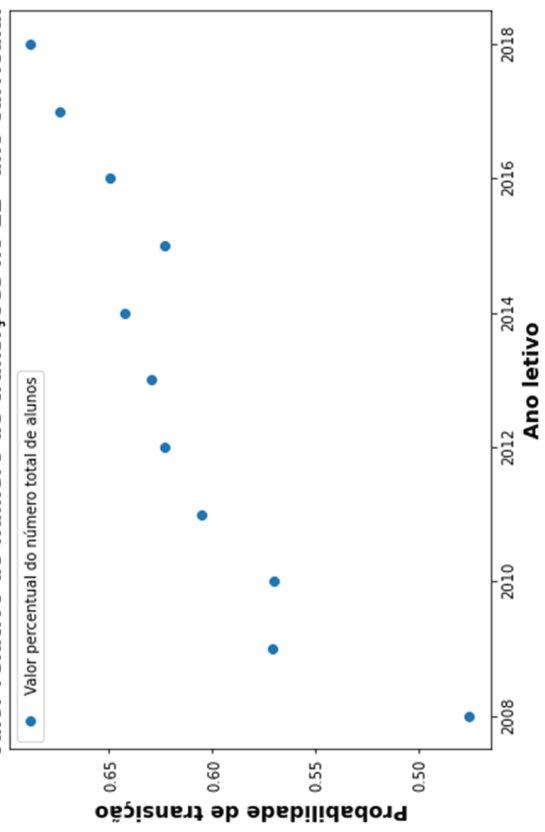
Valor relativo do número de transições no 10º ano curricular



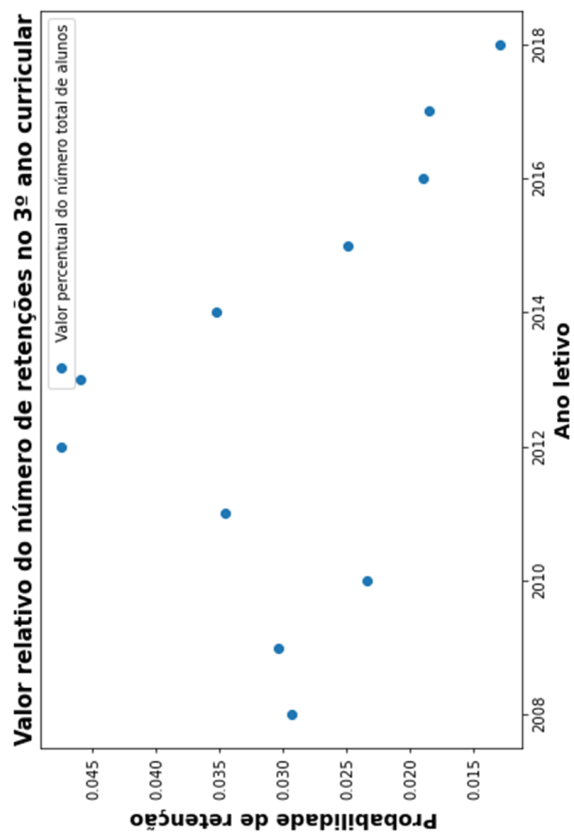
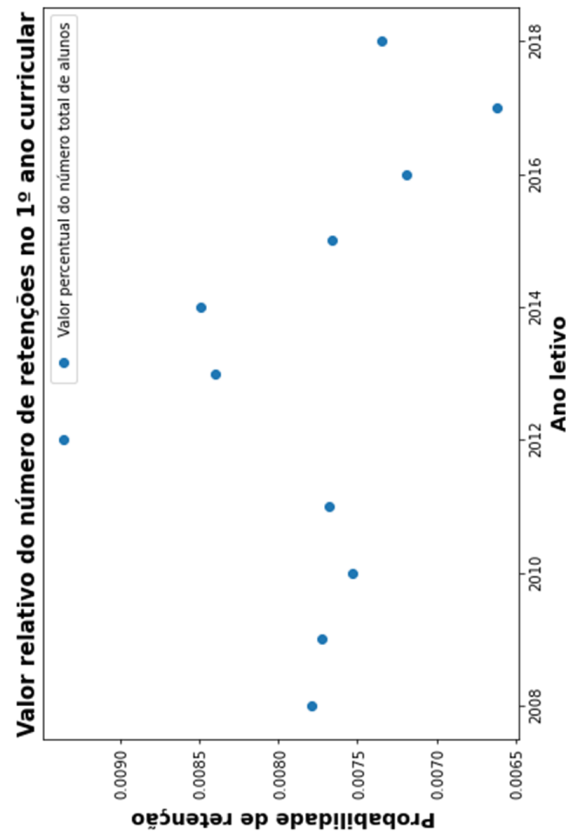
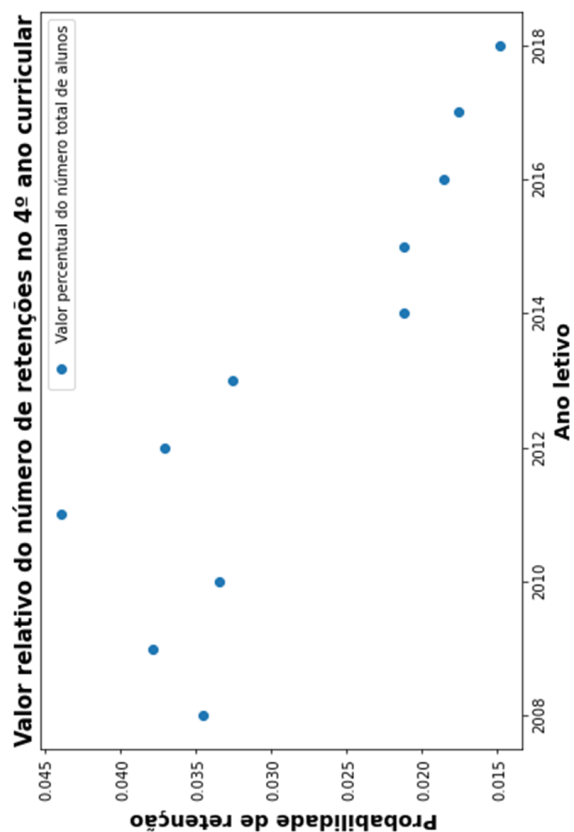
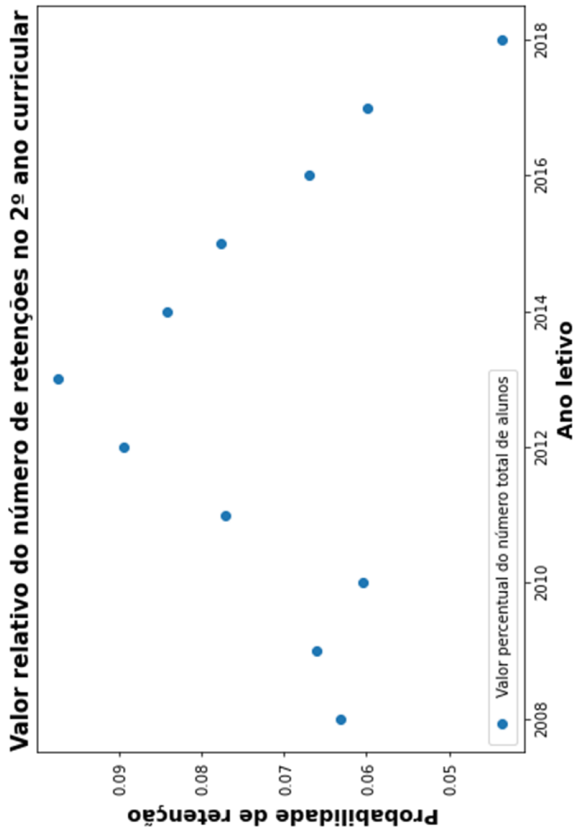
Valor relativo do número de transições no 11º ano curricular



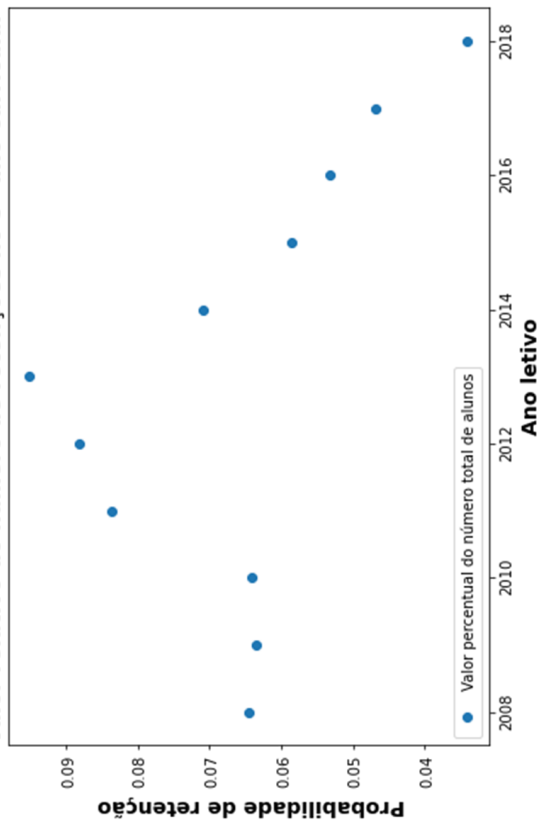
Valor relativo do número de transições no 12º ano curricular



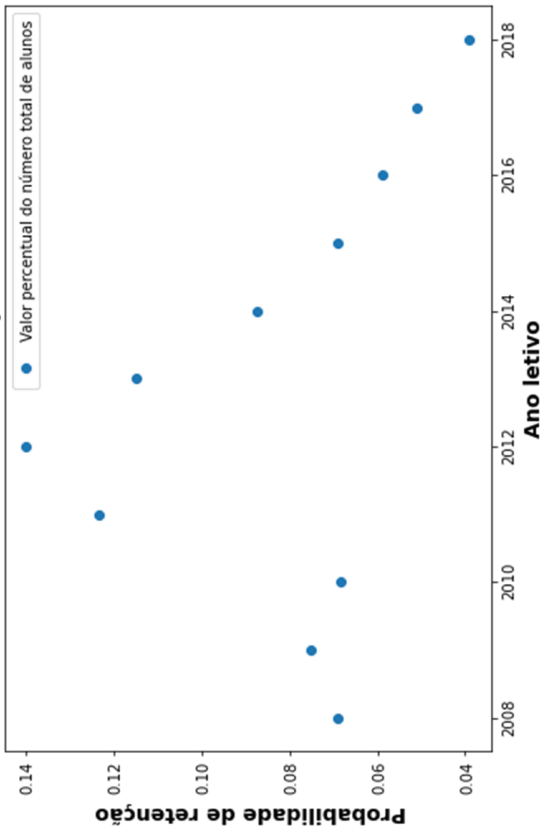
Retenções



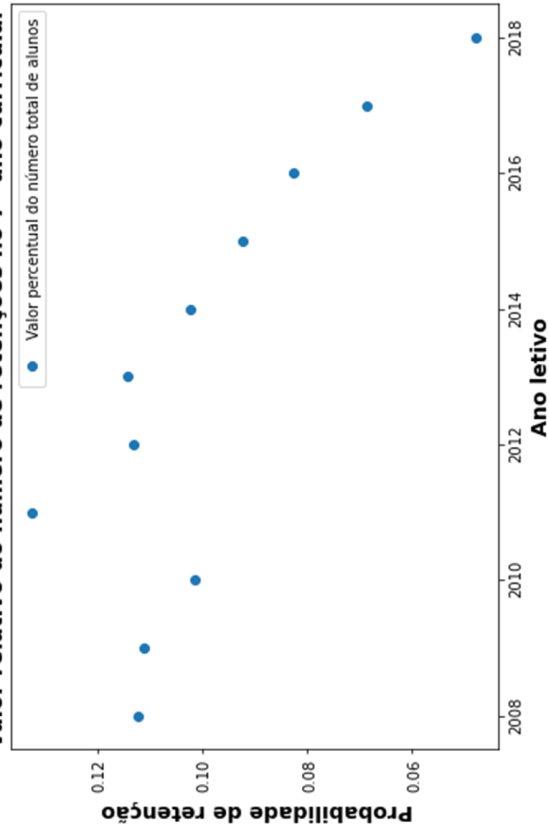
Valor relativo do número de retenções no 5º ano curricular



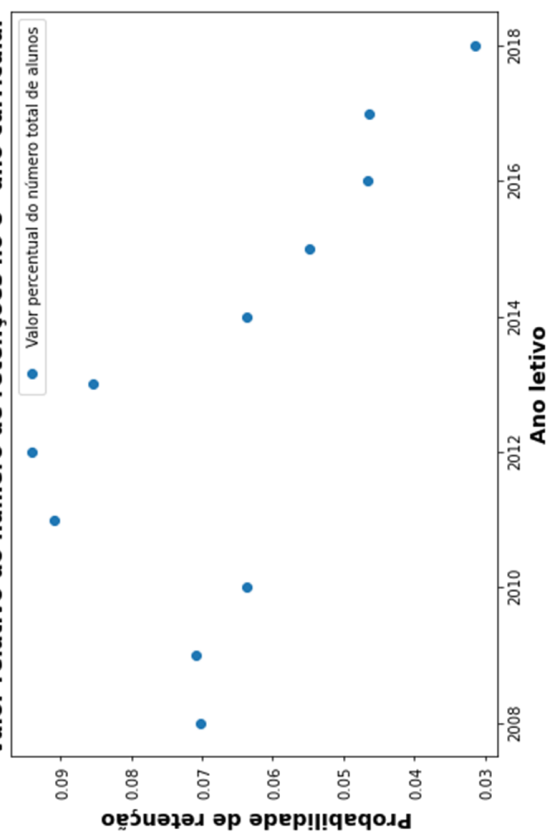
Valor relativo do número de retenções no 6º ano curricular



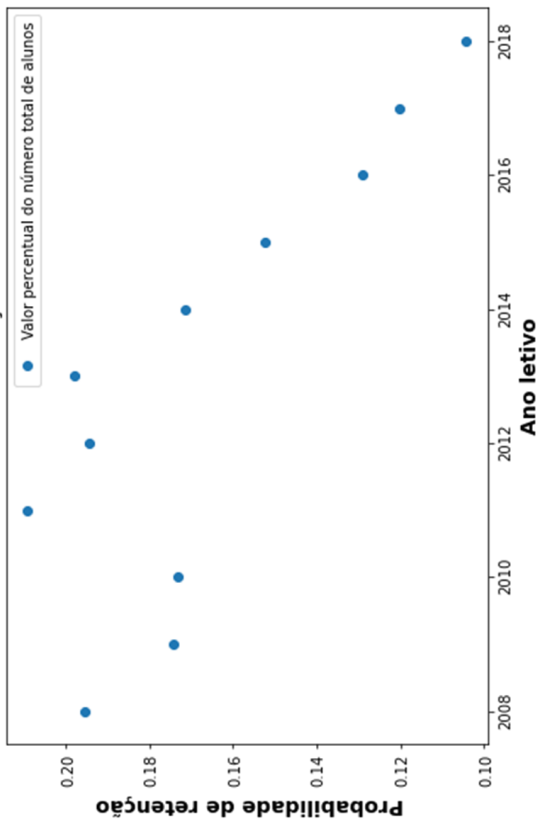
Valor relativo do número de retenções no 7º ano curricular



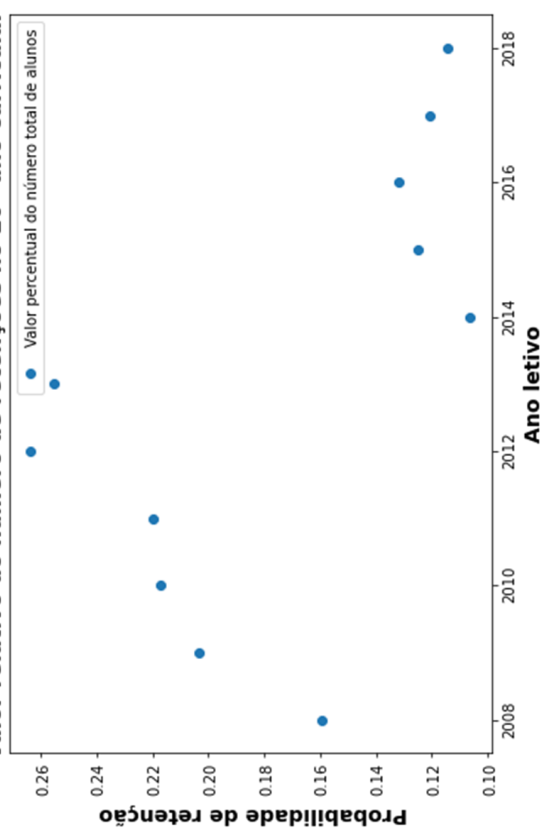
Valor relativo do número de retenções no 8º ano curricular



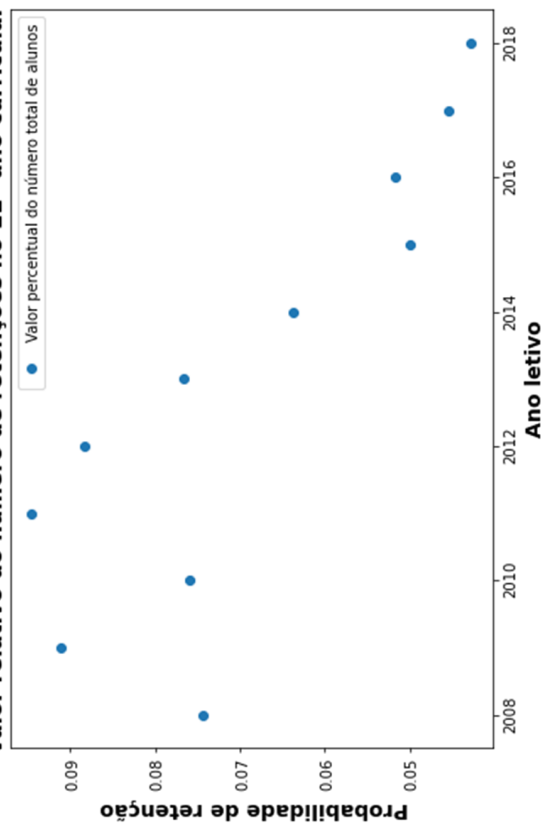
Valor relativo do número de retenções no 9º ano curricular



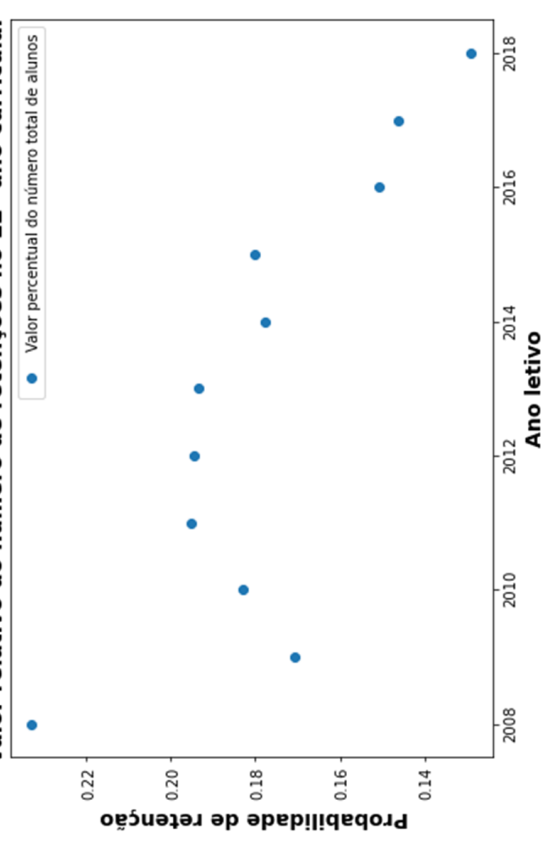
Valor relativo do número de retenções no 10º ano curricular



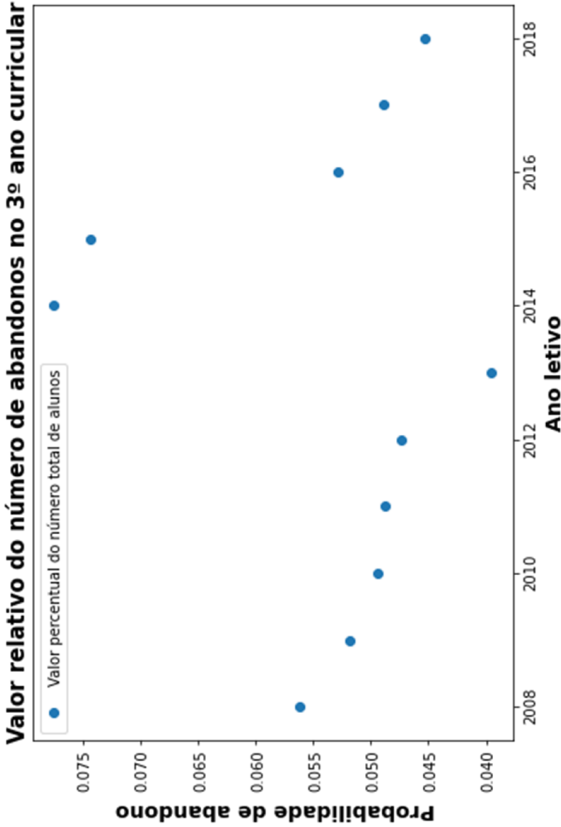
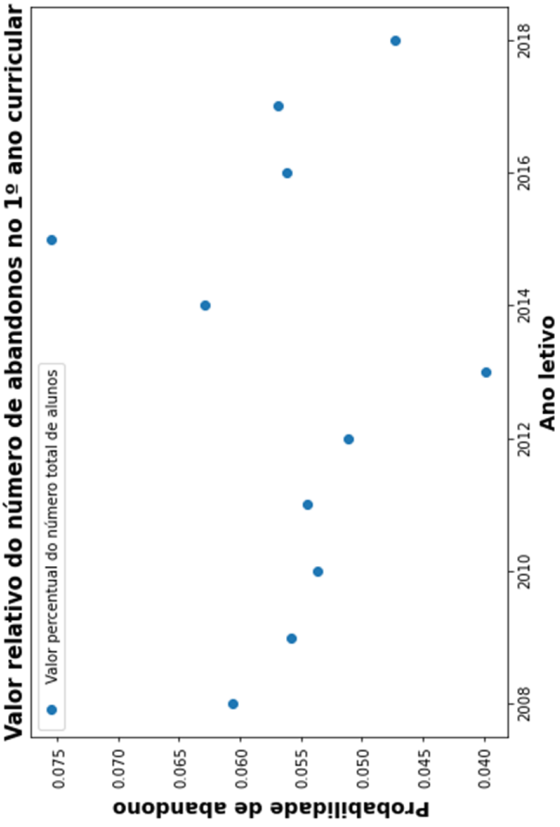
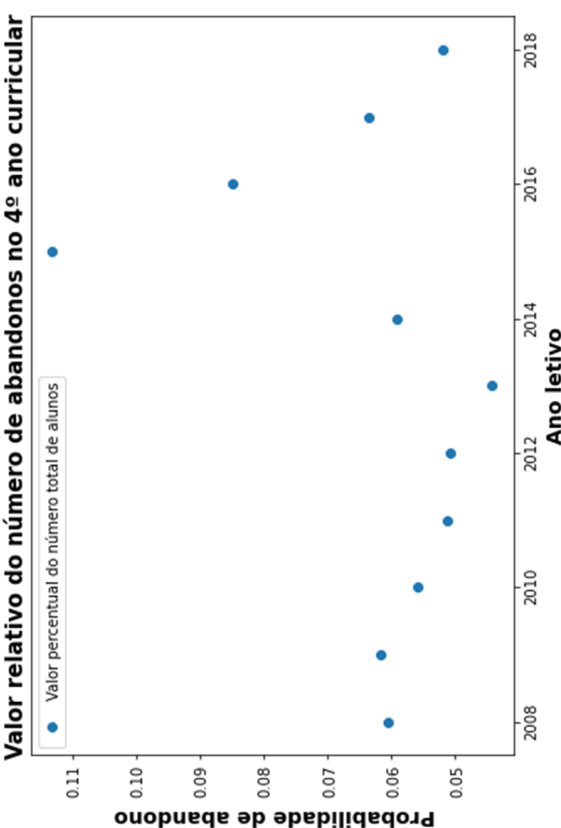
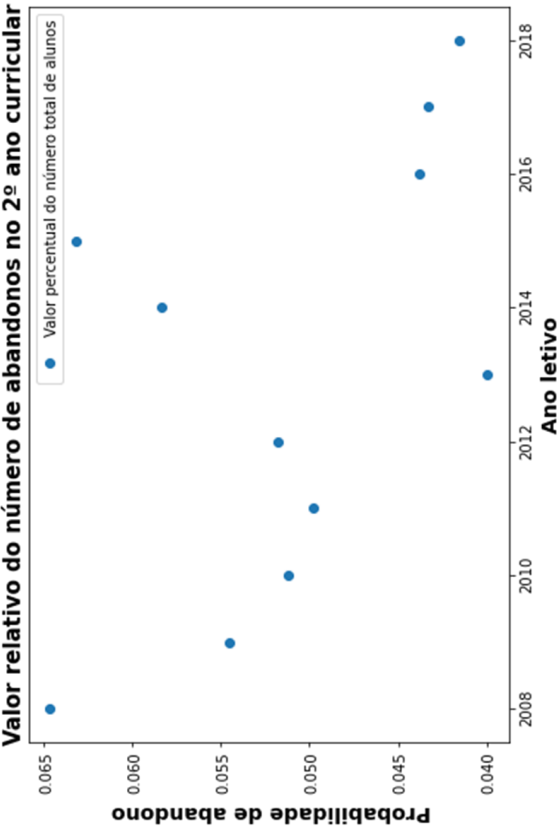
Valor relativo do número de retenções no 11º ano curricular



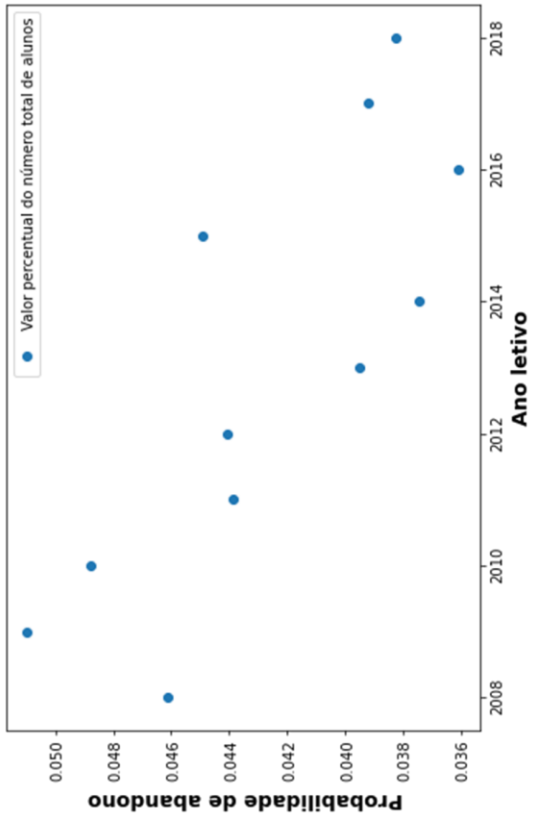
Valor relativo do número de retenções no 12º ano curricular



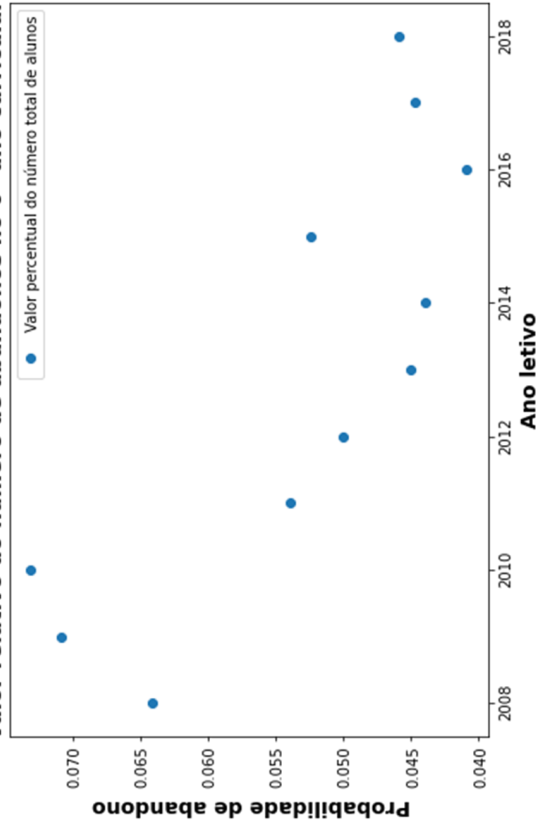
Abandonos



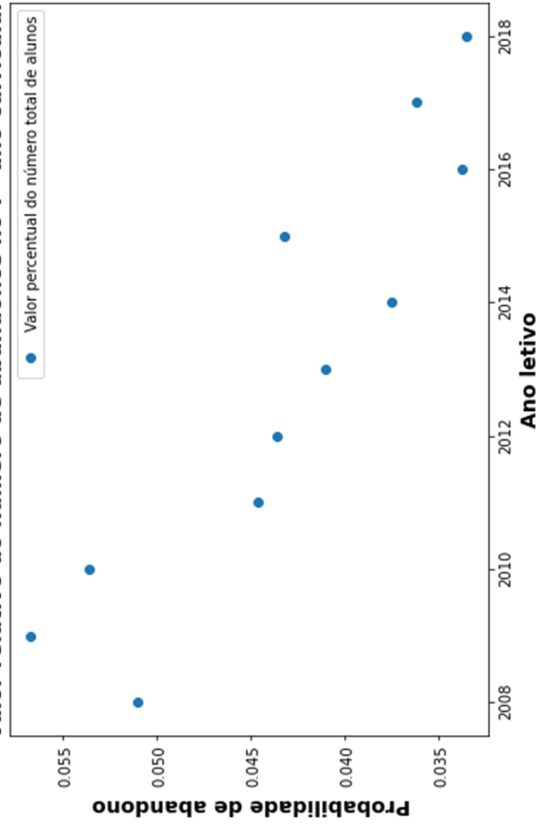
Valor relativo do número de abandonos no 5º ano curricular



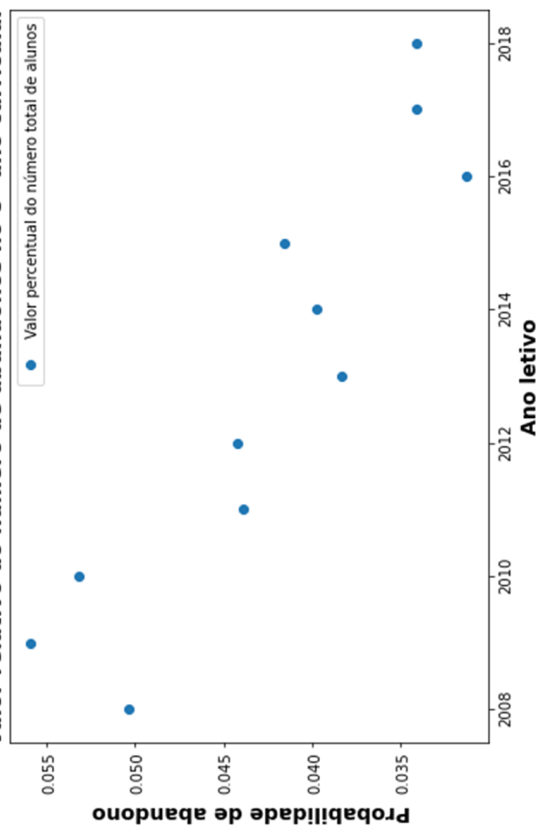
Valor relativo do número de abandonos no 6º ano curricular



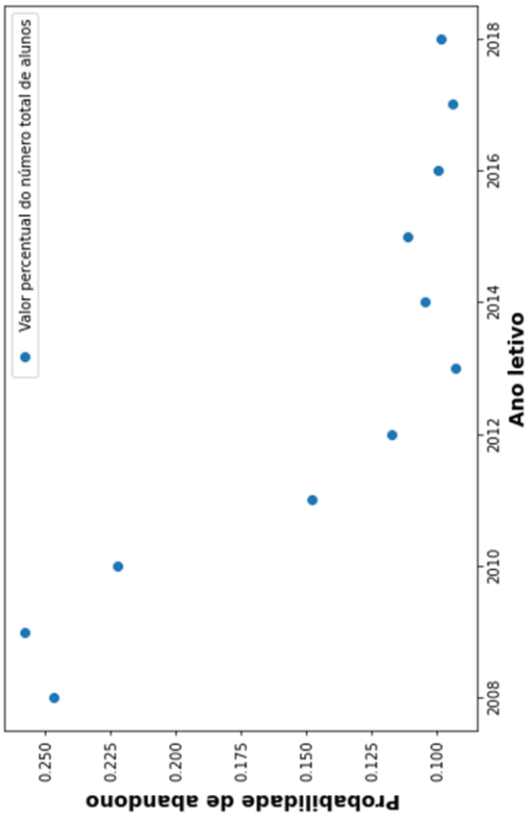
Valor relativo do número de abandonos no 7º ano curricular



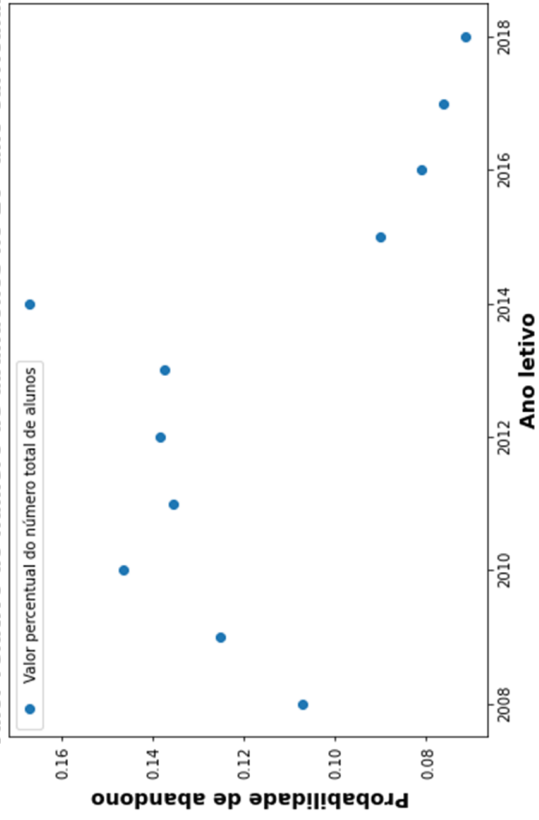
Valor relativo do número de abandonos no 8º ano curricular



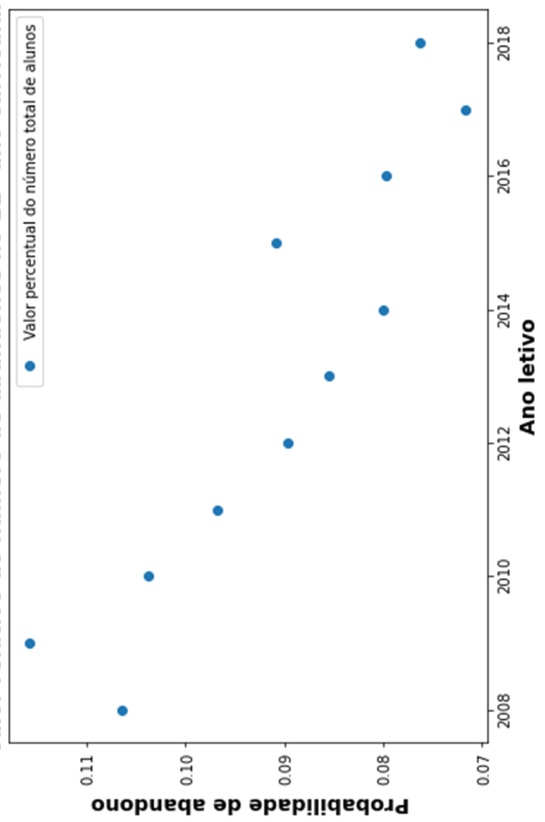
Valor relativo do número de abandonos no 9º ano curricular



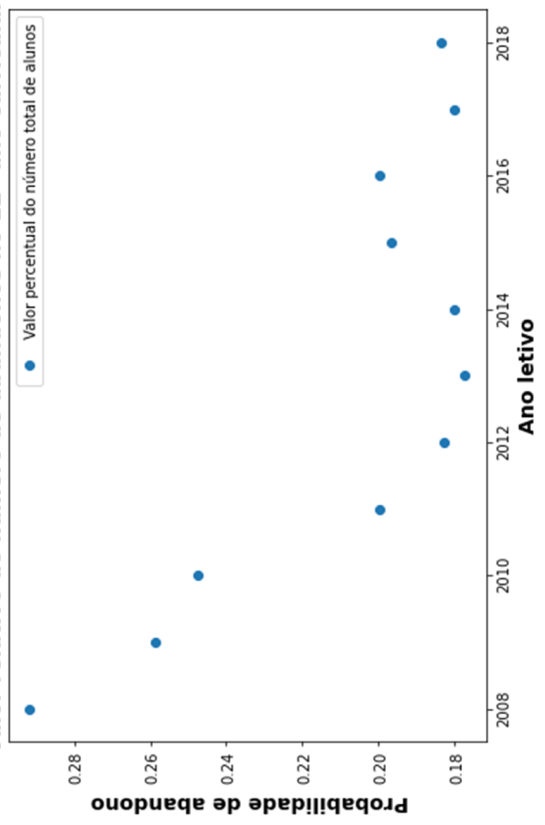
Valor relativo do número de abandonos no 10º ano curricular



Valor relativo do número de abandonos no 11º ano curricular



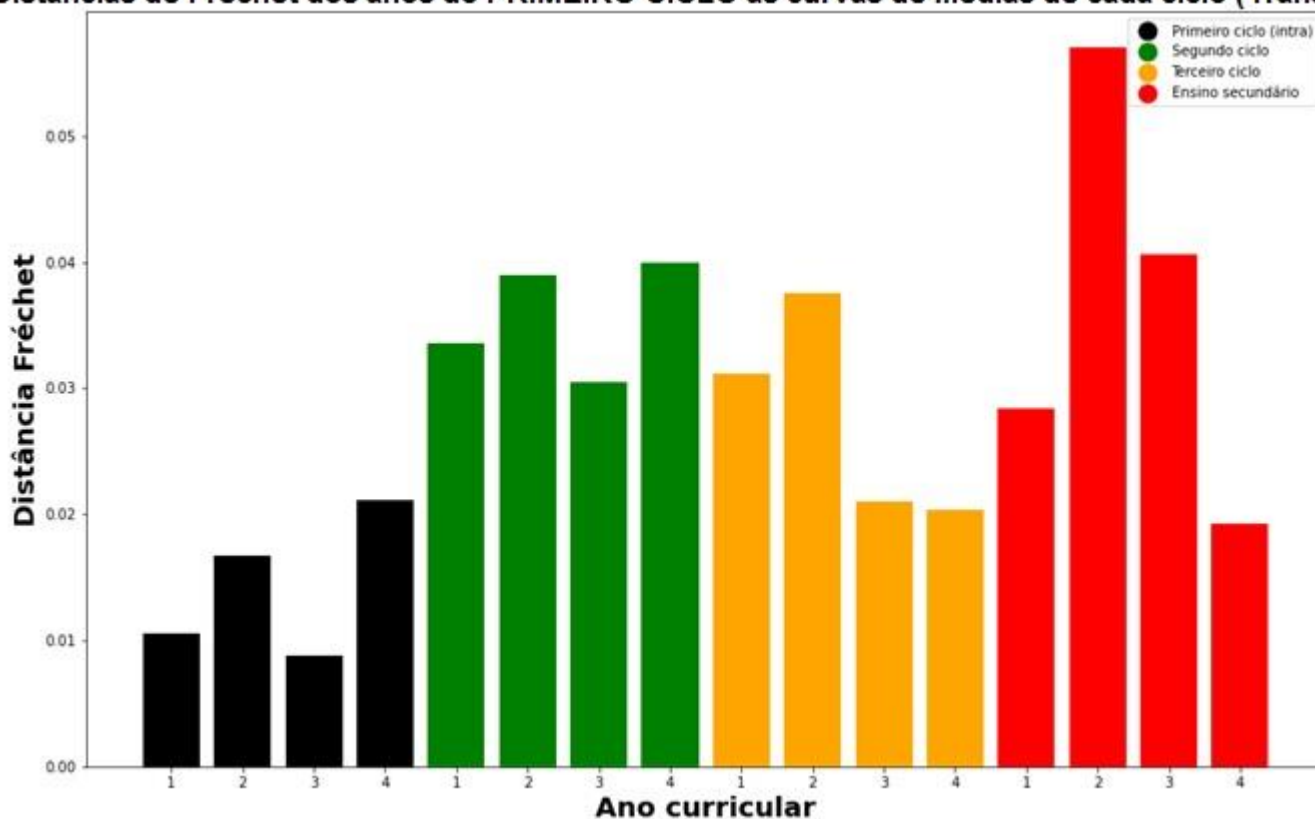
Valor relativo do número de abandonos no 12º ano curricular



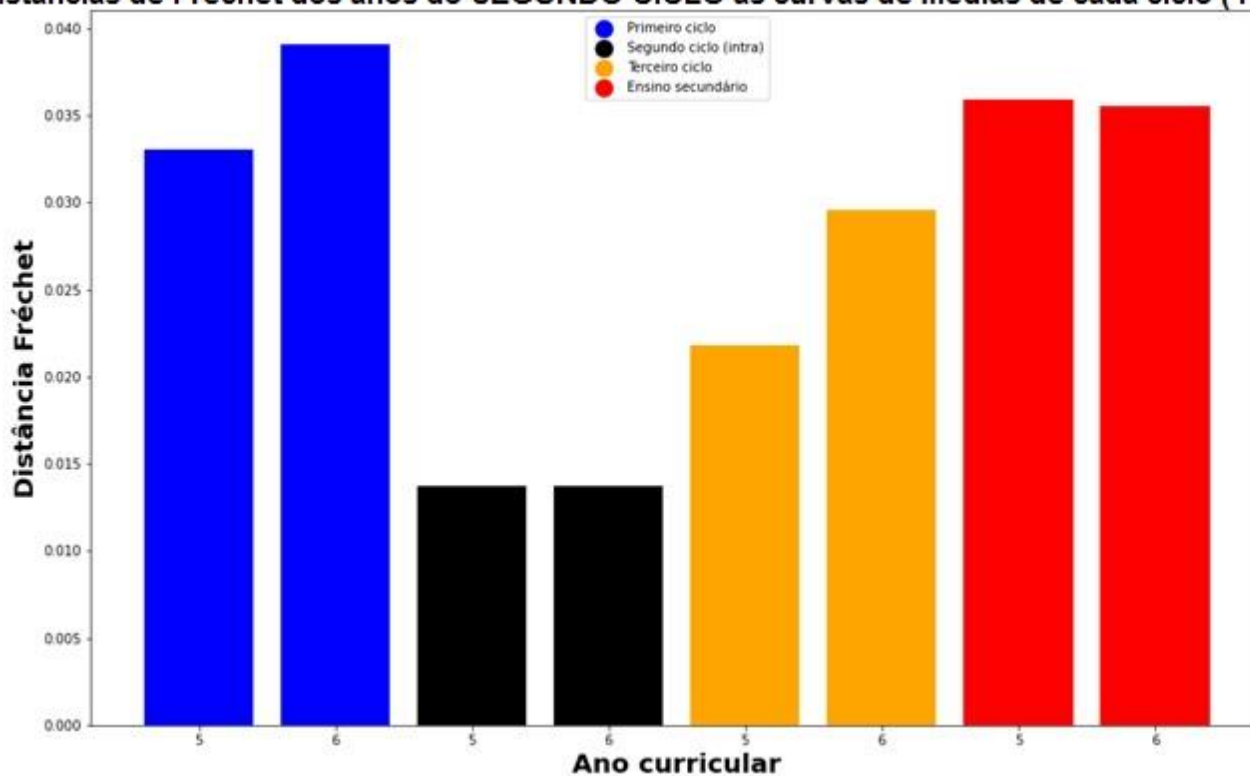
Apêndice E. Resultados das distâncias de Fréchet às curvas médias de cada ciclo

Distâncias de Fréchet às curvas de transição médias de cada ciclo

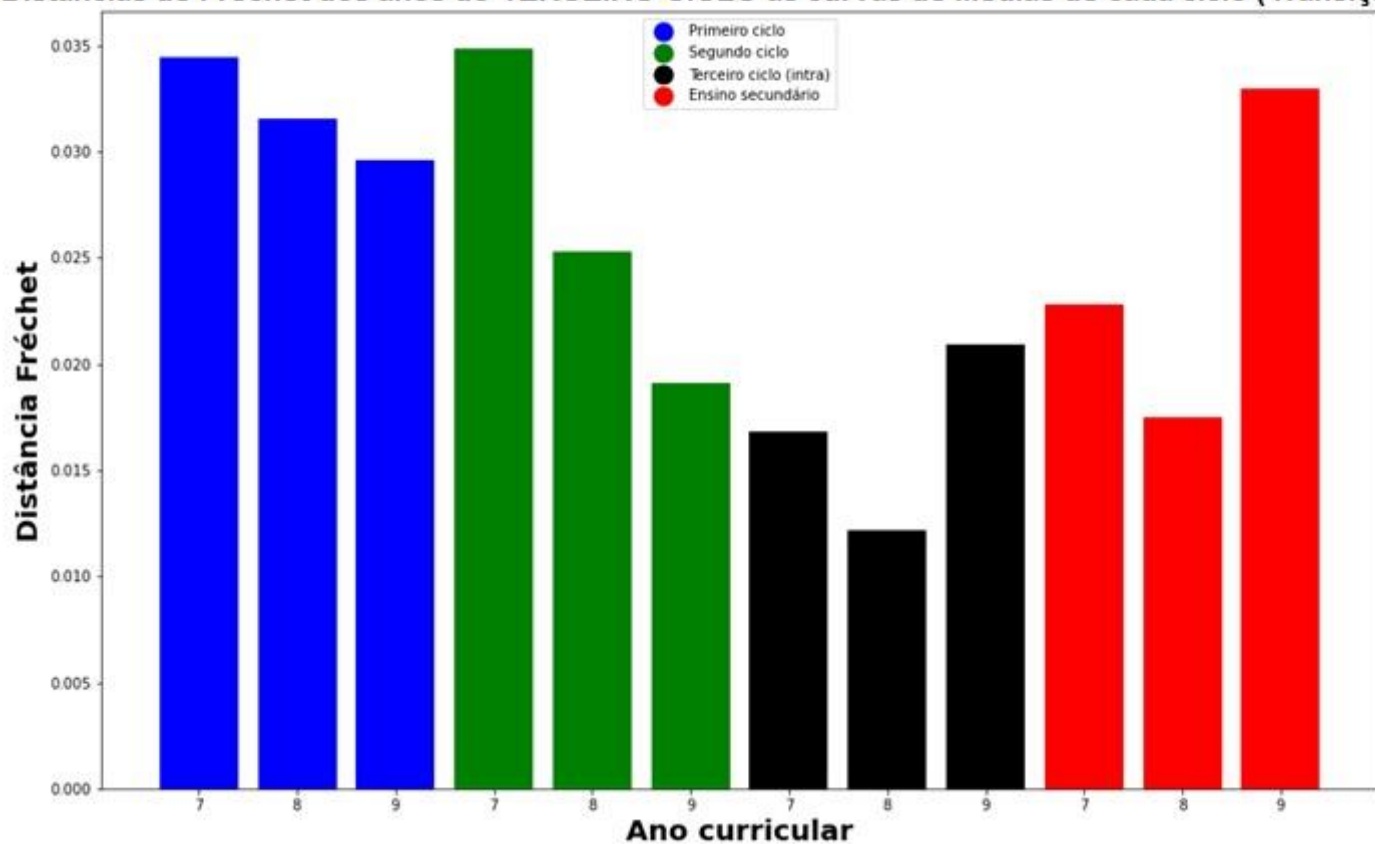
Distâncias de Fréchet dos anos do PRIMEIRO CICLO às curvas de médias de cada ciclo (Transição)



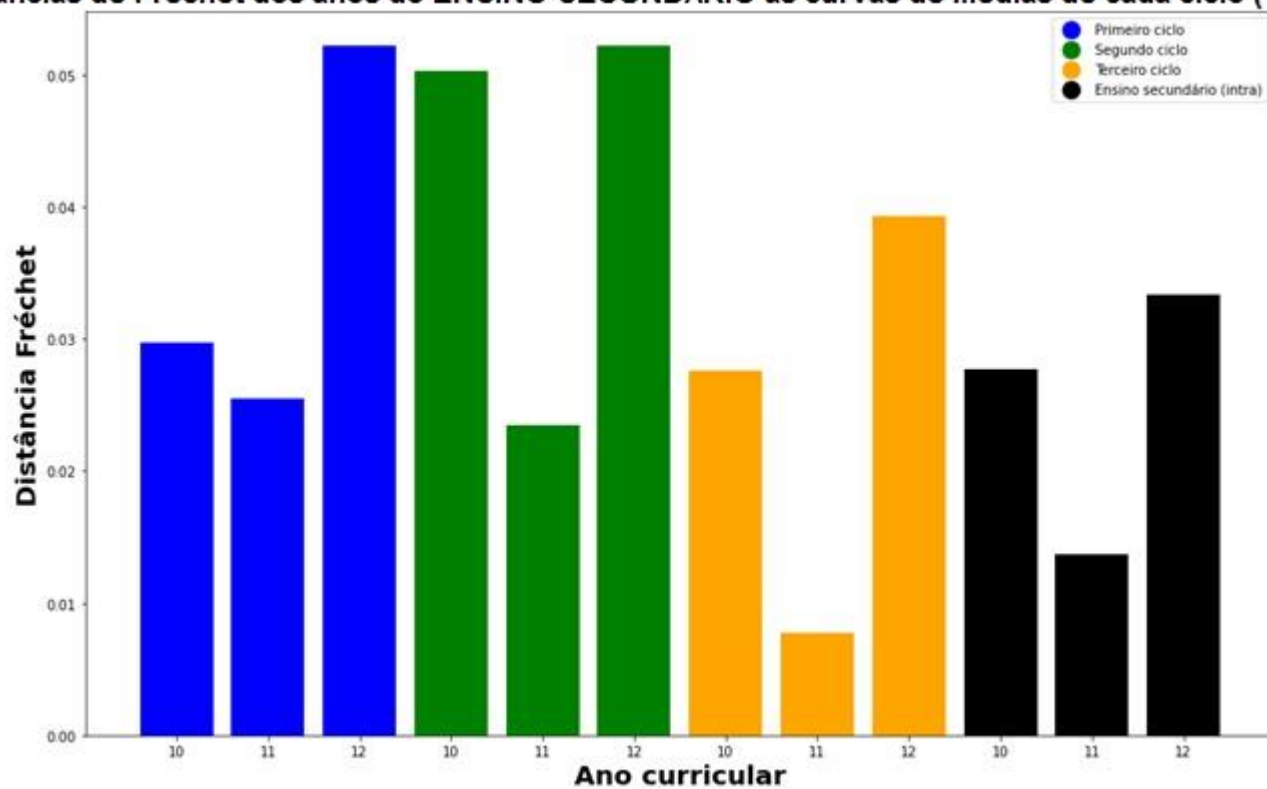
Distâncias de Fréchet dos anos do SEGUNDO CICLO às curvas de médias de cada ciclo (Transição)



Distâncias de Fréchet dos anos do TERCEIRO CICLO às curvas de médias de cada ciclo (Transição)

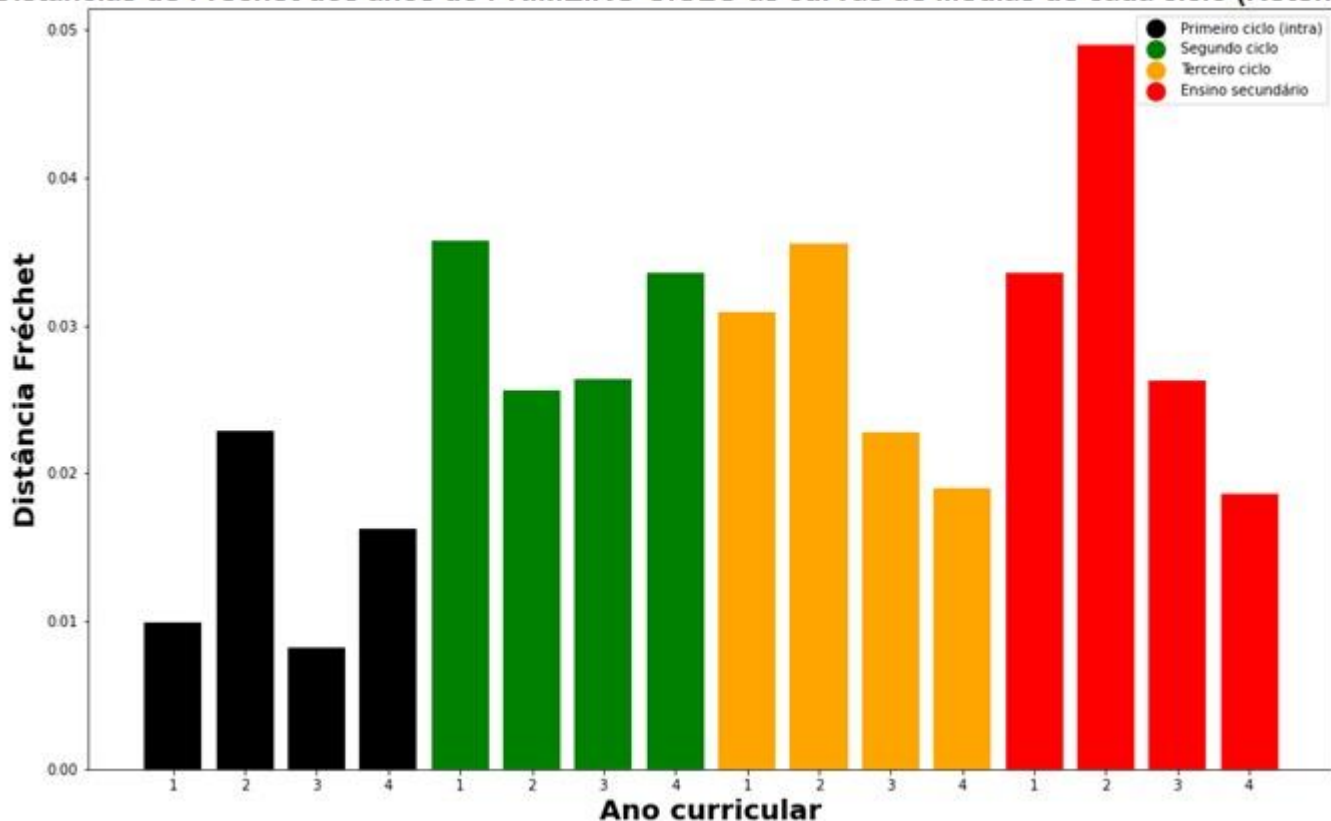


Distâncias de Fréchet dos anos do ENSINO SECUNDÁRIO às curvas de médias de cada ciclo (Transição)

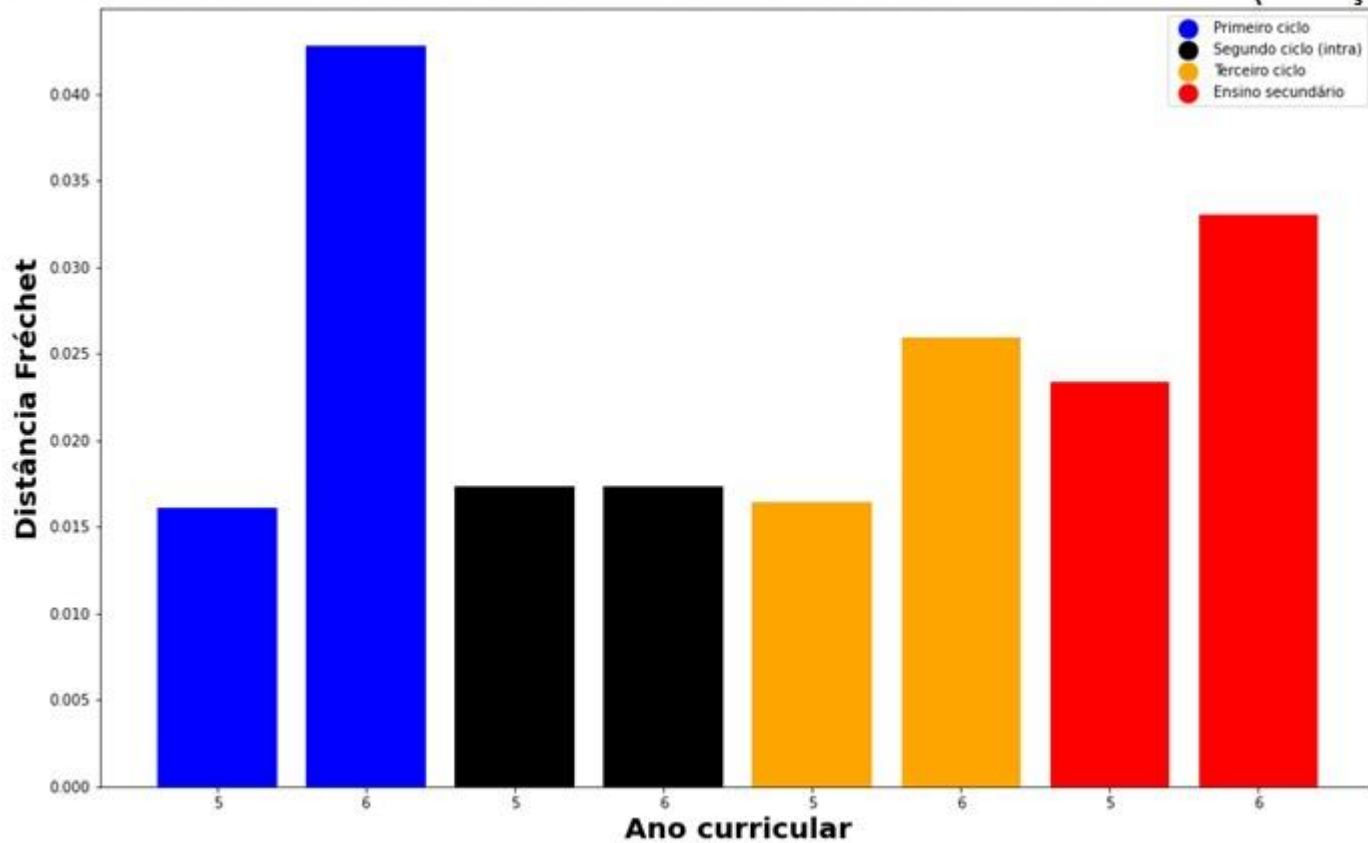


Distâncias de Fréchet às curvas de retenção médias de cada ciclo

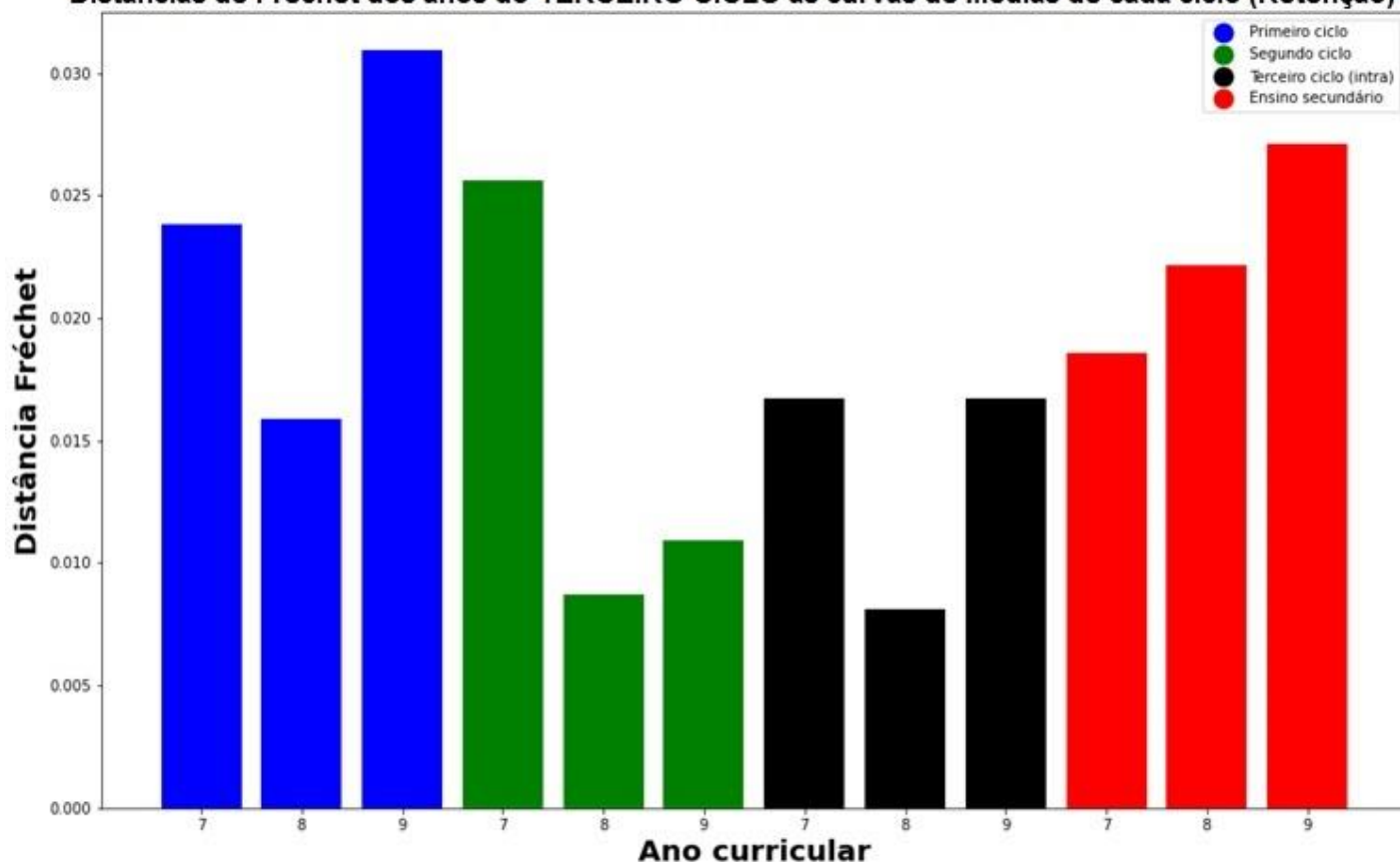
Distâncias de Fréchet dos anos do PRIMEIRO CICLO às curvas de médias de cada ciclo (Retenção)



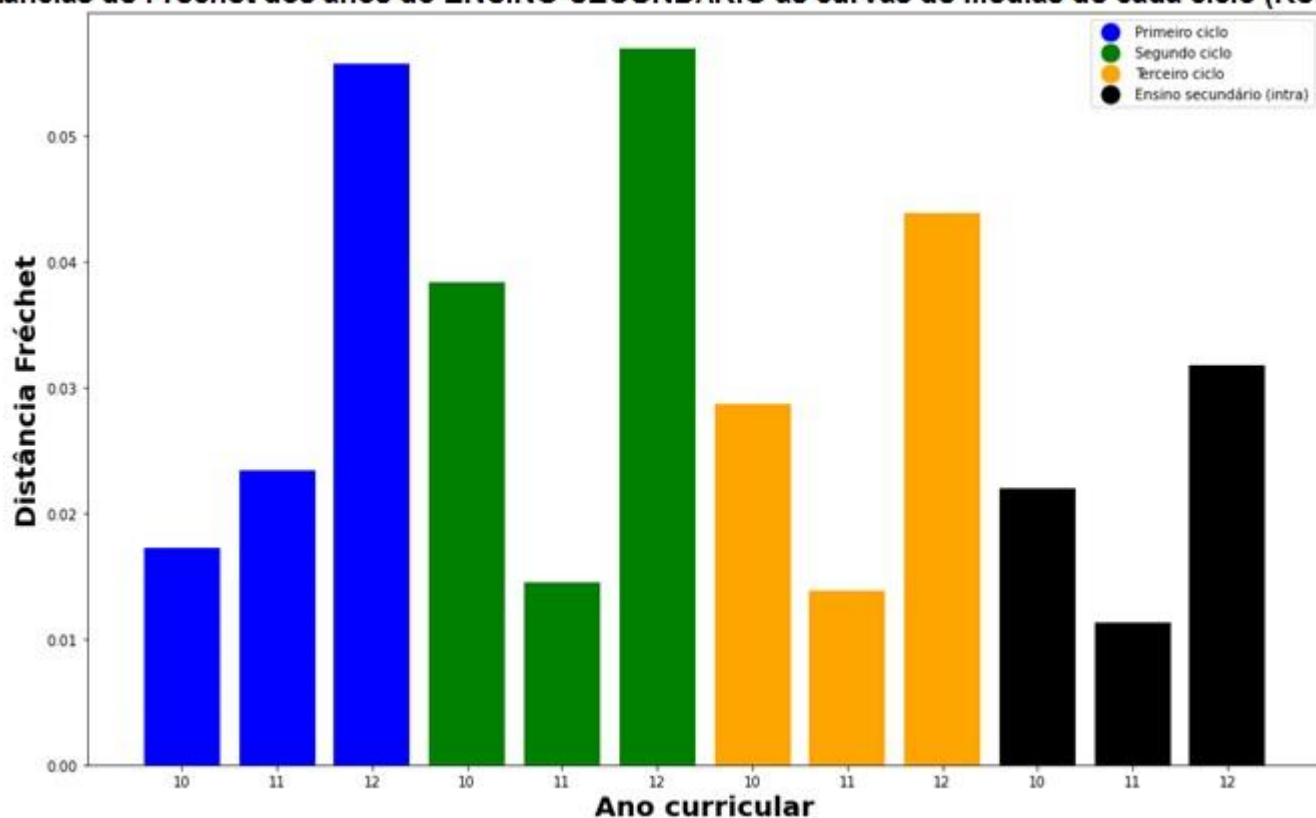
Distâncias de Fréchet dos anos do SEGUNDO CICLO às curvas de médias de cada ciclo (Retenção)



Distâncias de Fréchet dos anos do TERCEIRO CICLO às curvas de médias de cada ciclo (Retenção)

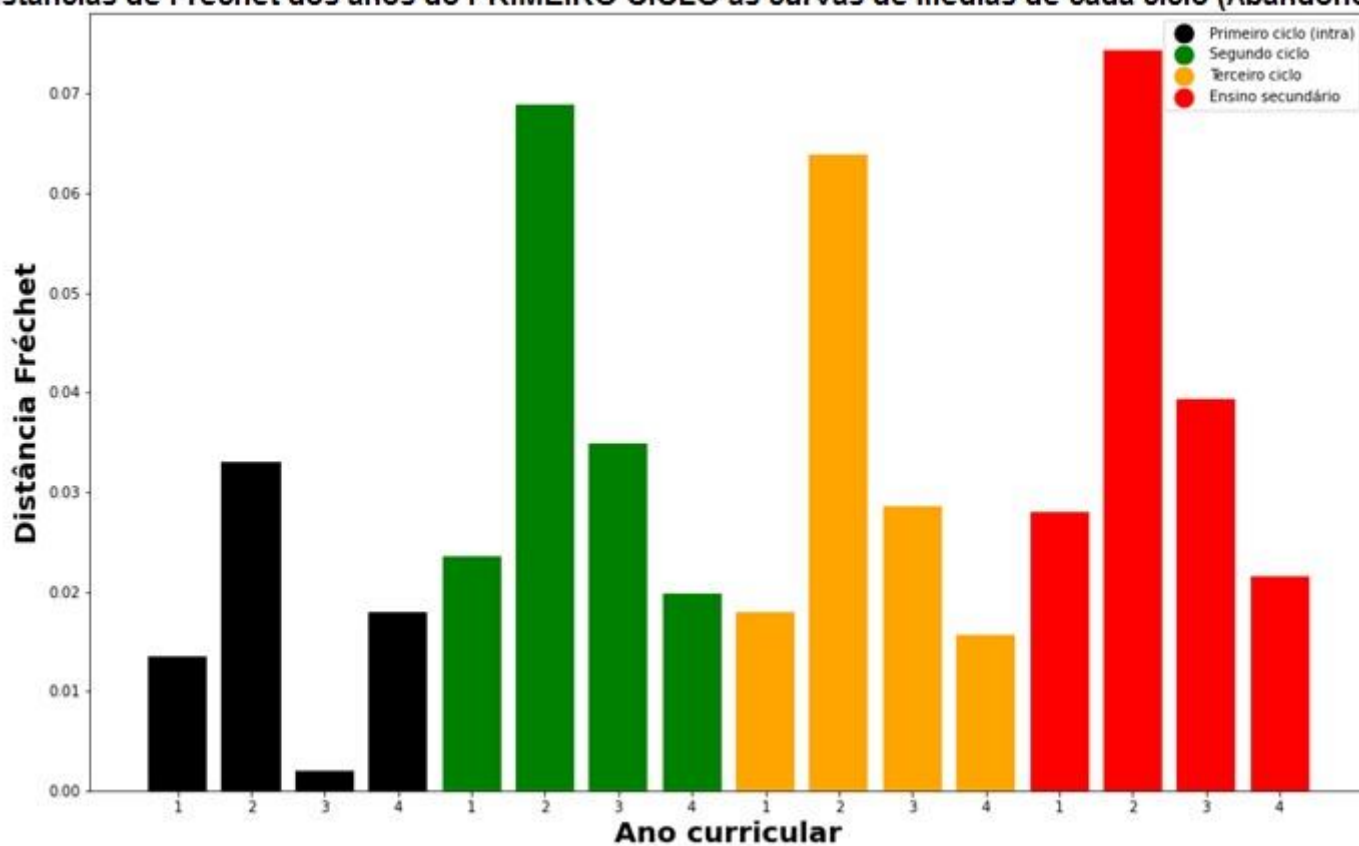


Distâncias de Fréchet dos anos do ENSINO SECUNDÁRIO às curvas de médias de cada ciclo (Retenção)

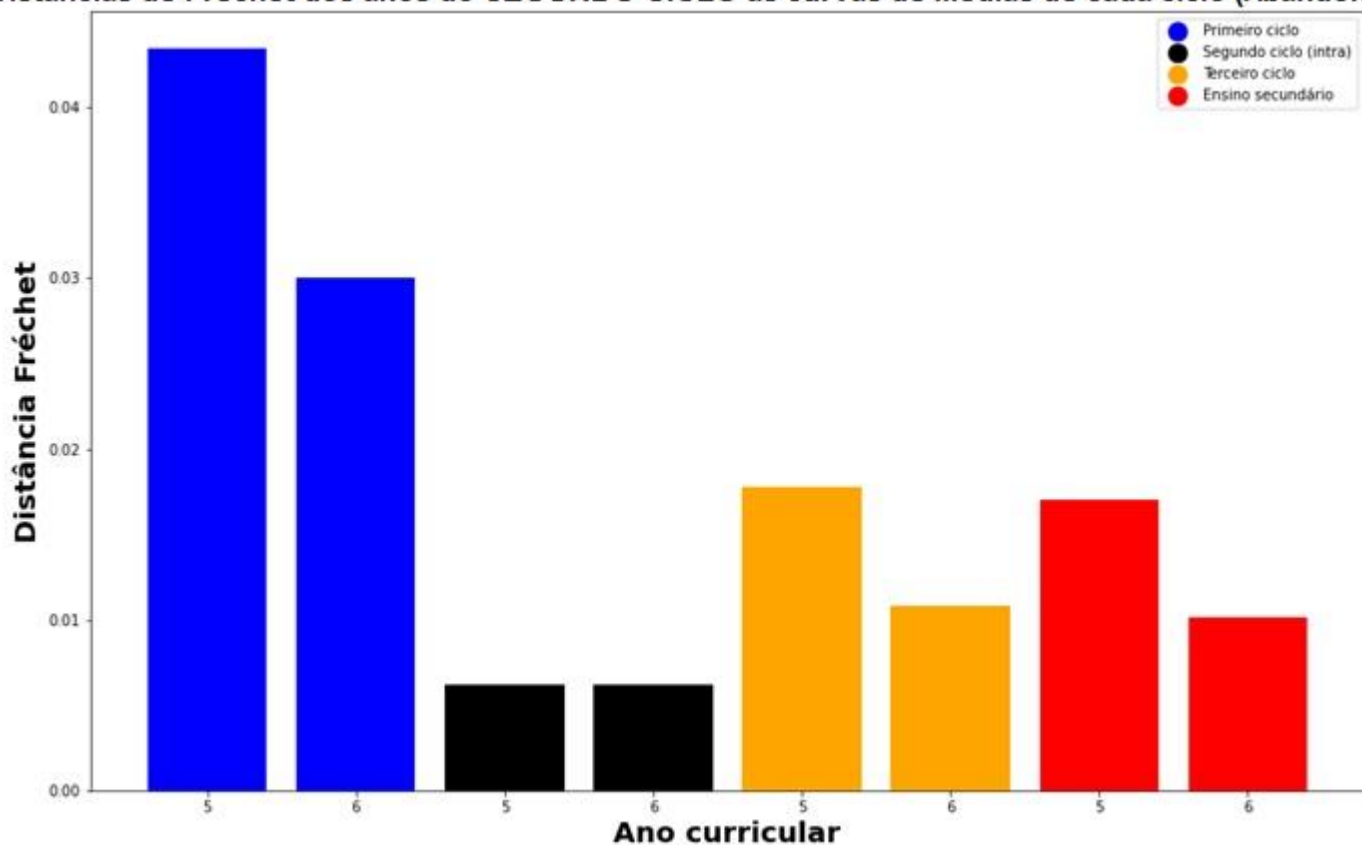


Distâncias de Fréchet às curvas de abandono médias de cada ciclo

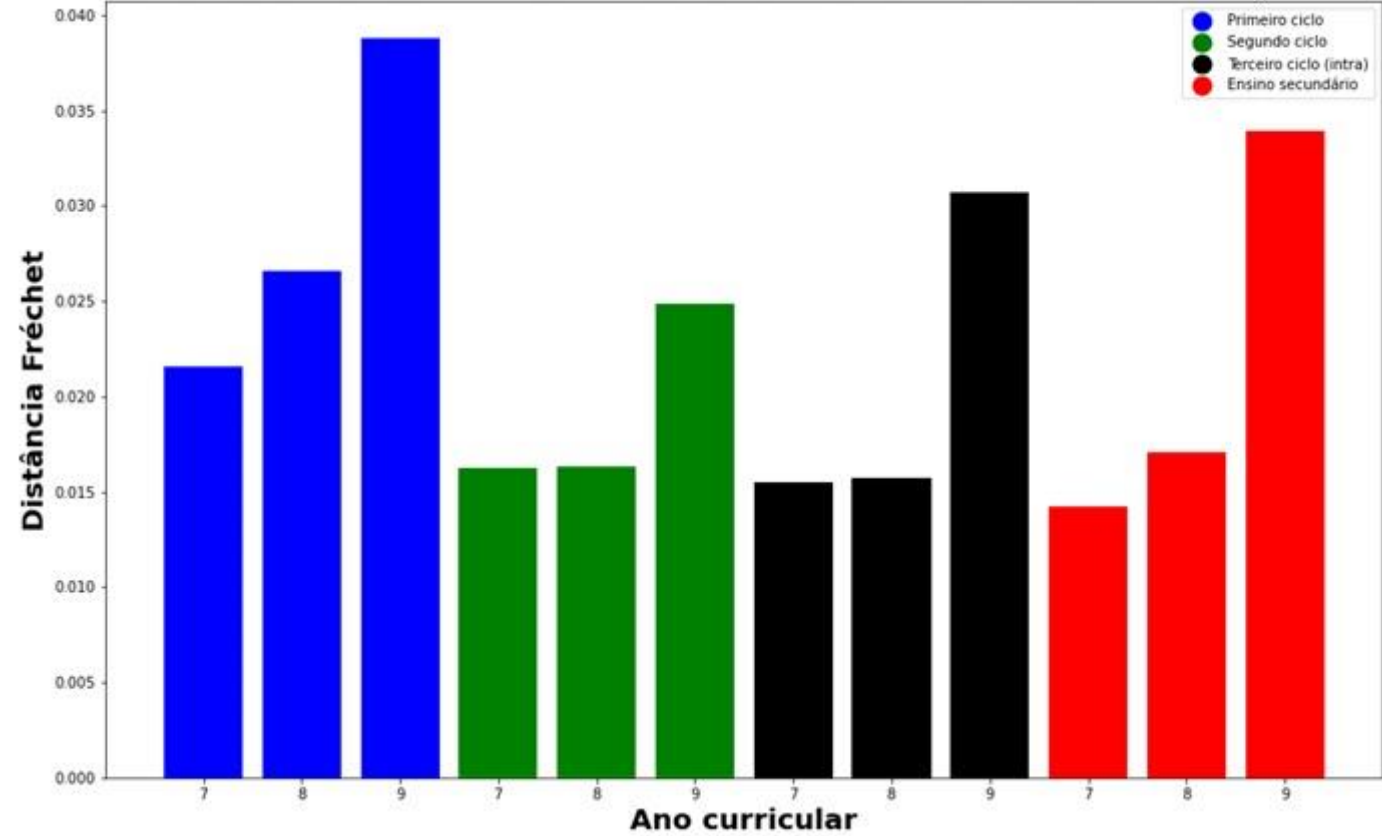
Distâncias de Fréchet dos anos do PRIMEIRO CICLO às curvas de médias de cada ciclo (Abandono)



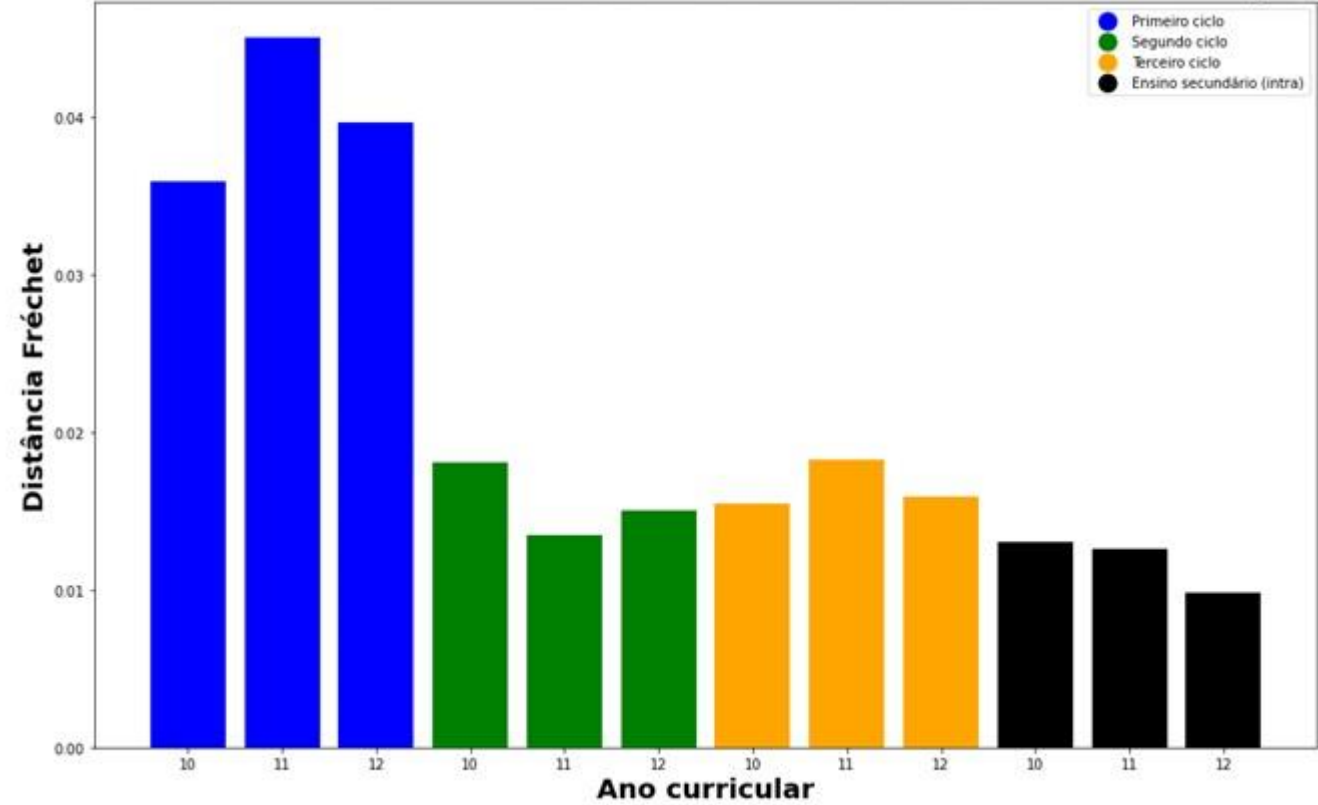
Distâncias de Fréchet dos anos do SEGUNDO CICLO às curvas de médias de cada ciclo (Abandono)



Distâncias de Fréchet dos anos do TERCEIRO CICLO às curvas de médias de cada ciclo (Abandono)

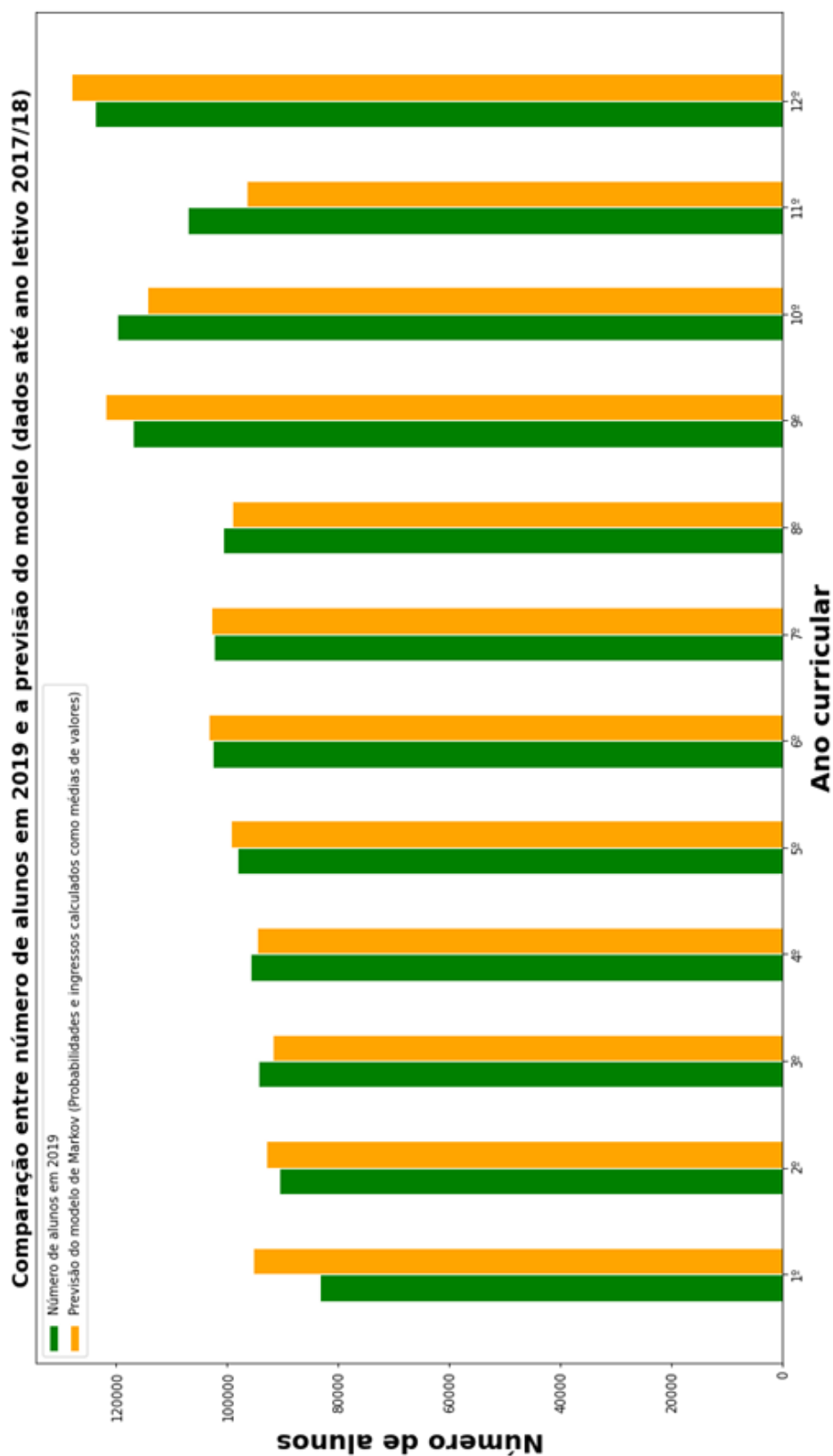


Distâncias de Fréchet dos anos do ENSINO SECUNDÁRIO às curvas de médias de cada ciclo (Abandono)

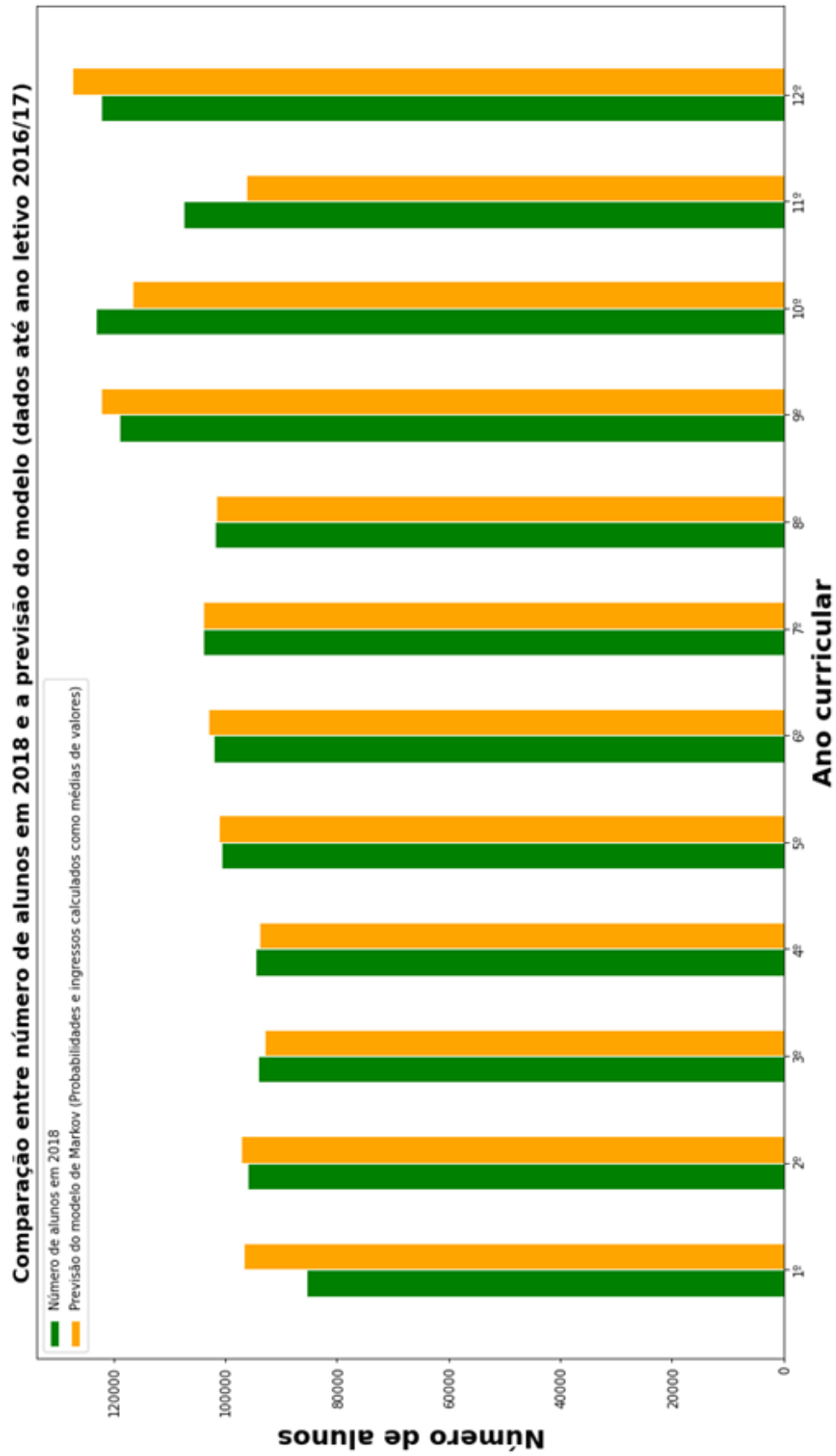


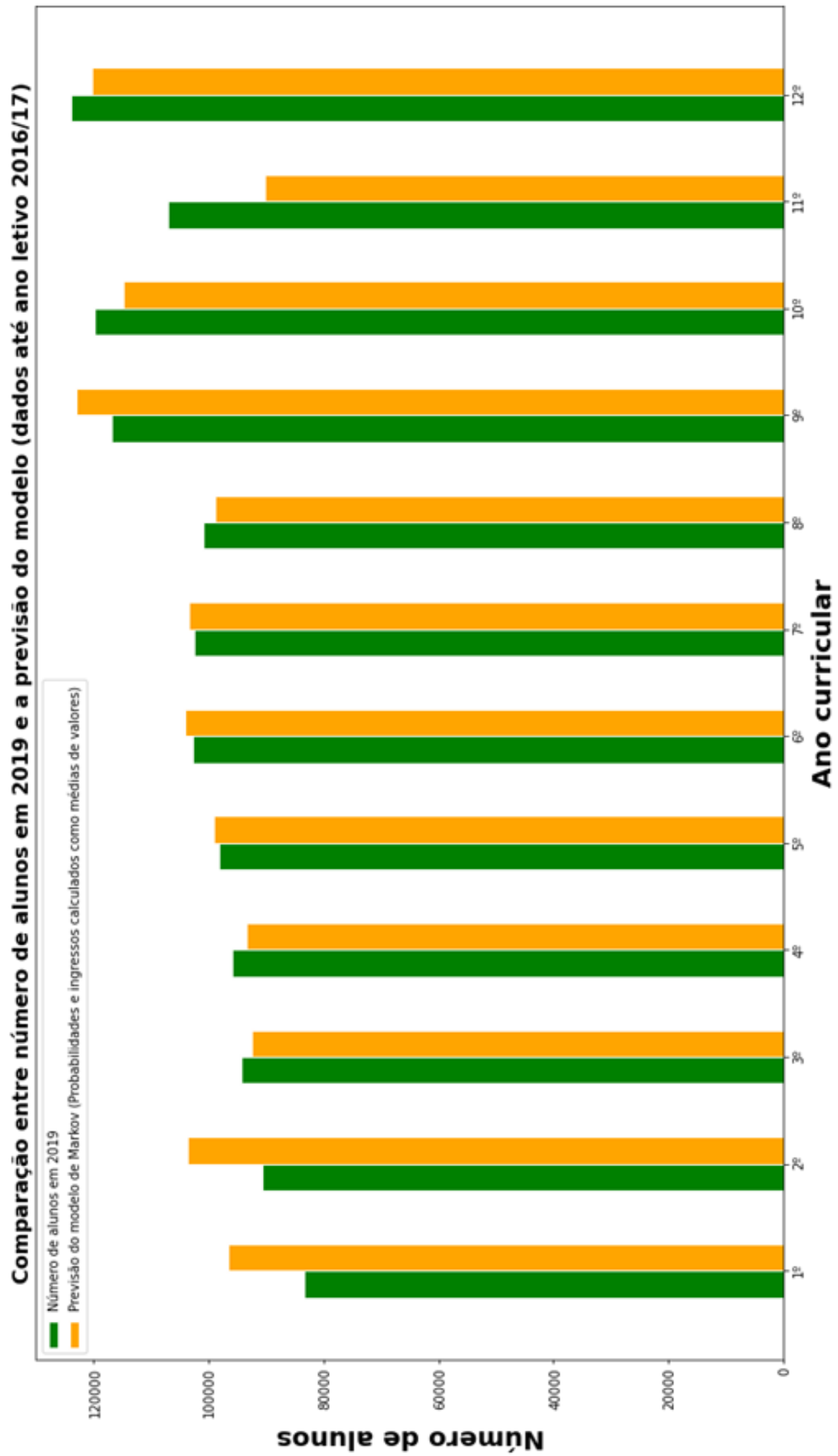
Apêndice F. Resultados dos modelos de probabilidades estáticas (todo ensino português)

Modelo com dados até o ano letivo 2017/18

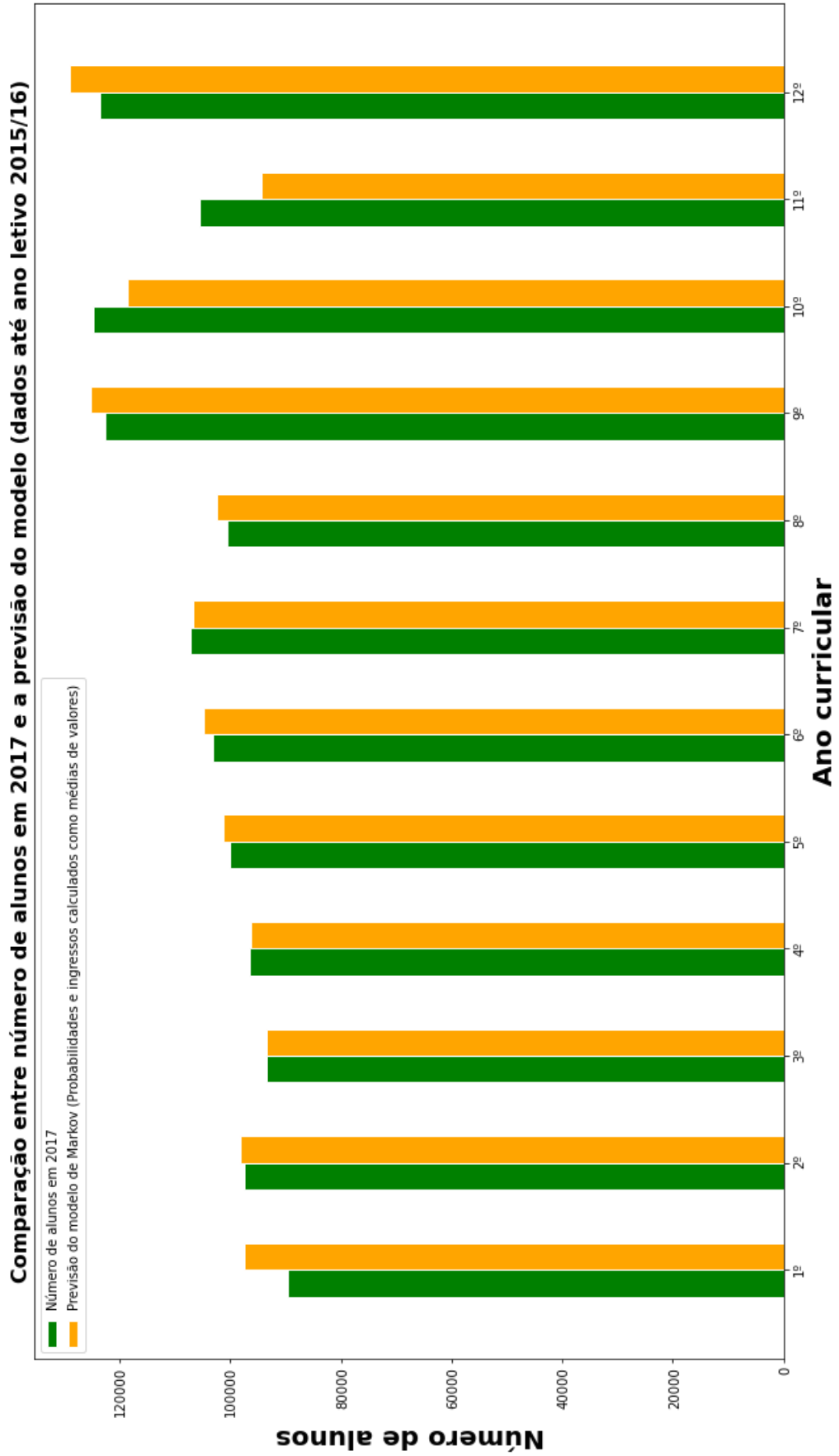


Modelos com dados até o ano letivo 2016/17

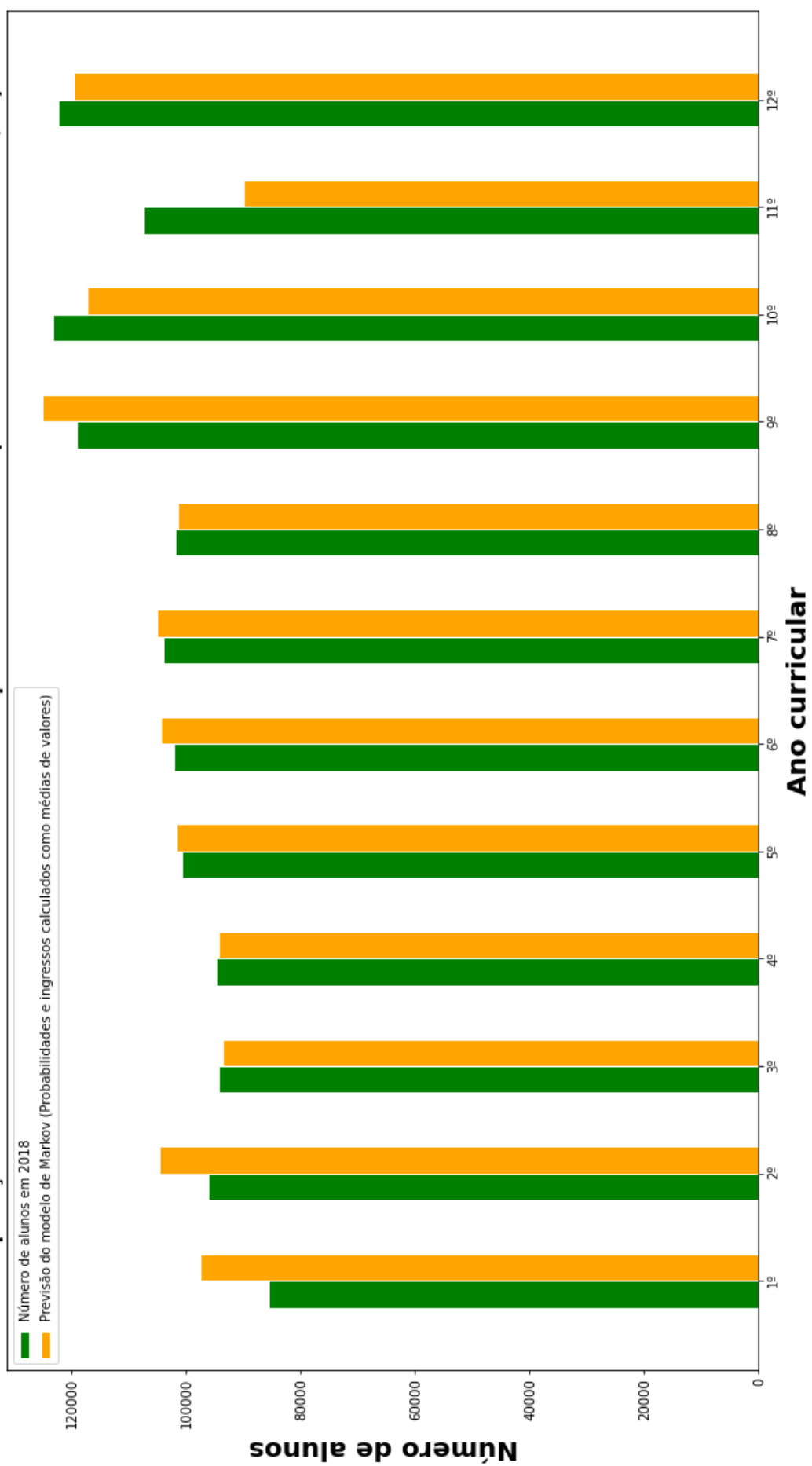


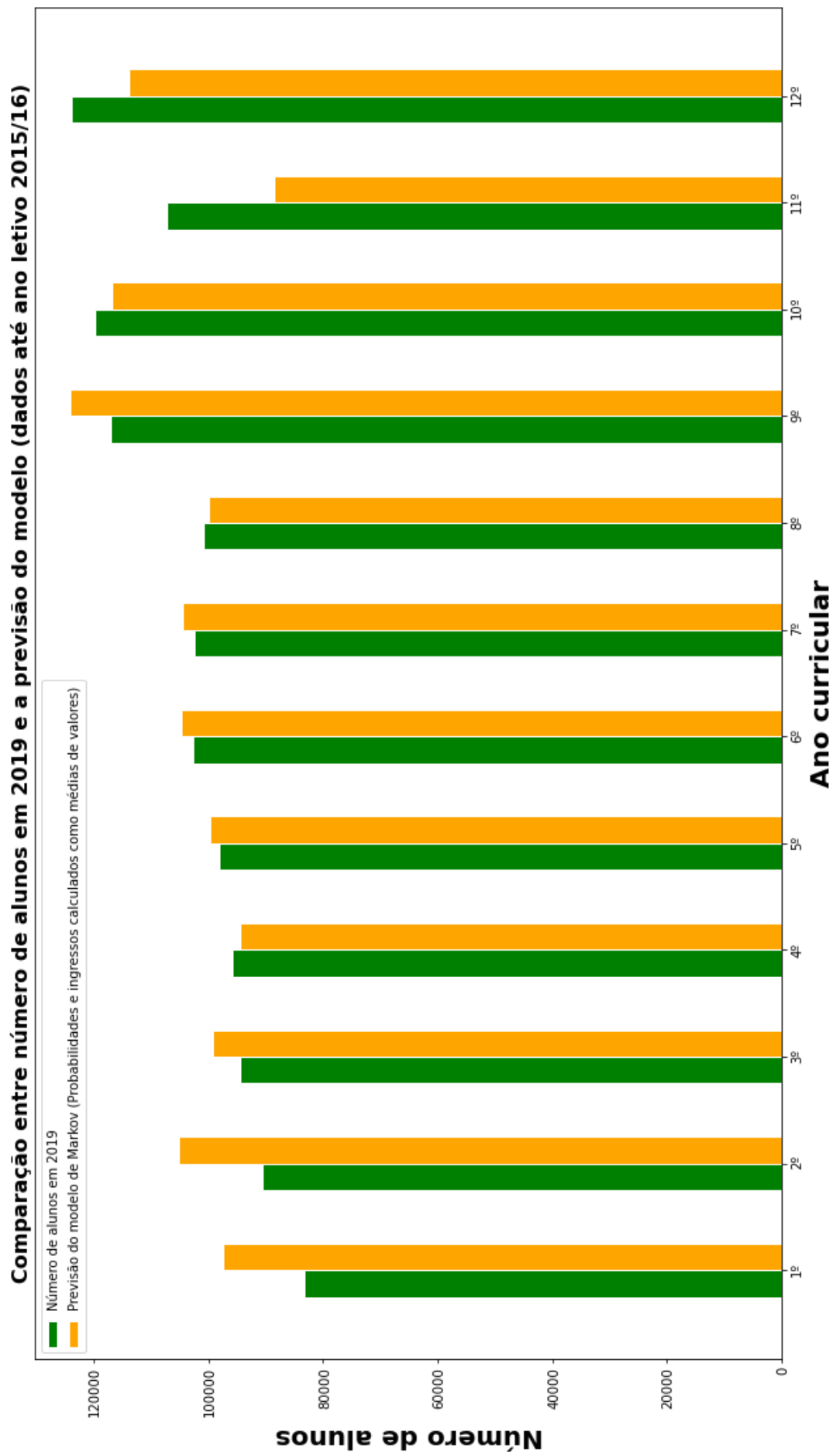


Modelos com dados até o ano letivo 2015/16



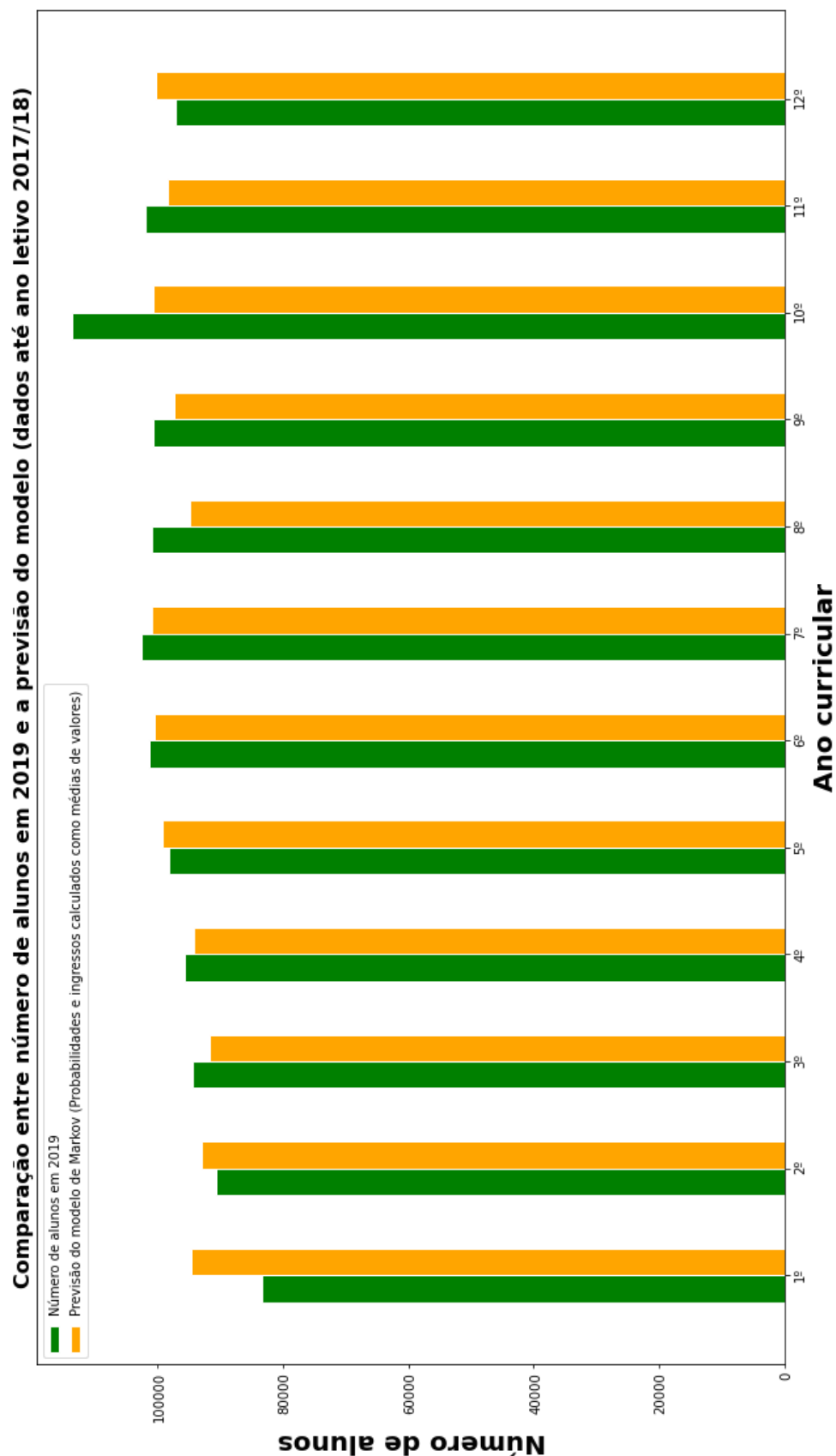
Comparação entre número de alunos em 2018 e a previsão do modelo (dados até ano letivo 2015/16)



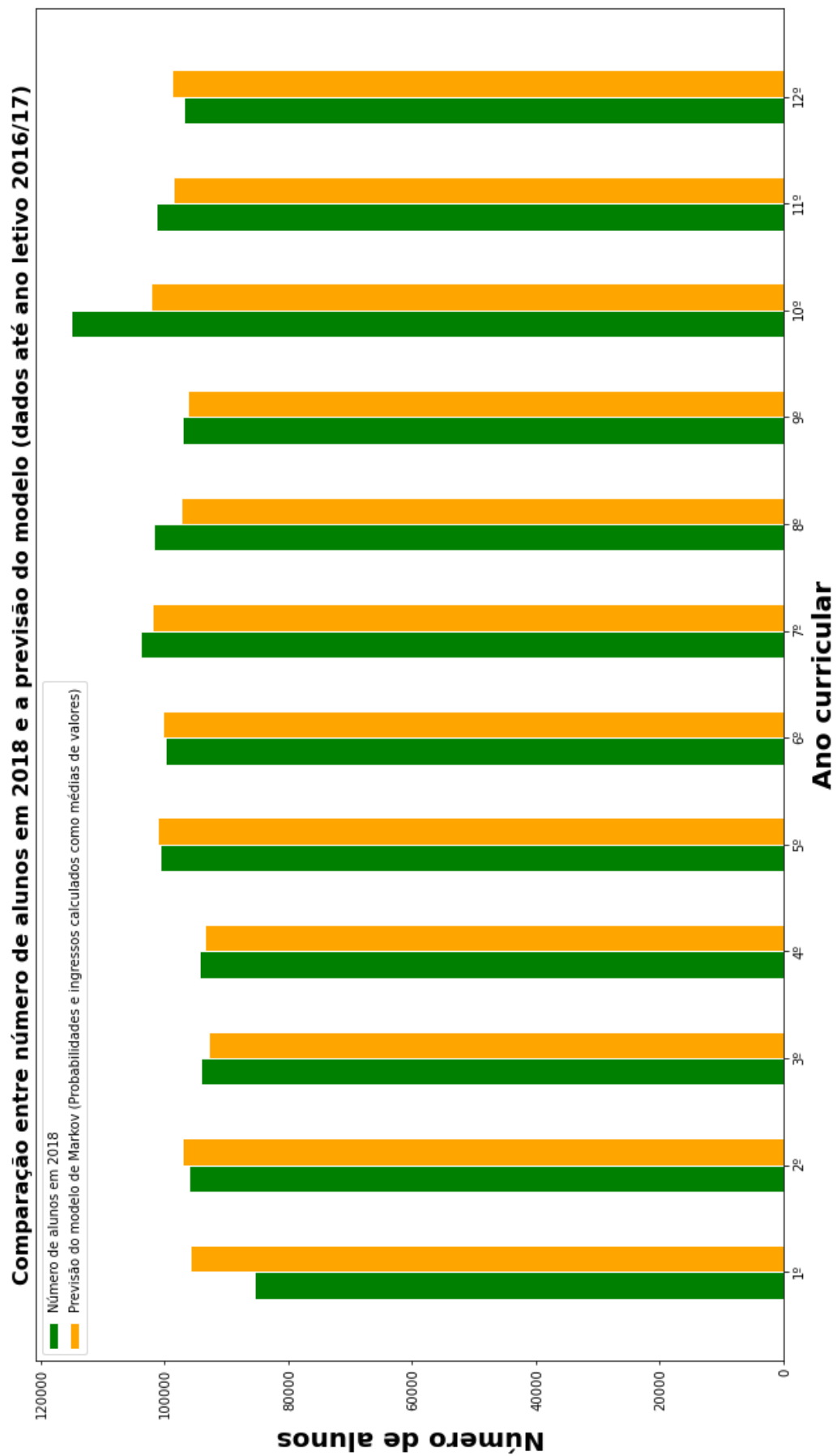


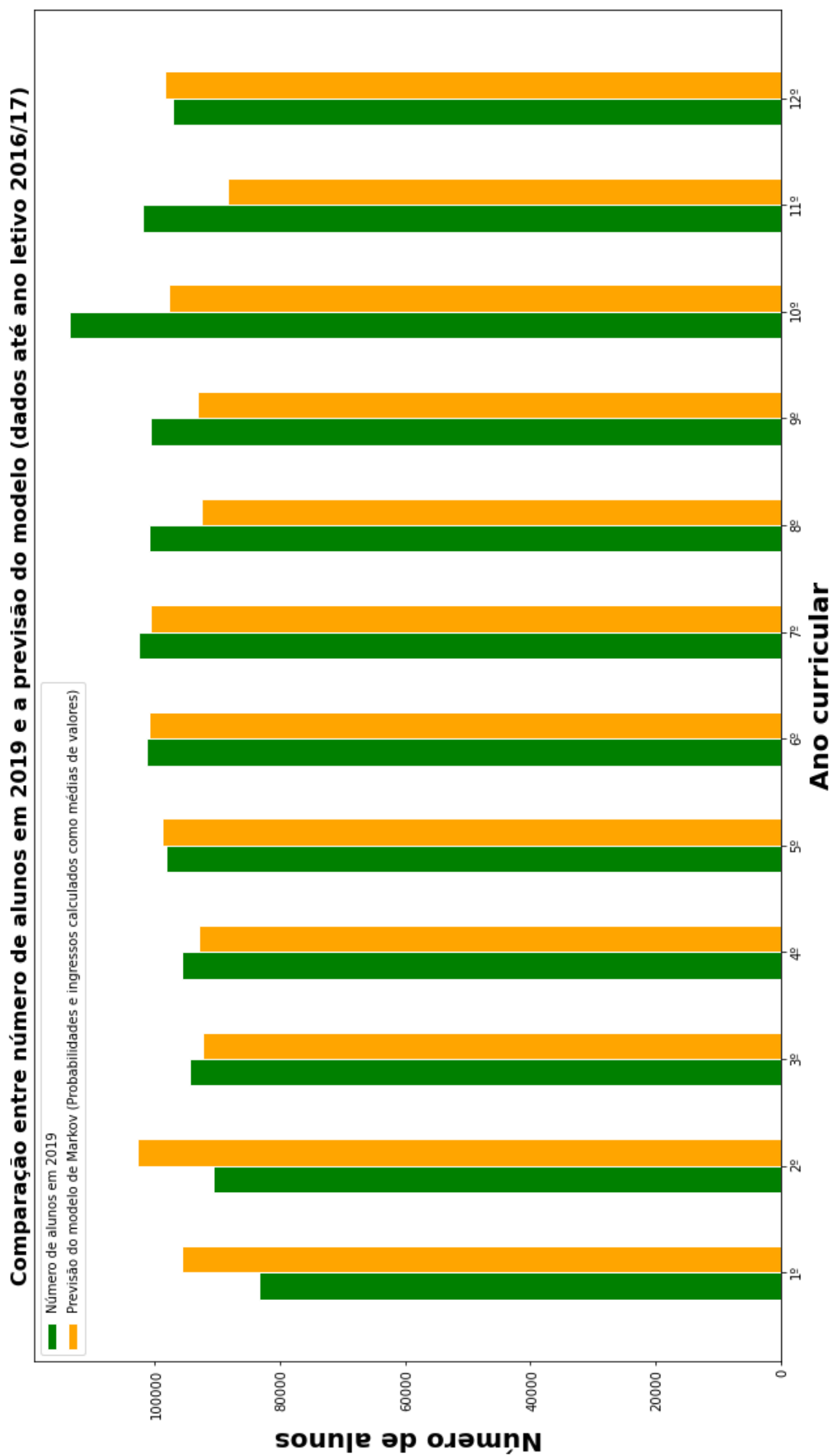
Apêndice G. Resultados dos modelos de probabilidades estáticas (ensino comum com cursos profissionais)

Modelo com dados até o ano letivo 2017/18

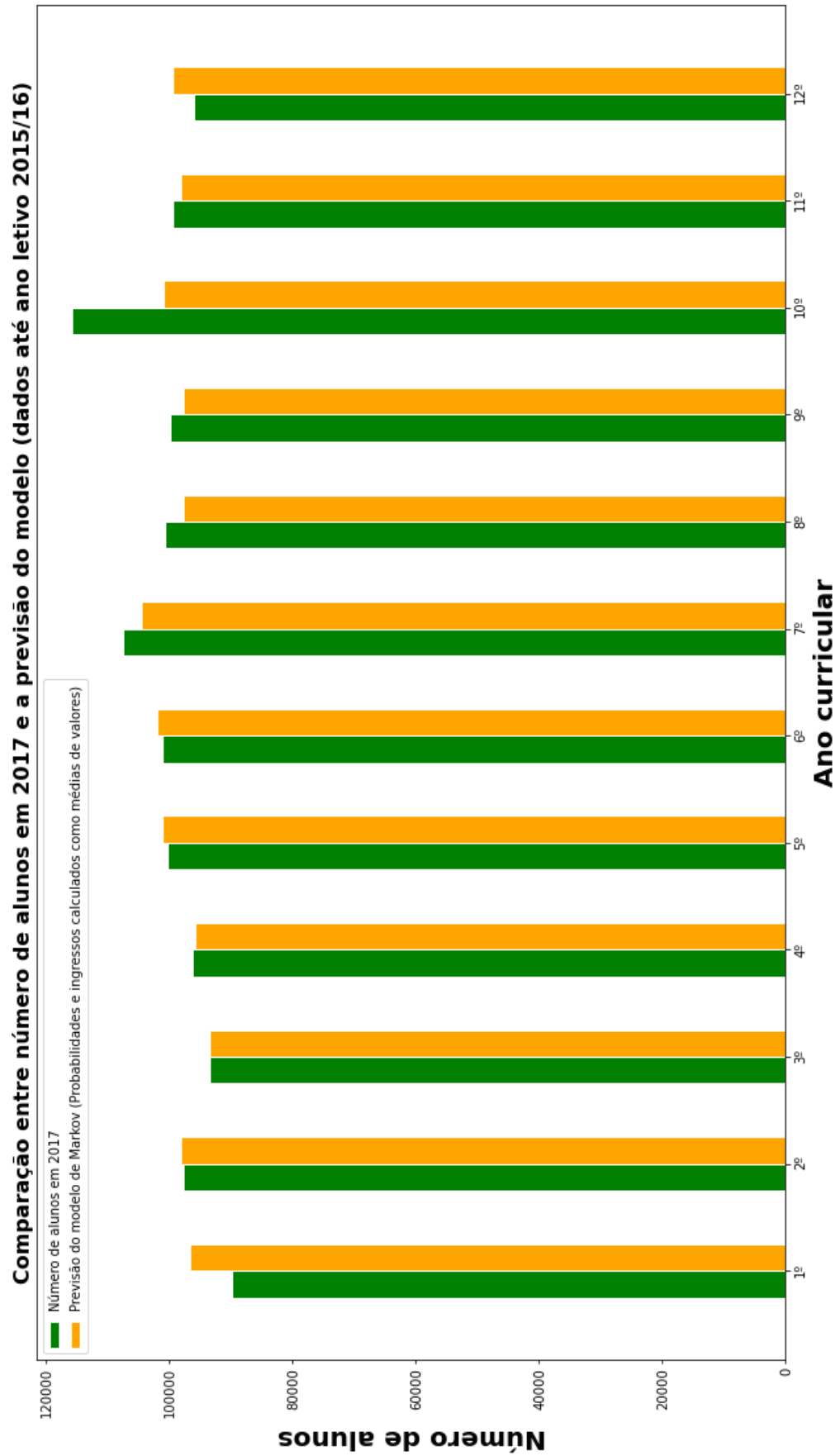


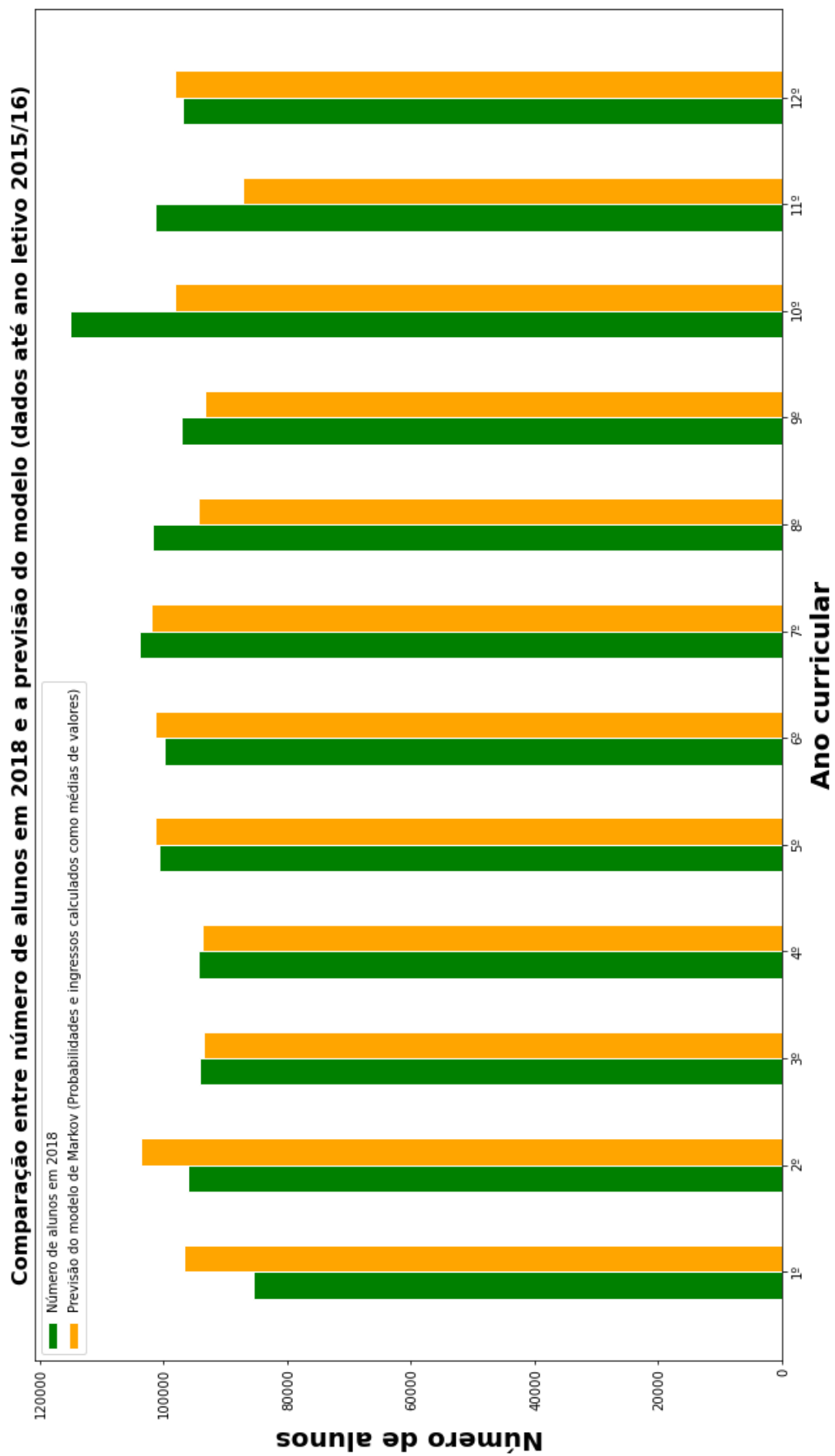
Modelos com dados até o ano letivo 2016/17

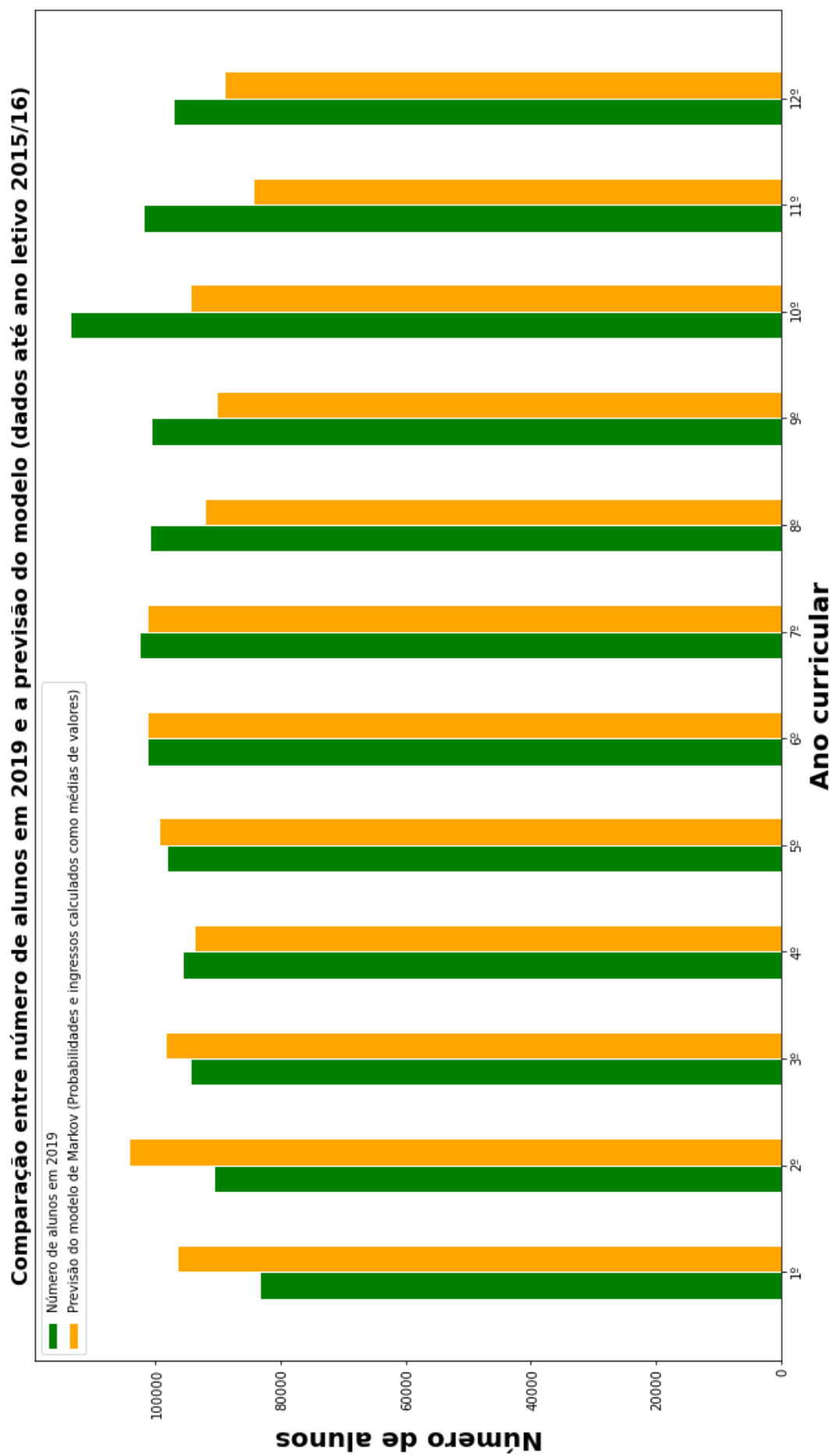




Modelos com dados até o ano letivo 2015/16

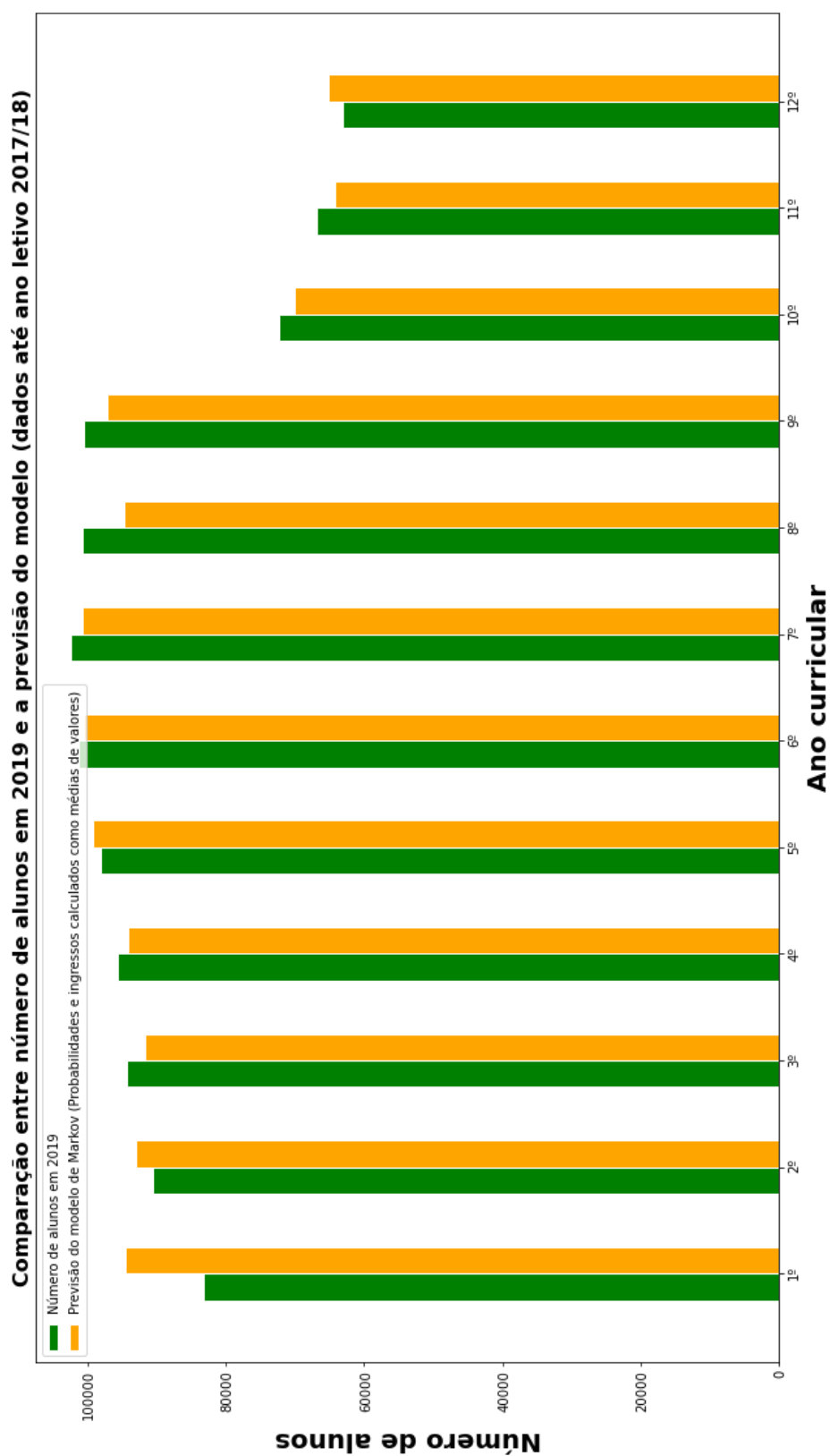




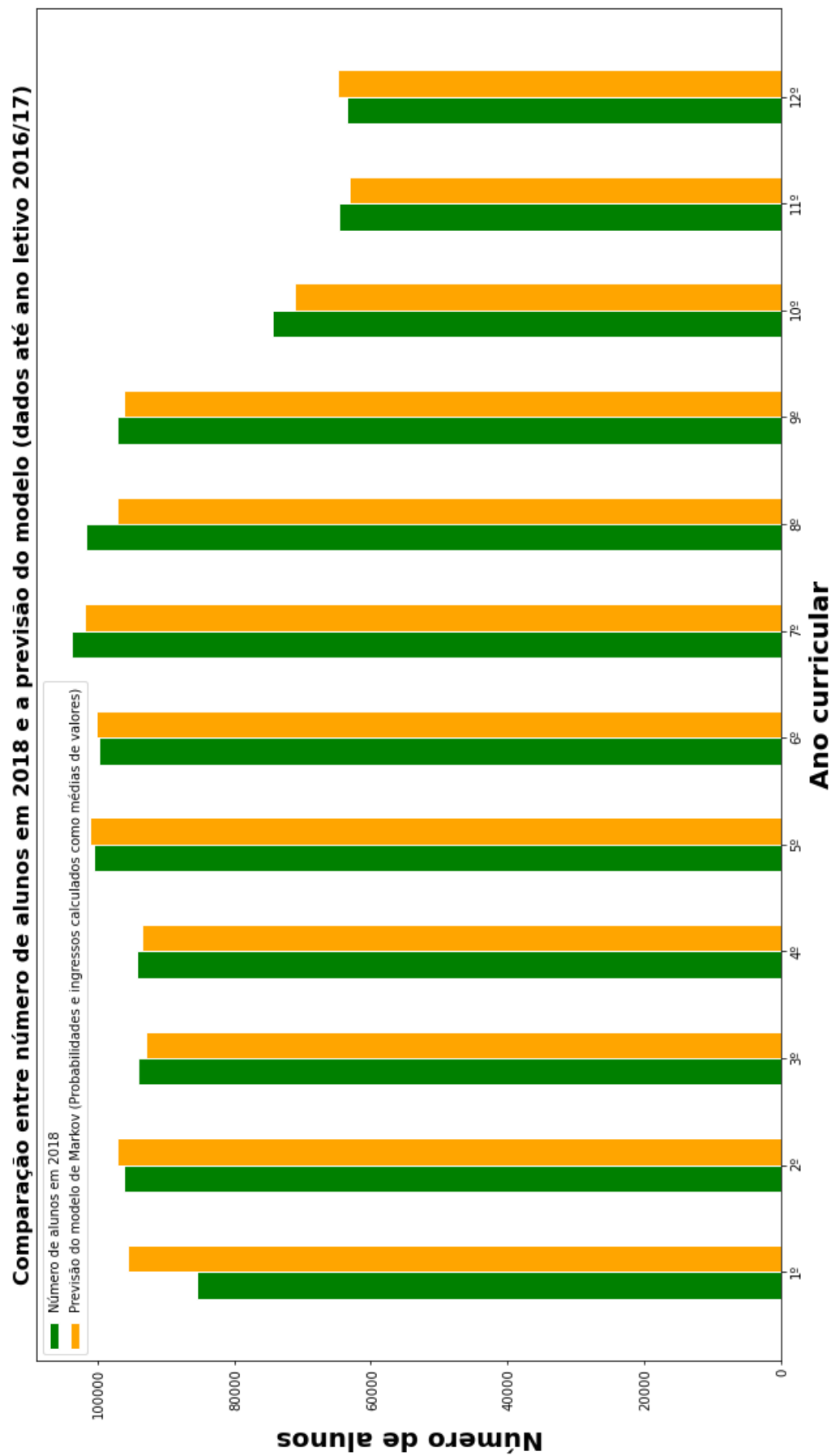


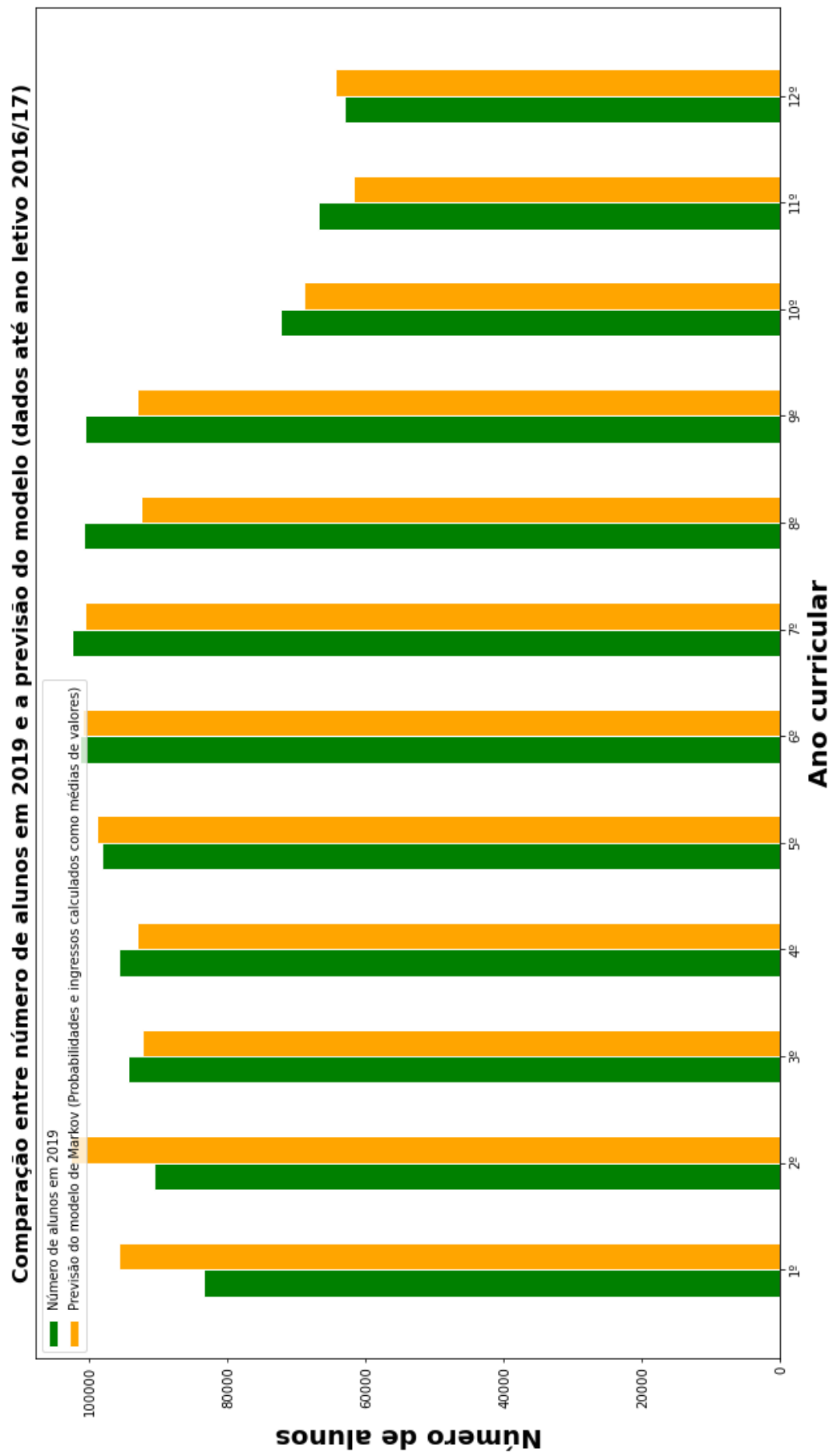
Apêndice H. Resultados dos modelos de probabilidades estáticas (ensino comum sem cursos profissionais)

Modelo com dados até o ano letivo 2017/18

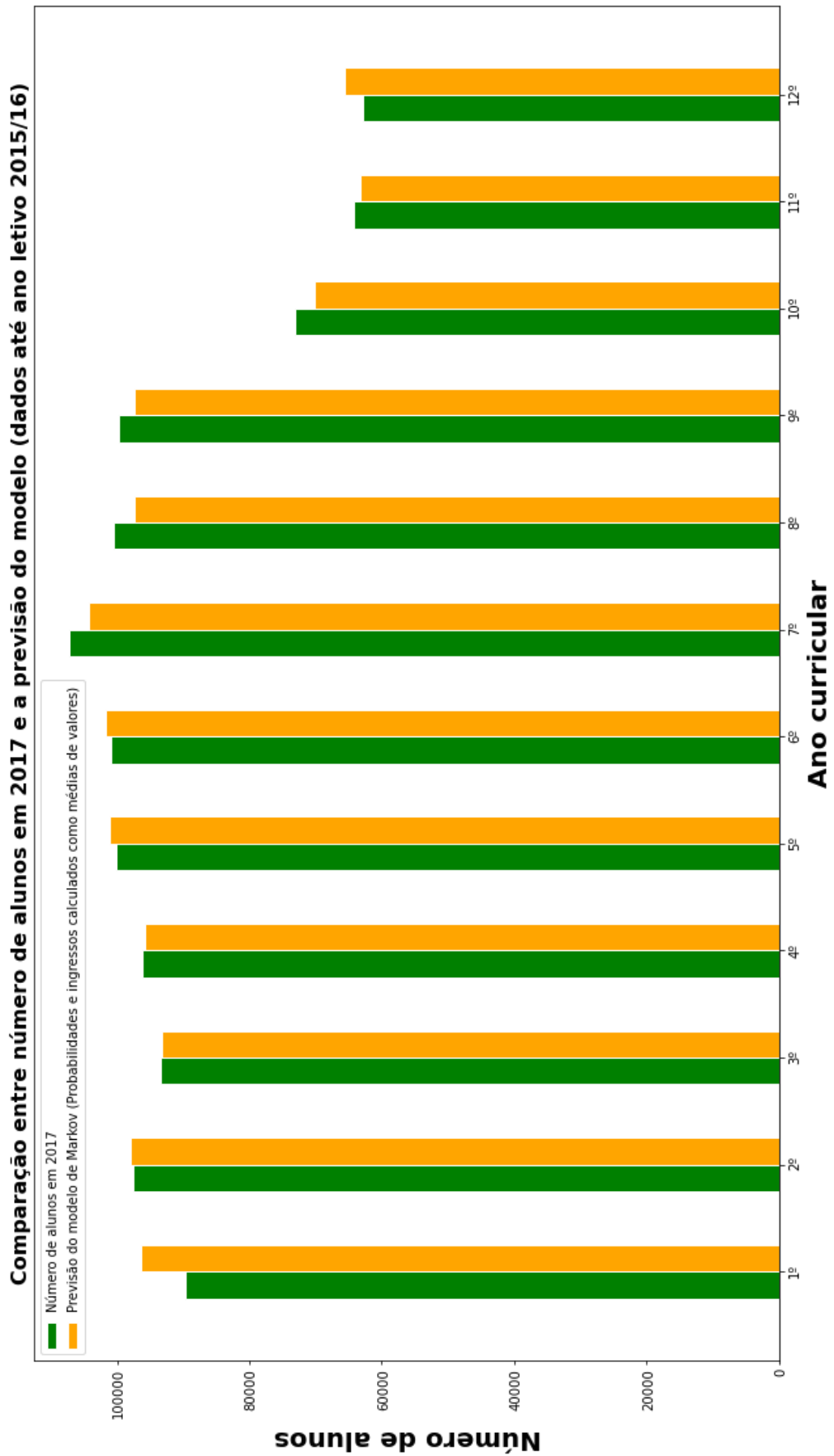


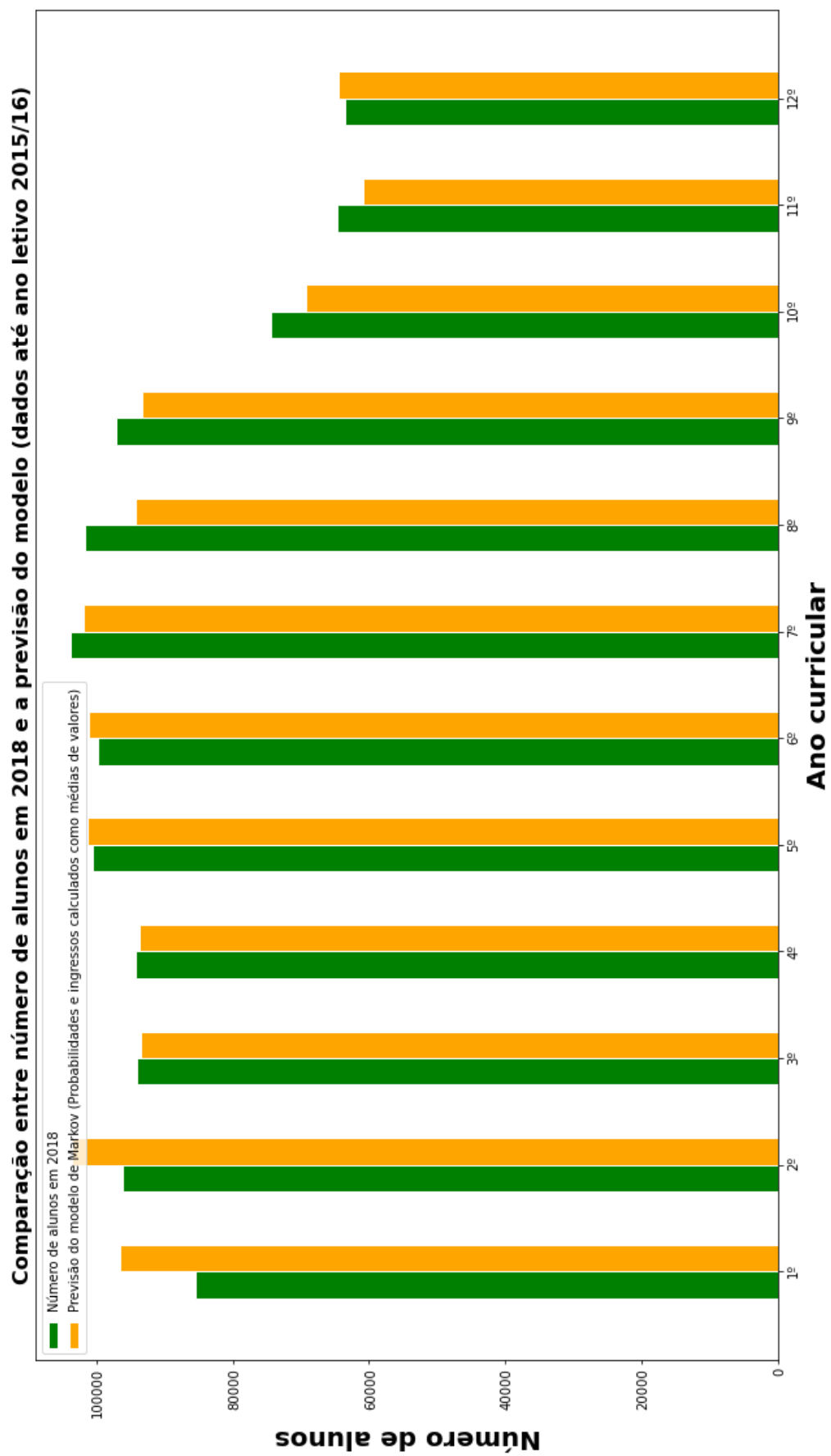
Modelos com dados até o ano letivo 2016/17

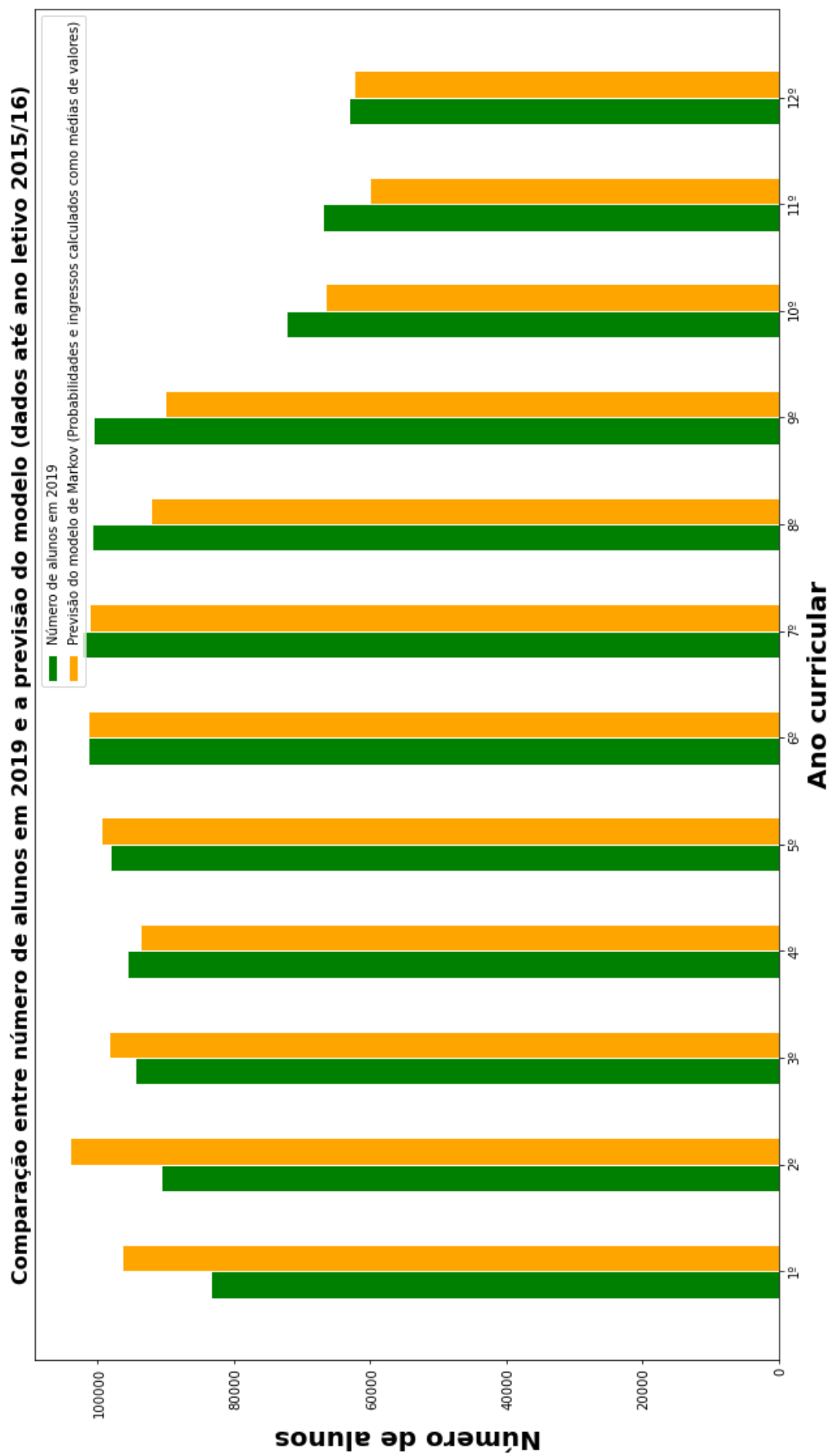




Modelos com dados até o ano letivo 2015/16







Apêndice I. MAPEs dos modelos com probabilidades estáticas

Dados de todo ensino português															
Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18 2016/17	2019/20	14,57	2,75	2,87	1,25	1,16	0,64	0,27	1,73	4,28	4,45	9,95	3,34	3,94	4,23
	2018/19	13,07	1,02	1,34	0,62	0,61	1,07	0,04	0,08	2,77	5,18	10,36	4,13	3,36	4,25
	2019/20	16,01	14,37	2,08	2,51	0,89	1,32	0,89	1,91	5,30	4,04	15,87	2,85	5,67	6,02
2015/16	2017/18	8,79	0,50	0,01	0,22	1,11	1,57	0,52	1,75	2,03	4,90	10,52	4,34	3,02	3,47
	2018/19	14,14	8,89	0,59	0,36	0,93	2,41	1,19	0,44	5,08	4,74	16,32	2,37	4,79	5,50
	2019/20	17,04	16,12	5,21	1,57	1,56	1,94	2,02	0,96	6,09	2,38	17,31	8,18	6,70	6,48

Ensino comum (com cursos profissionais)															
Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18 2016/17	2019/20	13,57	2,68	2,92	1,52	0,99	0,82	1,59	5,95	3,28	11,33	3,51	3,21	4,28	4,08
	2018/19	12,02	0,96	1,39	0,74	0,47	0,38	1,92	4,55	0,90	11,15	2,66	1,92	3,26	4,06
	2019/20	14,93	13,37	2,19	2,84	0,61	0,40	1,84	8,22	7,47	14,00	13,42	1,42	6,73	5,86
2015/16	2017/18	7,71	0,45	0,07	0,35	0,96	0,82	2,77	3,04	2,23	12,92	1,40	3,53	3,02	3,75
	2018/19	13,00	7,88	0,70	0,54	0,68	1,46	1,78	7,38	3,92	14,75	14,02	1,29	5,62	5,60
	2019/20	15,87	14,97	4,25	1,95	1,24	0,02	1,17	8,58	10,29	16,91	17,31	8,28	8,40	6,65

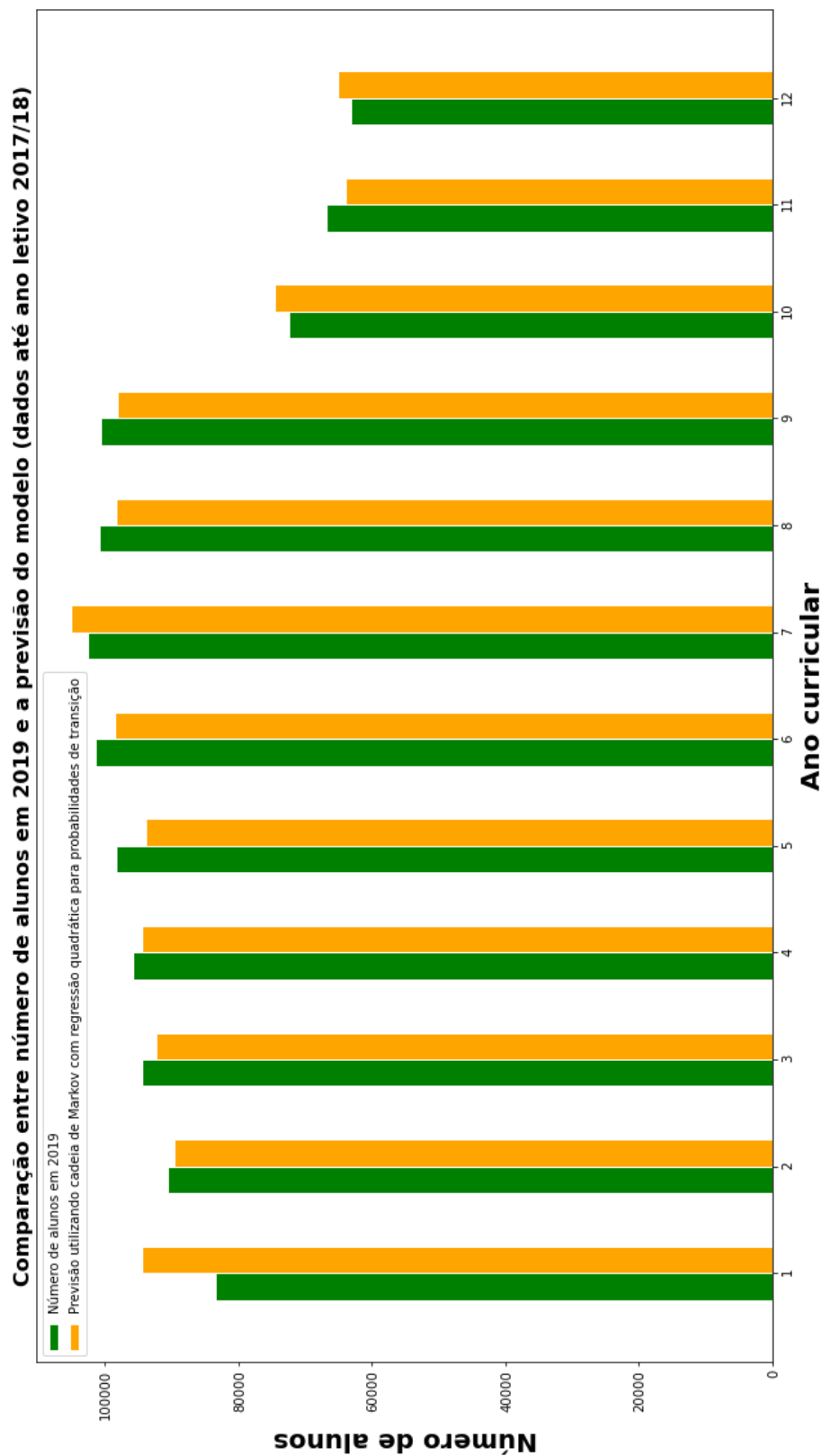
Ensino comum (sem cursos profissionais)															
Último ano utilizado no modelo	Ano de previsão	Ano curricular												MAPE	Desvio-padrão
		1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º		
2017/18 2016/17	2019/20	13,47	2,67	2,93	1,53	0,98	0,82	1,57	5,97	3,31	3,07	4,06	3,37	3,65	3,41
	2018/19	11,92	0,95	1,40	0,75	0,45	0,38	1,88	4,57	0,94	4,25	2,44	1,97	2,66	3,22
	2019/20	14,82	13,27	2,20	2,85	0,58	0,41	1,81	8,21	7,52	4,66	7,89	1,92	5,51	4,86
2015/16	2017/18	7,60	0,44	0,07	0,36	0,97	0,82	2,75	3,06	2,25	4,09	1,50	4,47	2,36	2,20
	2018/19	12,88	7,77	0,72	0,56	0,66	1,45	1,74	7,38	3,96	6,90	5,94	1,44	4,28	3,90
	2019/20	15,75	14,84	4,15	1,97	1,21	0,00	1,14	8,57	10,32	7,82	10,26	1,13	6,43	5,59

**Apêndice J. Predicted R^2 para os acontecimentos possíveis
(Ensino comum)**

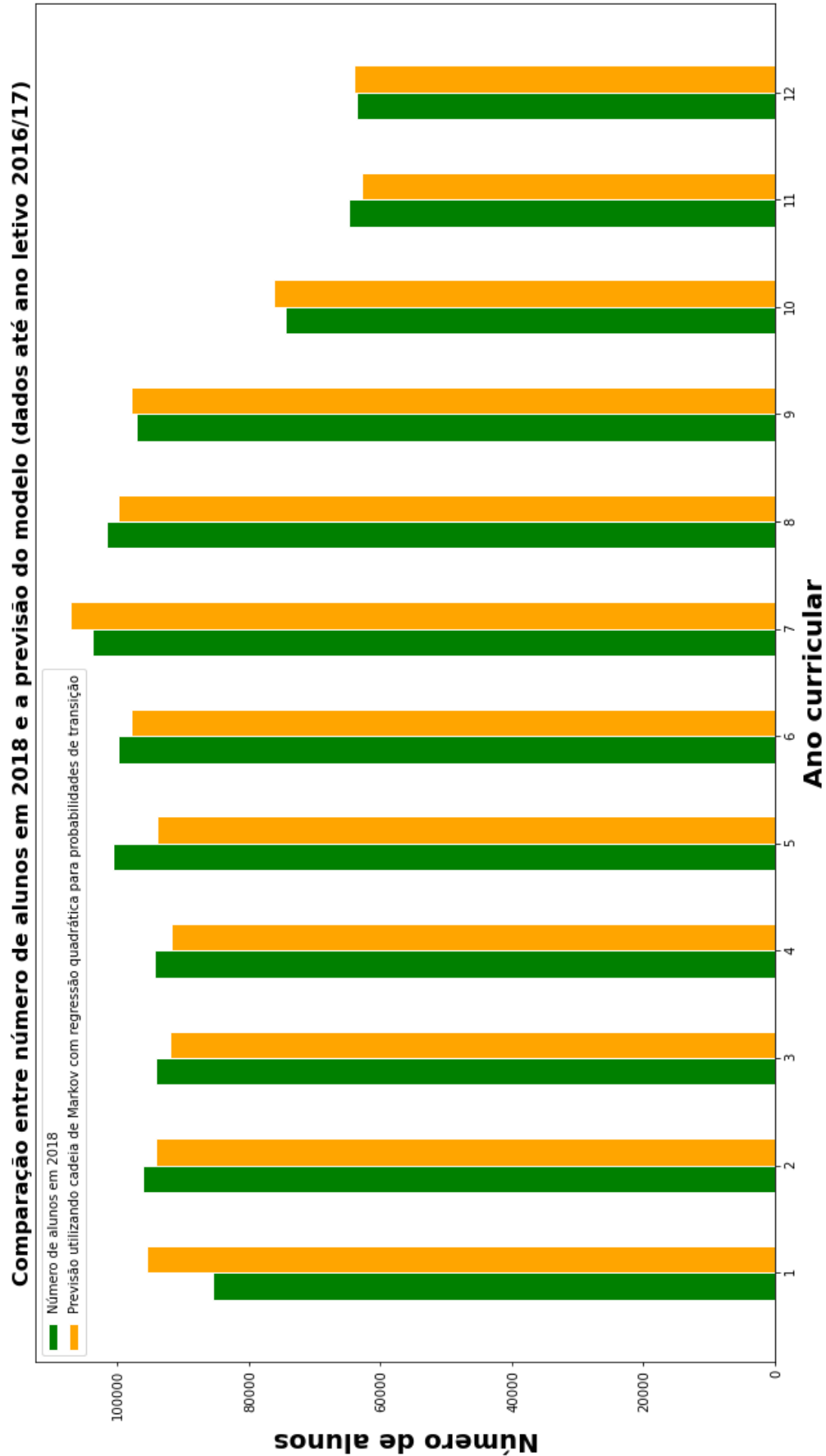
[illegible][illegible][illegible]

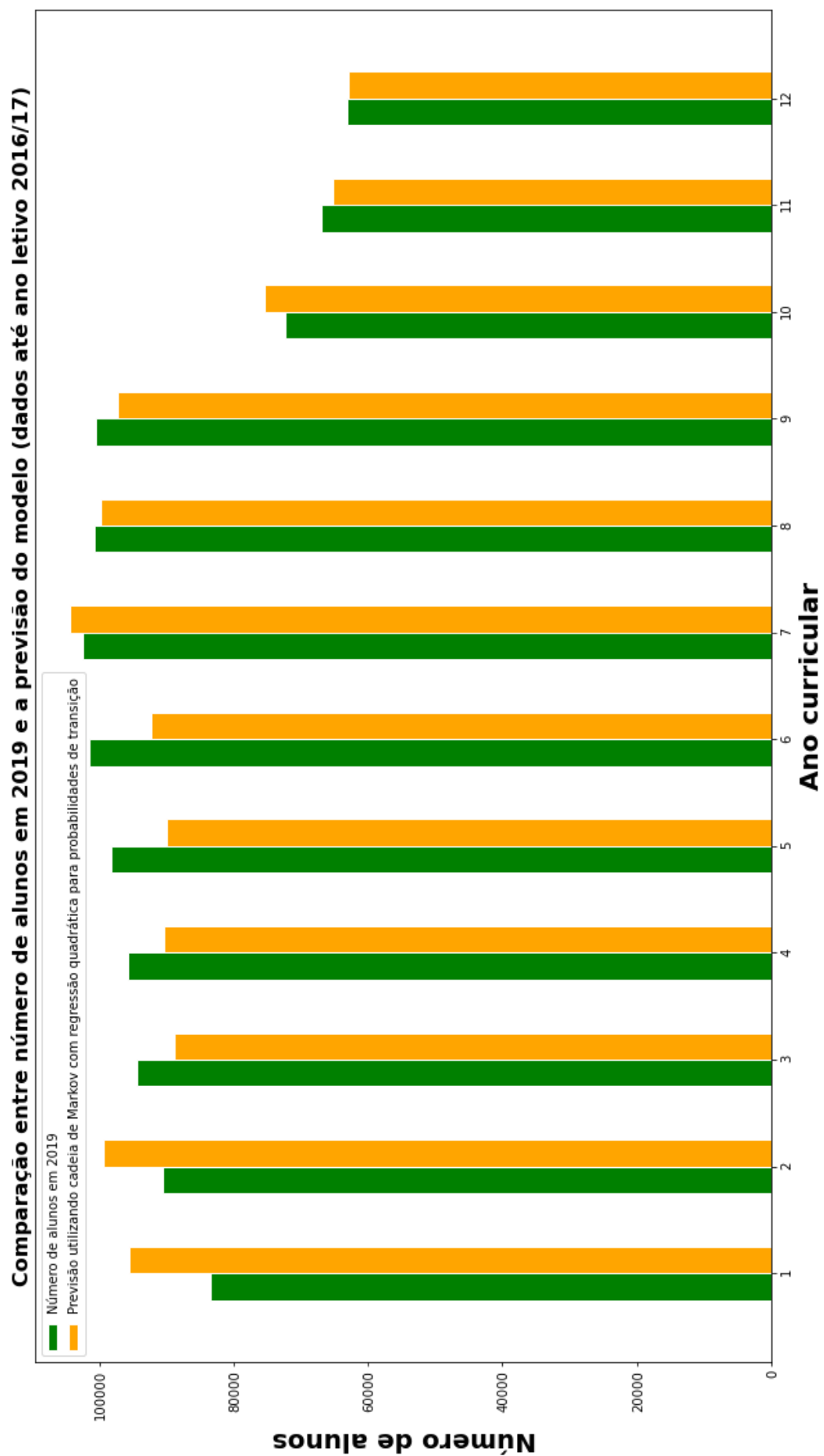
Apêndice K. Resultados dos modelos de probabilidades dinâmicas utilizando regressões polinomiais (ensino comum)

Modelo com dados até o ano letivo 2017/18



Modelos com dados até o ano letivo 2016/17





Modelos com dados até o ano letivo 2015/16

