INSTITUTO UNIVERSITÁRIO DE LISBOA

2020 / 2021 (2º semestre)

Projeto Aplicado a Ciência de Dados I
Professora Diana Mendes

Análises de Regressão para os sectores da Alta Tecnologia, Média-Alta Tecnologia e Conhecimento Intensivo, com Indicadores de Investimento, Emprego e Pesquisa e Desenvolvimento Enquanto Preditores

Turma LCD-PL

Catarina Castanheira N. 92478
João Martins N. 93259
Joel Paula N. 93392

Índice

Introdução

Lançado o desafio de explorar os dados relativos a indicadores de impacto no PIB dos sectores da alta e média-alta tecnologia e serviços de conhecimento intensivo, procurámos inicialmente complementar os dados com outros indicadores económicos. Pareceu-nos interessante explorar indicadores que pudessem estar relacionados ou mesmo até influenciar aqueles.
Com isto em mente, decidimos integrar também dados relativos a: investimentos privados em start-ups; emprego segregado por sector; investimento em pesquisa e desenvolvimento, discriminados pelos sectores empresarial privado, governo, ensino superior, empresas privadas sem fins lucrativos.
O nosso objetivo passa por procurar definir modelos de previsão para as variáveis sobre o valor acrescentado em alta/média alta tecnologia e também em sectores de conhecimento intensivo com base nos restantes indicadores económicos.
A questão principal que guia este estudo é: quais destas medidas têm melhor potencial explicativo do PIB por sector?
Neste relatório começamos por fazer uma análise exploratória dos dados em jogo, passando brevemente pelo método de transformação e "otimização" dos dados, para finalmente construir, adaptar e interpretar os modelos obtidos.
Datasets
Para este projecto foi determinado o seguinte conjunto de *datasets*:
RIO - *Value added for knowledge intensive services* (KIS) *and high-tech and medium-high-tech industries as % of total value added* – o nosso dataset principal, contend os indicadores de valor acrescentado ao PIB por setor;
Employment - *Employment in high-tech and medium-high-tech industries as % of total value added* - O dataset sobre a percentagem da força de trabalho contida nestes setores;
Venture - *Venture Capital Investments* – investimento de capital de risco;
GERDS - *Gross Expenditure in Research & Development* – investimento em Investigação e Desenvolvimento, por setor (privado, educação universitária, público).
O *dataset* principal deste projeto foi retirado da plataforma EUROSTAT (*Statistical Office of the European Union*). É um *dataset* com dados de 2000 até 2018, relativos ao valor acrescentado por *knowledge intensive services* (KIS) e *high-tech* and *medium-high-tech* industries como percentagem do valor acrescentado total.
Designam-se *Knowledge Intensive Services* (KIS) todos os serviços que são muito dependentes em conhecimentos profissionais. Exemplos de setores económicos de atividade classificados como KIS pelo classificador NACE (classificador de atividades económicas na UE) são: Saúde Publica, Segurança e Investigação e Serviços de Informação.
Designam-se por *High-Tech Industries* todas as indústrias produtoras de alta tecnologia. Exemplos de indústrias classificadas como *High-Tech* pelo classificador NACE são: manufatura de produtos e preparados farmacêuticos, eletrónica e produtos óticos e manufatura de maquinarias relacionadas com aeronaves.
Por fim, designam-se *Medium-High Tech Industries* todas as indústrias que utilizam equipamentos de média-alta tecnologia. Exemplos de indústrias classificadas como *Medium-High-Tech* pelo classificador NACE são: manufatura de produtos químicos, veículos, reboques e semirreboques e equipamento elétrico.

Compreensão dos Dados

Durante a análise dos dados foi necessário fazer uma restruturação das tabelas para que tivessem os dados de uma forma consistente. Após as restruturações, obtivemos a seguinte síntese:

Dataset	**Anos**	**Nº Países**	**Nº variáveis**	**Nº observações**	***Missing values***
RIO	2000-2019	28	15	8575	192
Employment	2008-2019	35	4	912	27
Venture Capital	2007-2015	20	3	540	0
GERDS	2003-2019	41	18	1189	375

Dataset - Value added for knowledge intensive services (KIS) and high-tech and medium-high-tech industries as % of total value added

Neste *dataset* podemos observar a percentagem que cada setor – indústria de alta tecnologia, indústria de média-alta tecnologia e serviços de conhecimento intensivo – têm nos setores da indústria e dos serviços. O seu impacto mantém-se mais ou menos estável no período analisado, existindo uma tendência para aumentar o impacto da indústria de média-alta tecnologia e um desacelerar do impacto da indústria de alta tecnologia:

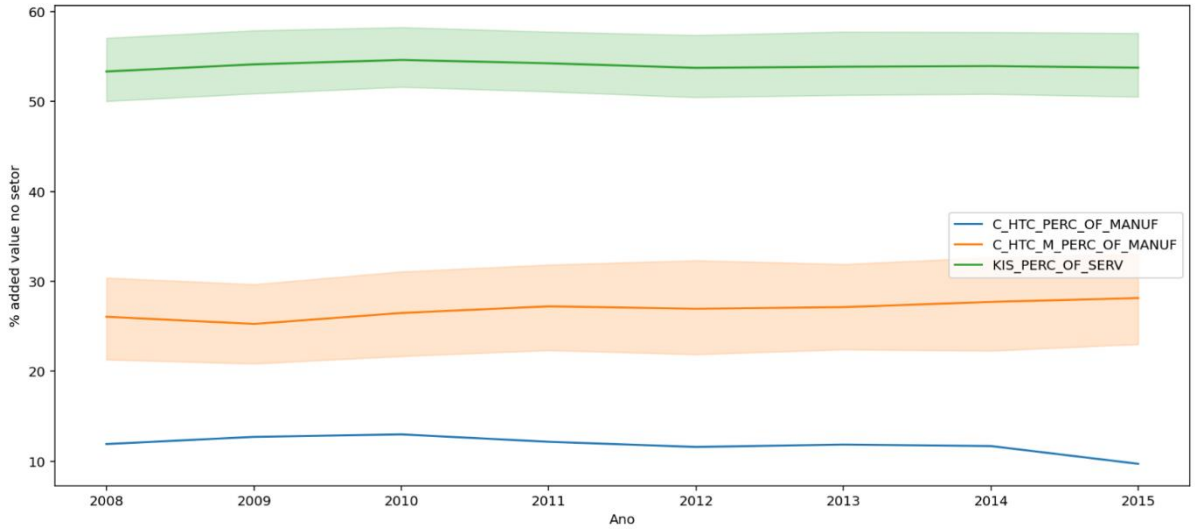


Figura 1 Percentagem média de valor acrescentado, dentro do seu setor

Olhando com mais detalhe para a média do impacto percentual da indústria de alta tecnologia no sector da manufatura em cada país:

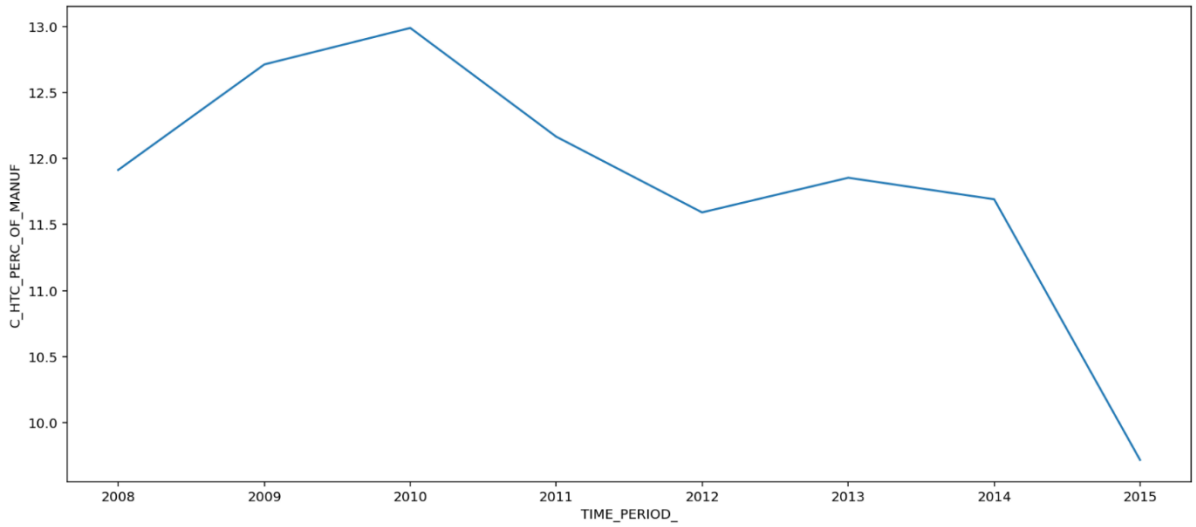


Figura 2 Percentagem média de valor acrescentado pela indústria de alta tecnologia no setor da indústria

Medindo agora em milhões de Euro:

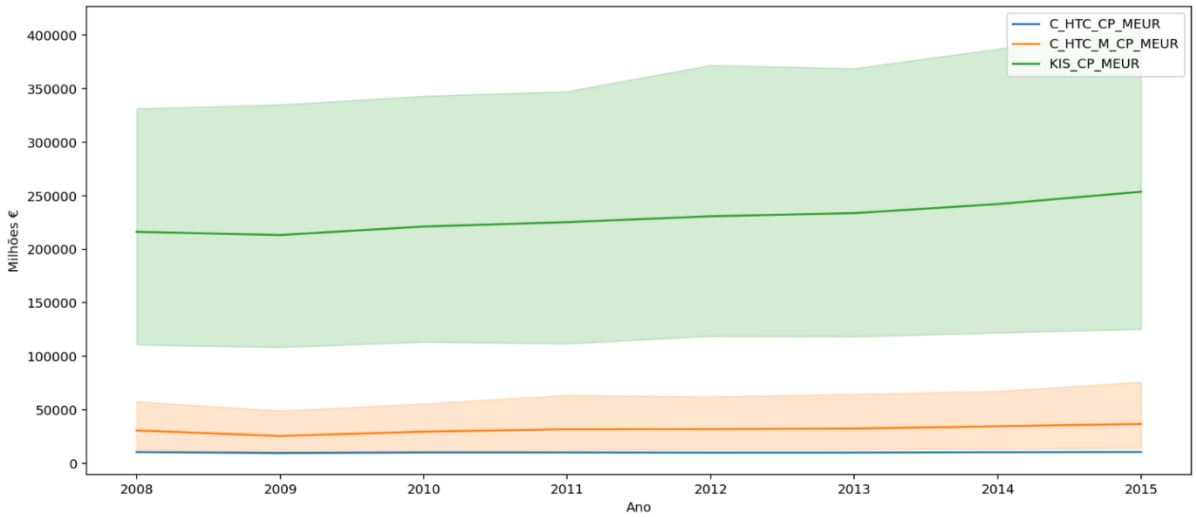


Figura 3 Valor médio adicionado à economia de cada país, por setor, em milhões de €

Verificamos que o setor dos serviços de conhecimento intensivo tem cada vez mais impacto nas economias dos respetivos países.
Por comparação, o impacto da indústria de Alta tecnologia é bem menor, embora muitas vezes os serviços de conhecimento intensivo sejam prestados exatamente a este tipo de indústrias.

Dataset - Employment in high- and medium-high technology manufacturing sectors and knowledge-intensive service sectors

Começamos por olhar para a média da percentagem de trabalhadores empregada no setor dos serviços de conhecimento intensivo (KIS – *Knowledge Intensive Services*):

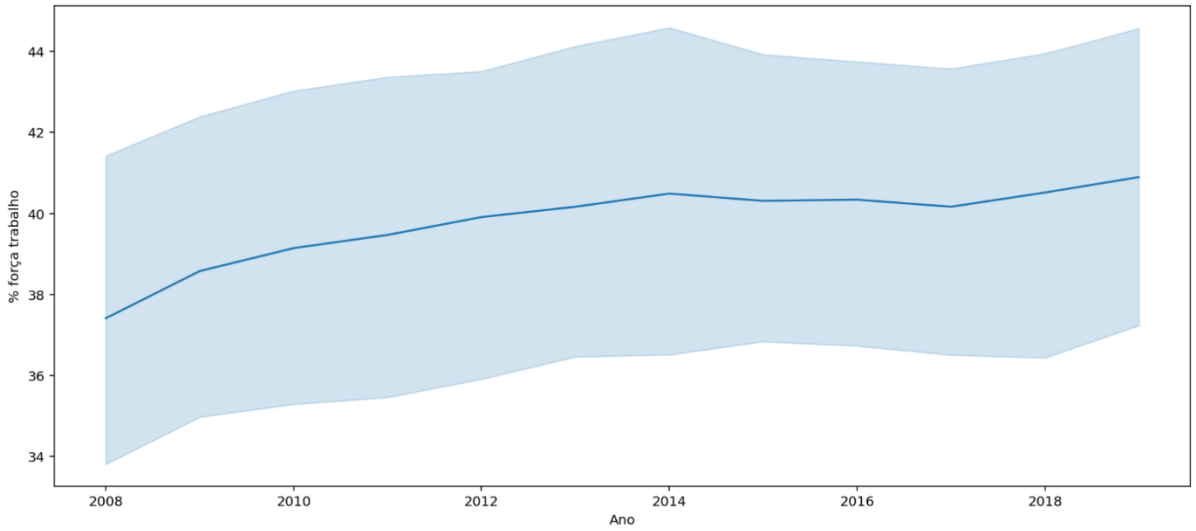


Figura 4 Média da percentagem de população trabalhadora empregada no setor dos serviços de conhecimento intensivo (KIS)

O setor dos serviços com trabalhadores qualificados tem vindo a crescer em todos os países, englobando entre 1/3 a quase metade do mercado de trabalho. Isto tem exigido investimento em qualificação por parte de todos os países.

Já os setores da indústria de alta e média-alta tecnologia têm-se mantido como empregadores estáveis:

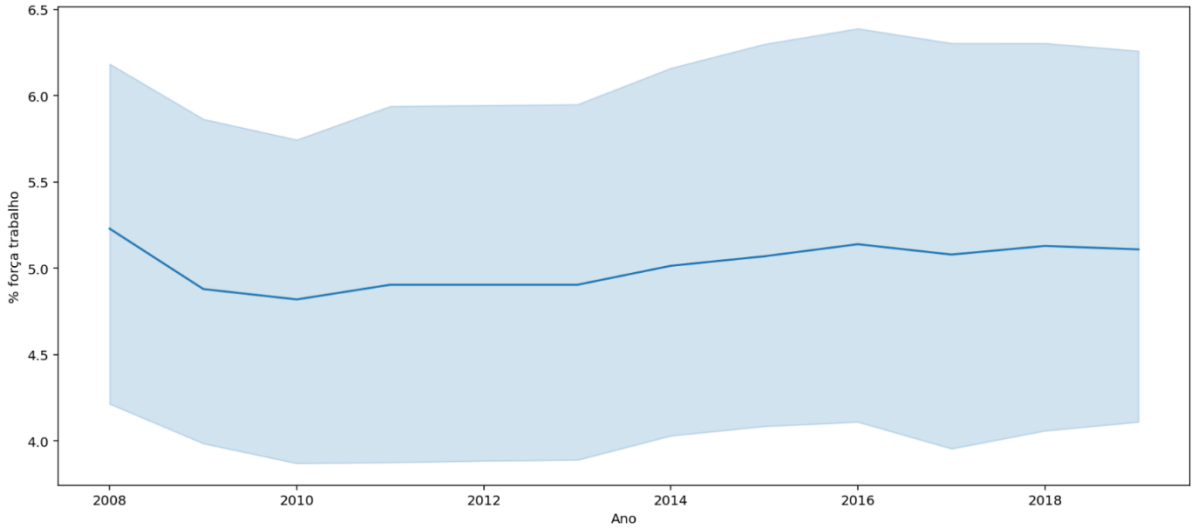


Figura 5 Média da percentagem de população empregada na industria de alta e média-alta tecnologia

Dataset - Venture Capital Investments

Gráfico que representa o investimento médio em Capital de Risco (*Venture Capital*) em cada país:

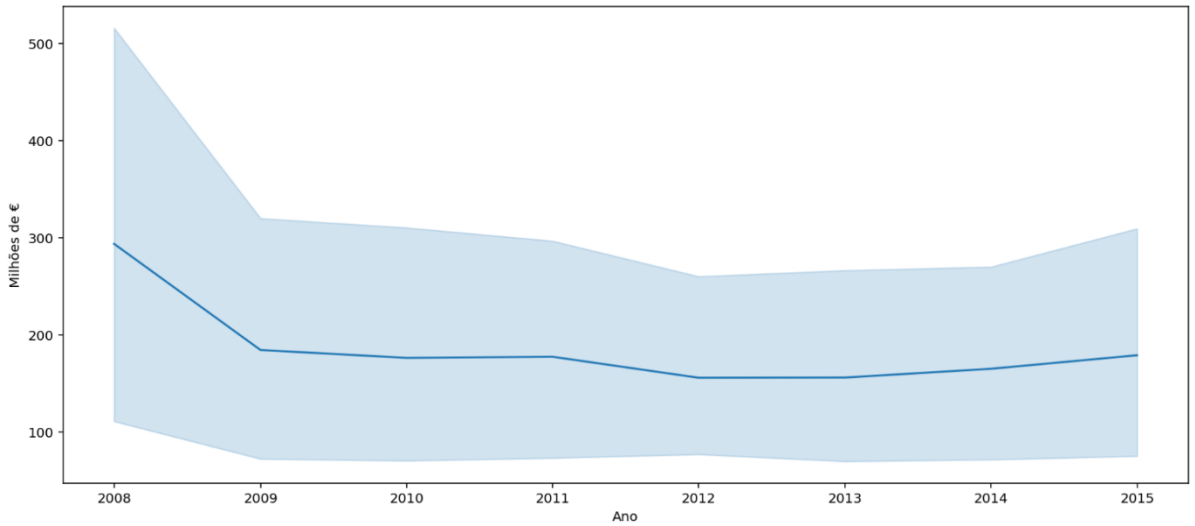


Figura 6 Investimento médio em Capital de Risco, em Milhões de Euro

Podemos observar aquilo que suspeitamos sejam os efeitos da crise de 2008, o impacto da crise da dívida soberana em 2011 e a recuperação a partir daí.
Já o número de investimentos parece ter sempre diminuído no período em análise, o que, tendo em conta o gráfico anterior, pode indicar um aumento do valor investido em cada caso.

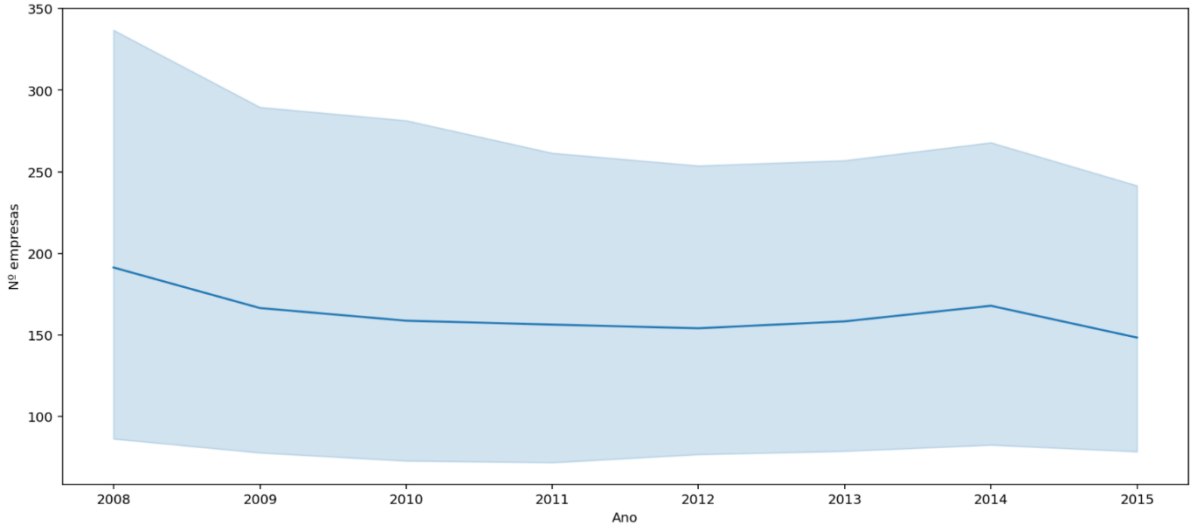


Figura 7 Nº de empresas com investimento de capital de risco

Tendo em conta os dois gráficos anteriores, podemos concluir que o pico no gráfico que expressa o investimento em percentagem do PIB poderá ser consequência da baixa do PIB de alguns países, na sequência da crise da dívida soberana.

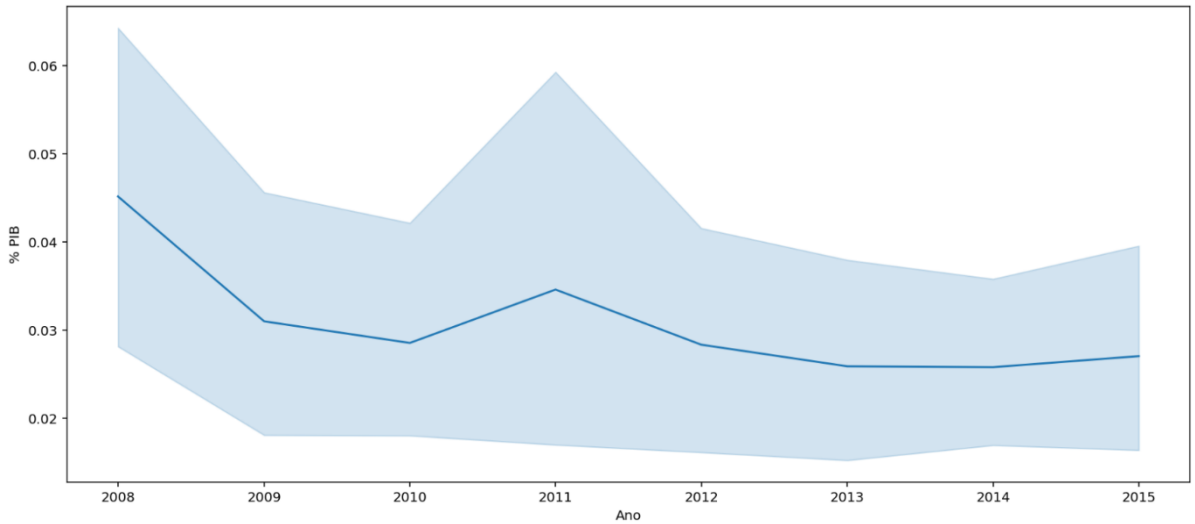


Figura 8 Média do Capital de Risco investido, em percentagem do PIB

Dataset - Gross Expenditure in Research & Development (GERD)

Usando a média de despesas em Euro por habitante dos 20 países:

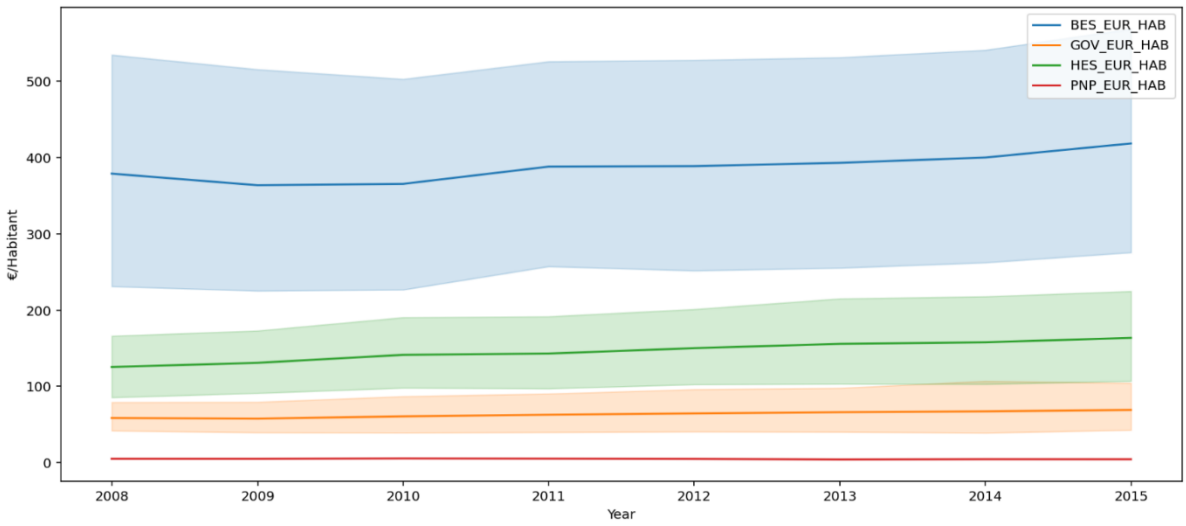


Figura 9 Despesa média em R&D por setor, em Euro por habitante

Podemos observar uma queda do investimento privado a partir da crise financeira de 2008 e um reforço do investimento em 2011 (durante a crise da dívida soberana de alguns países europeus) e o seu crescimento anual desde então.
No que toca ao investimento de instituições de ensino superior, têm crescido sistematicamente. Já no caso do investimento governamental nas atividades de Investigação e Desenvolvimento, os níveis de investimento parecem manter-se durante o período.

Olhando para as mesmas medidas, agora expressas em milhões de Euros:

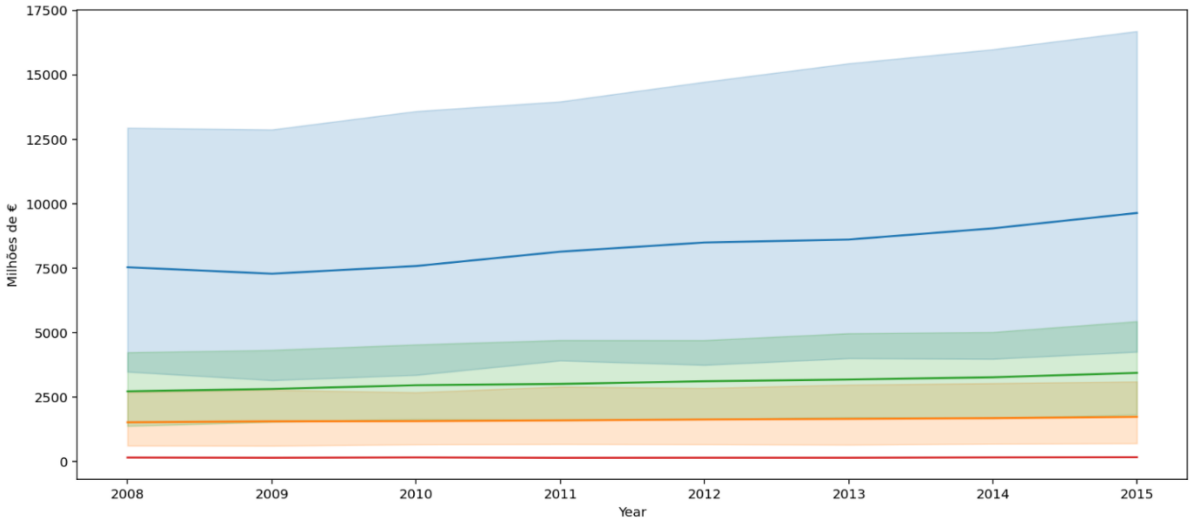


Figura 10 Despesa média em R&D por setor, em milhões de Euro

Olhando com mais detalhe ao setor das instituições privadas sem fins lucrativos e os investimentos em R&D, pois a sua escala é bem menor do que os outros setores:

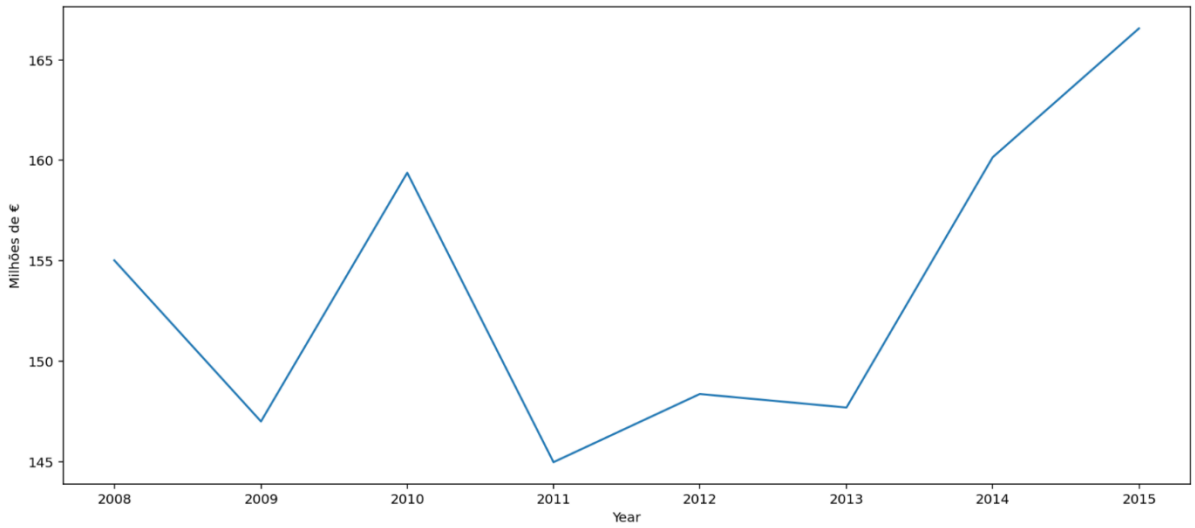


Figura 11 Despesa média em R&D no setor PNP (privadas sem fins lucrativos), em milhões de Euro

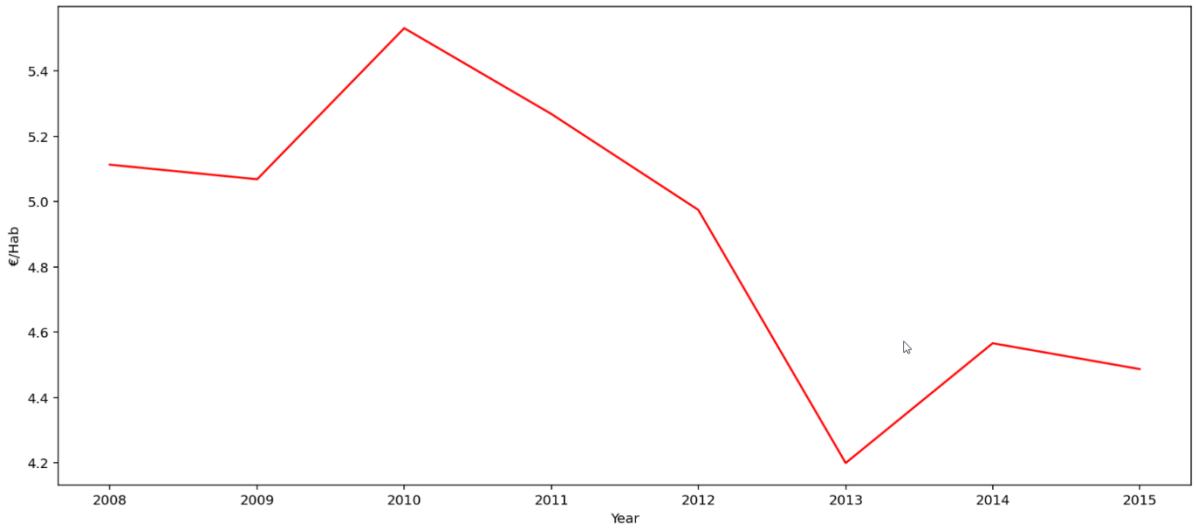


Figura 12 Despesa média em R&D no setor PNP (privadas sem fins lucrativos), em Euro por habitante

Nota-se uma tendência de aumento do investimento total neste setor, acima de tudo nos países maiores, pois o investimento por habitante não teve um impacto tão grande.

Preparação dos Dados

Para preparar os dados para a sua integração definimos um período comum a todos os *datasets*, desde 2008 até 2015, selecionando também apenas os países comuns ao *dataset* com o menor número de países (RIO_VENTURE) e selecionámos as variáveis mais interessantes para o nosso estudo em cada *dataset*:

- *RIO*: TIME_PERIOD, GEO, C_HTC_PERC_OF_MANUF, C_HTC_CP_MEUR, C_HTC_M_PERC_OF_MANUF, C_HTC_M_CP_MEUR, KIS_PERC_OF_SERV, KIS_CP_MEUR
- *Employment*: YEAR, GEO, EMPLOYMENT_C_HTC_MH, EMPLOYMENT_KIS
- *Venture*: TIME_PERIOD, GEO, MIO_EUR, NR_COMP, PC_GDP
- *GERDS*: TIME_PERIOD, GEO, BES_EUR_HAB, BES_MIO_EUR, GOV_EUR_HAB, GOV_MIO_EUR, HES_EUR_HAB, HES_MIO_EUR, PNP_EUR_HAB, PNP_MIO_EUR

Após a preparação das tabelas fizemos a integração das mesmas no *dataset* principal, RIO, ficando com um total de 160 observações e 32 variáveis.

Obtivemos 167 *missing values*, destes 167 destacaram-se duas situações:

- *Missing values* relacionados com o valor acrescentado em setores de *High-Tech Manufacturing* em Luxemburgo e na Irlanda, nalguns anos;
- *Missing values* relacionados com dados de despesas no setor privado sem fins lucrativos em R&D nalguns países em determinados anos.

Na primeira situação, após alguma pesquisa concluímos que Luxemburgo não tinha *High-Tech Manufacturing*, por isso substituímos esses *missing values* por 0. Quanto à Irlanda, a informação que obtemos sugere que este país tem *High-Tech Manufacturing.* Baseando-nos nisto, decidimos substituir os *missing values* com os valores referentes aos do ano anterior. Relativamente à segunda situação, podemos assumir que os *missing values* representam a ausência de despesas no setor privado sem fins lucrativos em R&D, nestes casos, substituímos os valores por 0.

Análise de Correlações

Ao analisar as correlações (consultar Anexo 1 – tabela de correlações), concluímos que existiam muitas elevadas e que existiam variáveis com a mesma informação que outras, indicado por uma correlação muito alta entre essas.

Outliers

Para identificar *outliers* usámos o limite de 3 desvios padrão da média para cada variável e com isto identificámos *outliers* em 12 variáveis. No entanto decidimos não remover esses *outliers* uma vez que, na maioria dos casos, iríamos estar a eliminar todas ou a maioria das observações de um país apenas por ter valores observados elevados relativamente às variáveis acima. Além disto, os *outliers* não correspondem a erros nos dados.

Modelização

A nossa escolha em termos de modelos recaiu sobre a Regressão Linear e usámos o modelo de regressão OLS (*Ordinary Least Squares*) tendo como objetivo minimizar as diferenças entre os dados observados e os dados previstos pela regressão linear.

Regressão Linear: OLS (*Ordinary Least Squares*)

Para fazer esse modelo de regressão linear, normalizámos as variáveis cujas medidas representam valores absolutos (número de empresas, ou milhões de euros).
Depois avançámos para a seleção das *features*, usando o método *SelectKBest* para escolher os melhores *preditores* para as variáveis *target* baseado no *f-score* e *p-value* de cada variável.
As variáveis referentes ao país e ao ano não foram consideradas nos modelos.
De seguida, dividimos os dados em amostra de treino e de teste, numa proporção de 80% e 20%, respetivamente.
Para evitar usar variáveis que não fossem significativas ou que fossem multicolineares nos modelos, fizemos uma eliminação das *features* (selecionadas previamente pelo *SelectKBest*) usando uma *stepwise feature elimination*.
Consultar as *features* selecionadas no anexo III.
Consultar também anexo IV, para verificar a performance de cada modelo na amostra de treino e na validação cruzada.

Teste do modelo com o *dataset* de teste.

Tendo em conta os resultados dos testes anteriores decidimos testar alguns dos modelos, para verificar a sua performance com os dados de teste:

Variável target	Modelo	R^2 ajustado	MAE	MSE
C_HTC_PERC_OF_MANUF	1 - Regressão linear	0.25	4.79	54.15
	2 – Regressão polinomial	-0.21	5.73	87.37
	3 – log da target	0.40	0.30	0.19

C_HTC_CP_MEUR	1 - Regressão linear	0.92	0.05	0.00
	2 – Regressão polinomial	0.93	0.04	0.00
	3 – log da target	0.67	0.64	0.68
C_HTC_M_PERC_OF_MANUF	1 - Regressão linear	0.73	4.06	35.28
	2 – Regressão polinomial	0.83	3.41	22.99
C_HTC_M_CP_MEUR	1 - Regressão linear	0.94	0.03	0.00
	2 – Regressão polinomial	0.99	0.01	0.00
KIS_PERC_OF_SERV	1 - Regressão linear	0.59	3.45	21.08
	2 – Regressão polinomial	0.62	3.24	19.84
KIS_CP_MEUR	1 - Regressão linear	0.96	0.04	0.00
	2 – Regressão polinomial	0.97	0.03	0.00

Visualizando os resultados para cada variável, para o modelo com melhor performance:

C_HTC_PERC_OF_MANUF

Percentagem de valor acrescentado à indústria pela indústria de alta tecnologia.

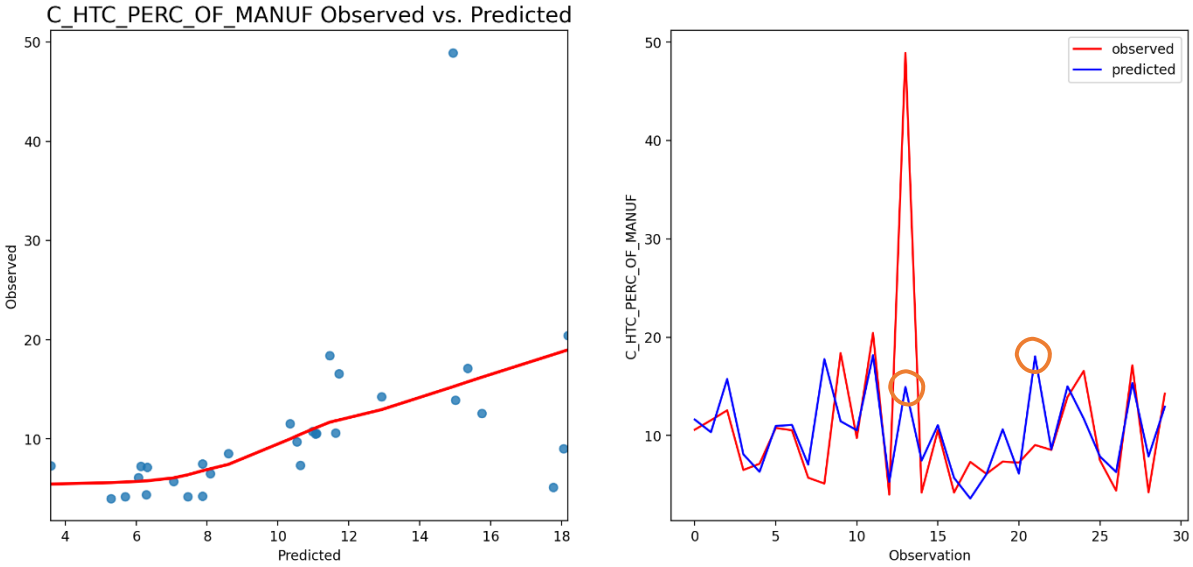


Figura 13 Preditos vs reais relativos a Percentagem de valor acrescentado à indústria pela indústria de alta tecnologia

Os dois pontos com maiores diferenças assinalados no gráfico, dizem respeito a Irlanda em 2011 e Suécia em 2012.

C_HTC_CP_MEUR

Valor acrescentado à indústria pela indústria de alta tecnologia, em milhões de Euro.

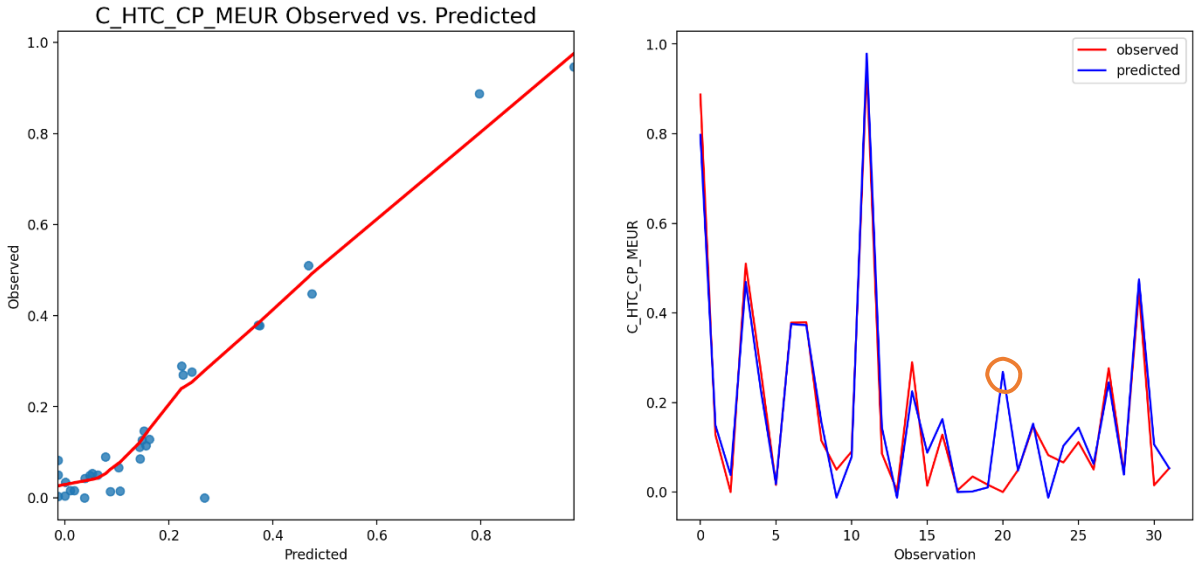


Figura 14Preditos vs Reais para Valor acrescentado à indústria pela indústria de alta tecnologia, em milhões de Euro

O erro assinalado no gráfico diz respeito ao Luxemburgo em 2015, que é um país que não tem indústria *High-Tech*.

C_HTC_M_PERC_OF_MANUF

Percentagem de valor acrescentado à indústria pela indústria de media-alta tecnologia.

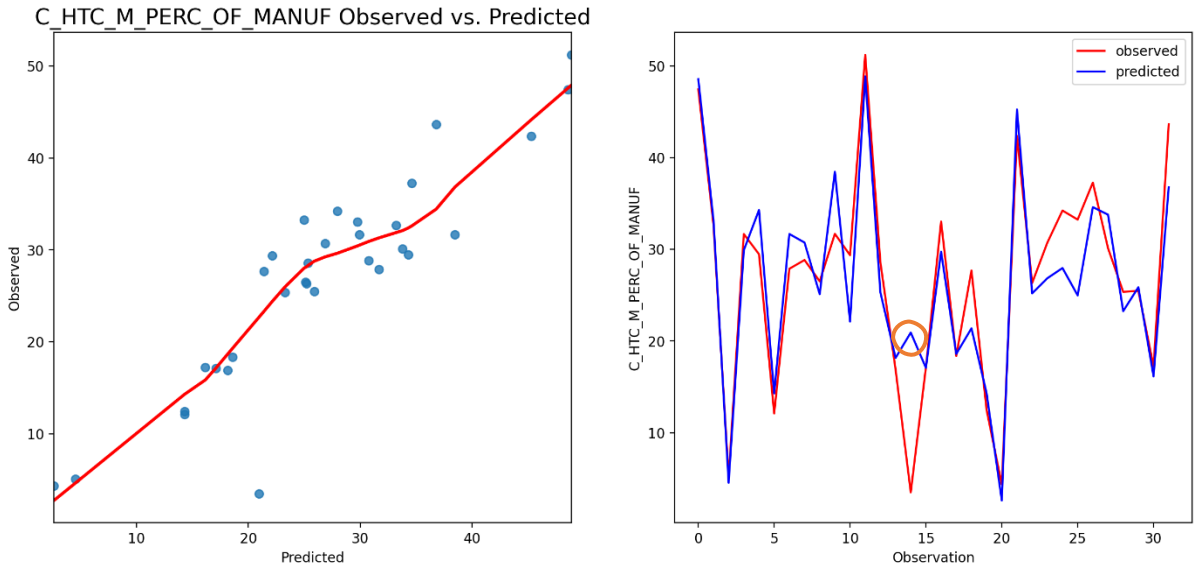


Figura 15 Preditos vs Reais para Percentagem de valor acrescentado à indústria pela indústria de media-alta tecnologia

O erro assinalado no gráfico diz respeito à Irlanda em 2011.

C_HTC_M_CP_MEUR

Valor acrescentado à indústria pela indústria de média-alta tecnologia, em milhões de Euro.

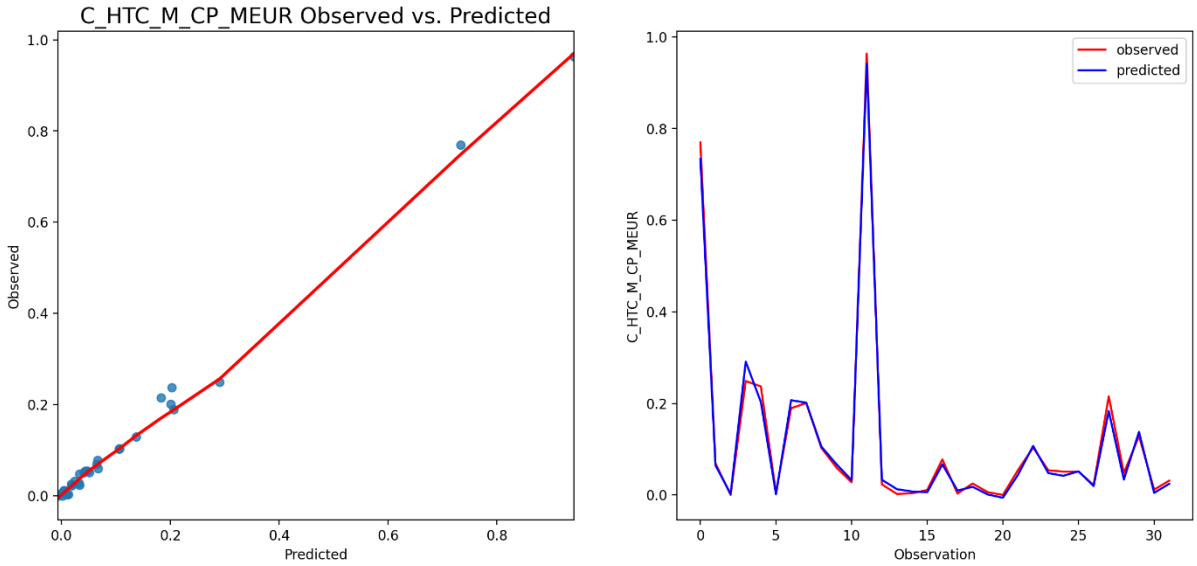


Figura 16 Preditos vs Reais para Valor acrescentado à indústria pela indústria de média-alta tecnologia, em milhões de Euro

KIS_PERC_OF_SERV

Percentagem de valor acrescentado ao setor dos serviços pelos serviços *knowledge-intensive*.

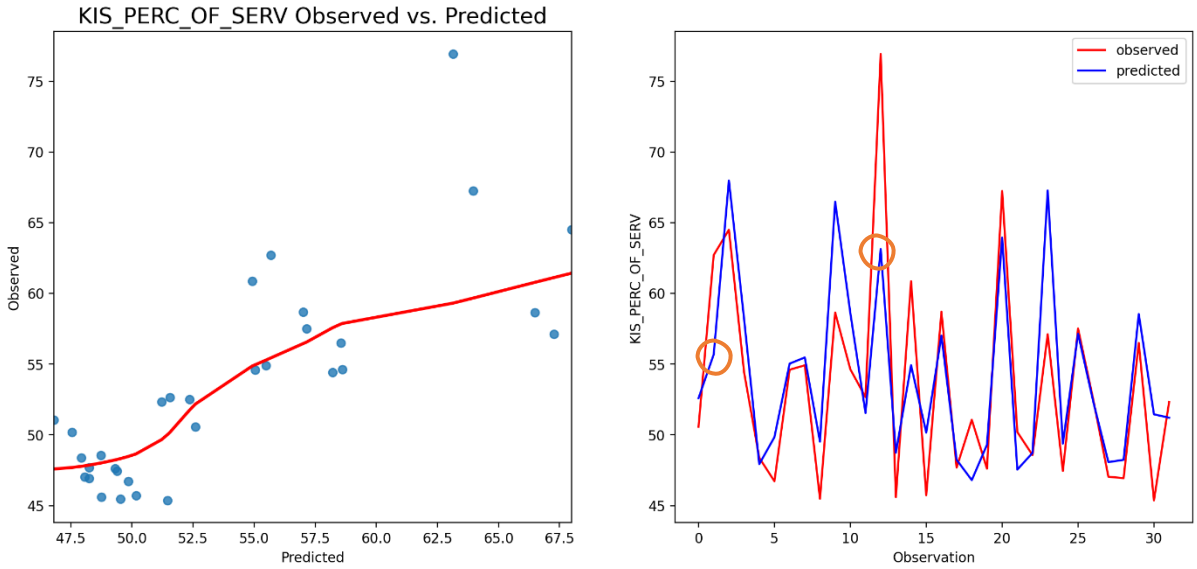


Figura 17 Preditos vs Reais para Percentagem de valor acrescentado ao setor dos serviços pelos serviços knowledge-intensive

Os erros assinalados no gráfico dizem respeito à Dinamarca em 2009 e Suécia em 2008.

KIS_CP_MEUR

Valor acrescentado ao setor dos serviços pelos serviços *knowledge-intensive*, em milhões de Euro.

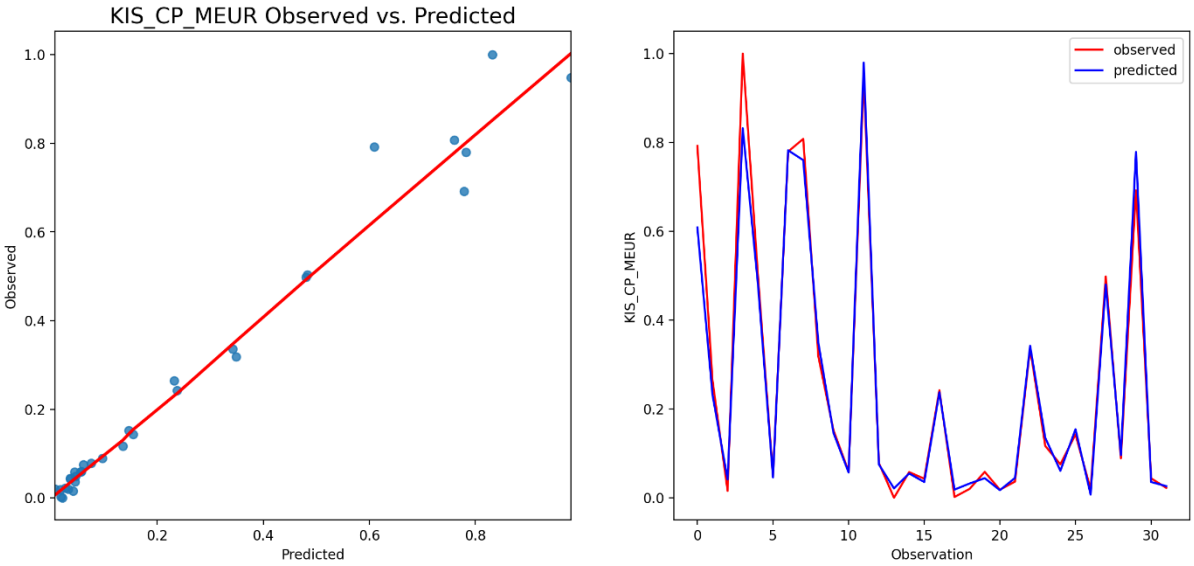


Figura 18 Preditos vs Reais para Valor acrescentado ao setor dos serviços pelos serviços knowledge-intensive, em milhões de Euro

Interpretação dos resultados

Através dos resultados das previsões efetuadas, é possível perceber que quanto mais complexo é o modelo, isto é, quantas mais as variáveis a que ele recorre para explicar a variabilidade da variável-alvo, maior é a sua eficácia. Basta olhar por exemplo para os modelos lineares das variáveis C_HTC_M_CP_MEUR e KIS_CP_MEUR cujos desempenhos sobre o conjunto de teste devolveram um R^2 de 0.99 e de 0.97, respetivamente. Estes modelos basearam a sua previsão em 6 variáveis cada um. Significará isto que ambas as variáveis têm a quase totalidade da sua informação explicada através das respetivas variáveis preditoras.
No caso do valor acrescentado da média-alta tecnologia (milhões de euros) ele consegue ser explicado em função de: capital investido em *start-ups*, a proporção de empregados no sector da média-alta tecnologia, e também em função de indicadores de investimento em pesquisa e desenvolvimento (de governo, instituições privadas sem fins lucrativos e empresas privadas).
As medidas relativas ao emprego e as de despesas em pesquisa e desenvolvimento surgem sempre como variáveis explicativas de todas as alvo que temos definidas, pese embora no primeiro caso em análise, não tenhamos encontrado nenhum modelo que explicasse mais que 41% do total da variabilidade no valor acrescentado obtido em alta-tecnologia, como uma proporção do total da indústria de produção. Nos modelos relativos a esta variável, não encontramos mesmo nenhuma medida relacionada com o investimento privado em *start-ups* a surgir como uma das variáveis explicativas. Poderá isto significar que não existem start-ups em alta-tecnologia, que não existe investimento de capital de risco em alta-tecnologia?
Nos modelos analisados, verificamos que a proporção de empregados no sector dos *knowledge intensive services* surge como um dos preditores nas medidas relacionadas com a alta-tecnologia, mas não com as medidas relacionadas com a média-alta tecnologia, denotando aqui possivelmente uma distinção nas importâncias deste tipo de medida de emprego num e noutro tipo de sectores. A criação de riqueza a partir dos sectores da alta tecnologia parece estar mais condicionado à proporção de emprego nos serviços com alta qualificação (KIS) do que os sectores da média-alta tecnologia.
Claramente os melhores modelos preditivos são os modelos polinomiais. Contudo, a "margem" do seu desempenho comparativamente aos modelos lineares só poderá ser considerada substancial no caso das previsões para a variável 3 – o valor acrescentado do sector da média-alta tecnologia como uma percentagem do total da indústria de produção. Neste caso, o R^2 obtido na amostra de teste é superior em quase 10 pontos percentuais que aquele obtido no modelo de regressão linear (0.826 vs 0.733). Sobre esta variável, poderemos especular que a sua variabilidade é determinada através de um mais complexo relacionamento das preditoras do que aquela que é possível obter através de um "simples" modelo linear.

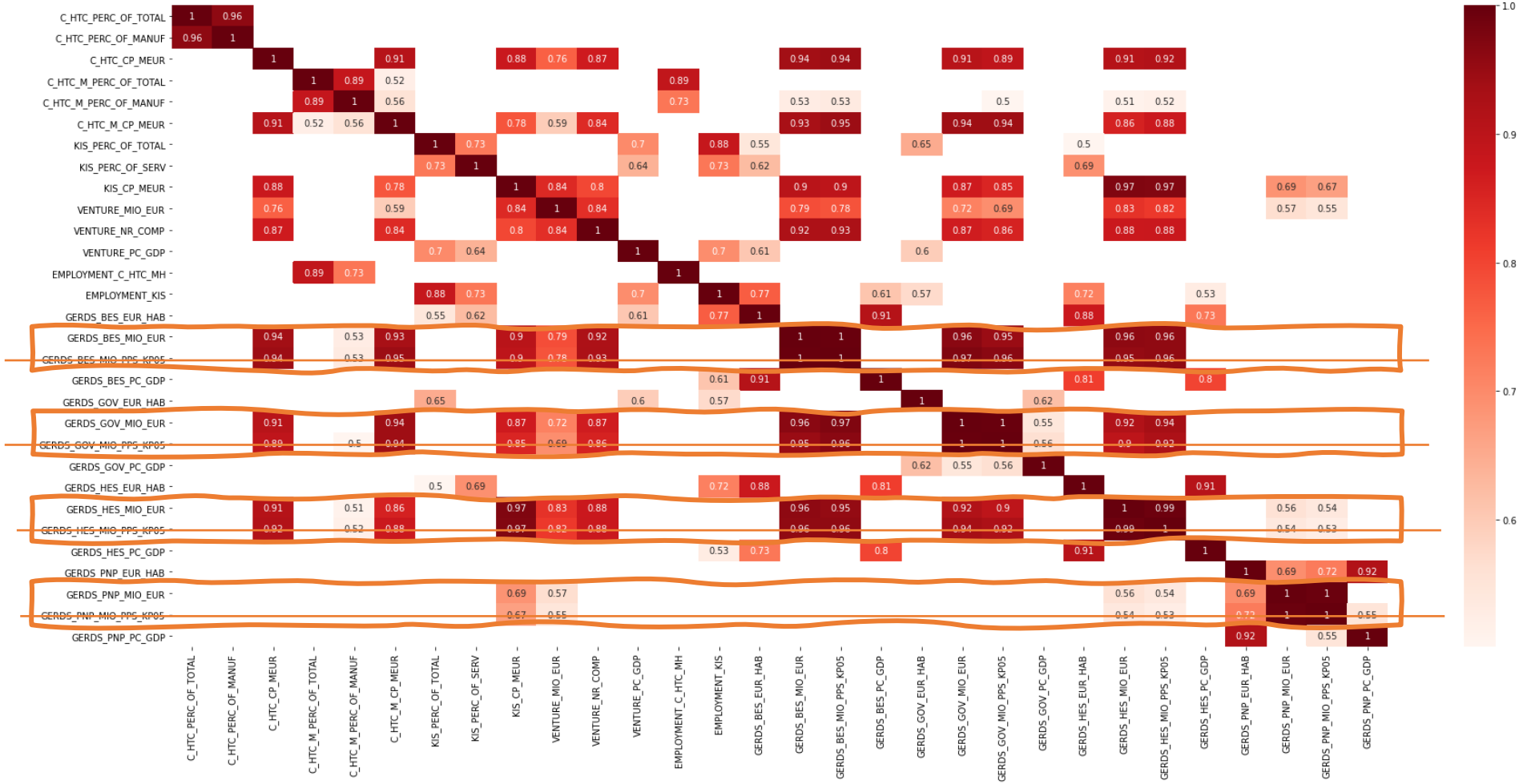
Ainda assim, o modelo linear consegue explicar cerca de 73% da variância da *target*, o que não deverá ser descurado.
Apesar desta melhoria obtida na previsão de uma das variáveis, os modelos polinomiais conduzem a uma enorme dificuldade na sua interpretação, pelo facto de os seus coeficientes dizerem também respeito à multiplicação de variáveis.
O melhor modelo de regressão linear que conseguimos obter foi o que dita a previsão do valor acrescentado em milhões de euros obtido com os KIS em função de medidas de investimento privado em *startups*, de medidas de empregabilidade tanto nos sectores da alta e média-alta tecnologia como nos sectores dos serviços de elevada qualificação, e de medidas de despesa em pesquisa e desenvolvimento (tanto governamentais como instituições privadas sem fins lucrativos). Conseguimos aqui obter um modelo que consegue explicar aproximadamente 96% da variabilidade observada nos dados de teste da nossa *target*. Já se olharmos para os modelos polinomiais, no caso da variável 4 (valor acrescentado em milhões de euros obtido através do sector da média-alta tecnologia) obtemos a quase perfeição na previsão (R^2 de 0.994), mas em termos interpretativos a única coisa que conseguimos fazer é afirmar que a relação ótima entre as variáveis para obter a melhor previsão não é linear.
Já o pior modelo de regressão linear que obtivemos foi o que tenta explicar a variável 1 (C_HTC_PERC_OF_MANUF) através da empregabilidade no sector da alta e média-alta tecnologia e também dos *knowledge intensive services*, e através das despesas em pesquisa e desenvolvimento levadas a cabo por instituições governamentais. Apesar de estes terem sido os melhores preditores para a variável, a variância explicada obtida é relativamente baixa. Existirá aqui a probabilidade ou de os dados dos preditores não terem uma relação linear com a variável *target*, ou então de não termos na nossa posse informação suficiente que explique melhor a variabilidade. A obtenção de dados de outras fontes poderá ser considerada em estudos futuros.
Um outro dado observado é o facto de todas as variáveis terem a elas associadas pelo menos um preditor relacionado com o investimento em pesquisa e desenvolvimento por parte de instituições governamentais ou de ensino superior, exceto no caso da variável 3 (medida de proporção do valor acrescentado do sector da média-alta tecnologia). Aqui, a componente de investimento em pesquisa e desenvolvimento que tem importância na previsão é a relacionada com instituições do ensino superior.
Numa análise mais refinada aos resultados da previsão em cada modelo, constatamos que existem observações que de forma sistemática integram o topo daquelas com maior erro na previsão. Um exemplo disto são os dados relativos à Irlanda, no ano de 2011. Outro caso são as observações relativas ao Luxemburgo, para as medidas de alta-tecnologia: aqui a previsão dos modelos é consistentemente errada, porque este país não tem nenhuma representação deste sector na sua economia.
Finalmente, conseguimos verificar o seguinte:

- As variáveis-alvo representadas com valores absolutos em milhões de euros veem nos seus modelos de previsão sempre a presença de 5 ou 6 variáveis preditoras; mas no caso em que as variáveis-alvo são representadas com proporções, existe somente a presença de 3 ou de 4 variáveis preditoras. Poderá isto significar que têm mais informação subjacente que as primeiras?
- A determinação dos melhores modelos de regressão linear para a variável KIS_PERC_OF_SERV (valor acrescentado por parte dos KIS como uma proporção da totalidade dos serviços no país) inclui sempre como preditores variáveis que representam também elas proporções e não valores absolutos;
- A execução do método de validação cruzada na amostra de treino em qualquer um dos modelos levou sempre à obtenção de resultados para o R^2 inferiores àqueles que acabamos por obter quando executamos a previsão nos dados de teste. Ou seja, os modelos têm na sua generalidade um melhor desempenho que o inicialmente previsto.

Em estudos futuros, existem alguns aspetos que serão interessantes explorar. O primeiro será analisar o potencial preditivo de cada uma das variáveis preditoras, de forma individualizada, ou por grupos de variáveis (variáveis de capital de investimento, variáveis sobre o emprego por sector, e variáveis sobre o investimento em pesquisa e desenvolvimento por tipo de entidade).

Em segundo lugar, seria interessante compreender as relações de causalidade entre as variáveis ao longo do tempo: por exemplo, terá o investimento em pesquisa e desenvolvimento por parte de entidades governamentais impacto observável em medidas de PIB (representadas pelo valor acrescentado em alta-tecnologia, média-alta tecnologia e serviços de elevada qualificação (KIS)) em anos subsequentes? Infelizmente, no âmbito deste projeto, os dados anuais disponíveis (8 observações por ano em cada país) não são suficientes para esta análise. Fontes de dados alternativas teriam de ser exploradas.
Em terceiro lugar, através da análise feita na exploração inicial dos dados e na determinação de estatísticas descritivas, ficou clara a existência de alguns grupos de países em cada variável. Definindo estes grupos através de técnicas de *clustering*, que tipo de discriminação conseguimos obter? Regional (sul, centro, leste, norte da Europa)? Por nível de desenvolvimento económico? Por nível de investimento privado? Por nível de investimento em pesquisa e desenvolvimento?...
Por fim, este *clustering* dos países conseguiria trazer novos insights sobre a relação entre as variáveis alvo e as diversas variáveis preditoras que considerámos nos nossos modelos? Poderá o *clustering* de países trazer modelos de previsão mais refinados e adequados a cada contexto regional/económico?

Anexo I – Tabela de correlações



Anexo II – Definições

KIS - Knowledge-intensive services

De acordo com Glossary:Knowledge-intensive services (KIS) - Statistics Explained (europa.eu), a categorias KIS inclui (códigos do NACE Rev.2 entre parêntesis):

- High-tech knowledge-intensive services:
  - ☐ Motion picture, video and television programme production, sound recording and music publishing activities (59);
  - ☐ Programming and broadcasting activities (60);
  - ☐ Telecommunications (61);
  - ☐ Computer programming, consultancy and related activities (62);
  - ☐ Information service activities (63);
  - ☐ Scientific research and development (72)

- Knowledge-intensive market services (excluding financial intermediation and high-tech services):
  - ☐ Water transport (50);
  - ☐ Air transport (51);
  - ☐ Legal and accounting activities (69);
  - ☐ Activities of head offices; management consultancy activities (70);
  - ☐ Architectural and engineering activities; technical testing and analysis (71);
  - ☐ Advertising and market research (73);
  - ☐ Other professional, scientific and technical activities (74);
  - ☐ Employment activities (78);
  - ☐ Security and investigation activities (80)
- Knowledge-intensive financial services:
  - ☐ Financial service activities, except insurance and pension funding (64);
  - ☐ Insurance, reinsurance and pension funding, except compulsory social security (65);
  - ☐ Activities auxiliary to financial services and insurance activities (66)
- Other knowledge-intensive services:
  - ☐ Publishing activities (58);
  - ☐ Veterinary activities (75);
  - ☐ Public administration and defence; compulsory social security (84);
  - ☐ Education (85);
  - ☐ Human health activities (86);
  - ☐ Residential care activities (87);
  - ☐ Social work activities without accommodation (88);
  - ☐ Creative, arts and entertainment activities (90);
  - ☐ Libraries, archives, museums and other cultural activities (91);
  - ☐ Gambling and betting activities (92);
  - ☐ Sports activities and amusement and recreation activities (93)

In https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Knowledge-intensive_services_(KIS)

High-tech / Medium-High-Tech Industry

De acordo com Glossary:High-tech classification of manufacturing industries - Statistics Explained (europa.eu) estas categorias incluem (códigos do NACE Rev.2 entre parêntesis):

- High-technology:
  - Manufacture of basic pharmaceutical products and pharmaceutical preparations (21);
  - Manufacture of computer, electronic and optical products (26);
  - Manufacture of air and spacecraft and related machinery (30.3)
- Medium-high-technology:
  - Manufacture of chemicals and chemical products (20);
  - Manufacture of weapons and ammunition (25.4);
  - Manufacture of electrical equipment (27);
  - Manufacture of machinery and equipment n.e.c. (28);
  - Manufacture of motor vehicles, trailers and semi-trailers (29);
  - Manufacture of other transport equipment (30) excluding Building of ships and boats (30.1) and excluding Manufacture of air and spacecraft and related machinery (30.3);
  - Manufacture of medical and dental instruments and supplies (32.5)

In https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:High-tech_classification_of_manufacturing_industries

Anexo III – Tabela de *features* por cada *Target*

Na tabela em baixo encontram-se as *features* selecionadas para cada variável *target*:

Variável alvo	Preditores selecionados	Descrição
C_HTC_PERC_OF_MANUF Percentagem de valor acrescentado à indústria pela indústria de alta tecnologia	EMPLOYMENT_C_HTC_MH	Percentagem da população empregada, na indústria de *high-tech* e *medium-high-tech*
	EMPLOYMENT_KIS	Percentagem da população empregada, no setor dos serviços *knowledge-intensive*
	GERDS_GOV_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por governos/estado, expressos em Euro por habitante
C_HTC_CP_MEUR Valor acrescentado à indústria pela indústria de alta tecnologia, em milhões de Euro	EMPLOYMENT_C_HTC_MH	Percentagem da população empregada, na indústria de *high-tech* e *medium-high-tech*
	EMPLOYMENT_KIS	Percentagem da população empregada, no setor dos serviços *knowledge-intensive*
	GERDS_BES_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por empresas privadas, expressos em Euro por habitante
	GERDS_BES_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) por empresas privadas, expressos em milhões de Euro
	GERDS_GOV_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por governos/estado, expressos em Euro por habitante
C_HTC_M_PERC_OF_MANUF Percentagem de valor acrescentado à indústria pela indústria de media-alta tecnologia	VENTURE_PC_GDP	Investimento feito por venture capital, expresso em percentagem do PIB do país
	EMPLOYMENT_C_HTC_MH	Percentagem da população empregada, na indústria de *high-tech* e *medium-high-tech*
	GERDS_HES_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições de ensino superior, expressos em Euro por habitante
	GERDS_HES_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições de ensino superior, expressos em milhões de Euro
C_HTC_M_CP_MEUR Valor acrescentado à indústria pela indústria de média-alta tecnologia, em milhões de Euro	VENTURE_MIO_EUR	Investimento feito por *venture capital*, expresso em milhões de Euro
	EMPLOYMENT_C_HTC_MH	Percentagem da população empregada, na indústria de high-tech e medium-high-tech
	GERDS_BES_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por empresas privadas, expressos em Euro por habitante
	GERDS_BES_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) por empresas privadas, expressos em milhões de Euro
	GERDS_GOV_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por governos/estado, expressos em Euro por habitante
	GERDS_PNP_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições privadas sem fins lucrativos, expressos em milhões de Euro
KIS_PERC_OF_SERV	VENTURE_PC_GDP	Investimento feito por venture capital, expresso em percentagem do PIB do país

Percentagem de valor acrescentado ao setor dos serviços pelos serviços *knowledge-intensive*	EMPLOYMENT_KIS	Percentagem da população empregada, no setor dos serviços *knowledge-intensive*
	GERDS_HES_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições de ensino superior, expressos em Euro por habitante
KIS_CP_MEUR Valor acrescentado ao setor dos serviços pelos serviços knowledge-intensive, em milhões de Euro	VENTURE_MIO_EUR	Investimento feito por venture capital, expresso em milhões de Euro
	EMPLOYMENT_C_HTC_MH	Percentagem da população empregada, na indústria de *high-tech* e *medium-high-tech*
	EMPLOYMENT_KIS	Percentagem da população empregada, no setor dos serviços *knowledge-intensive*
	GERDS_GOV_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) pelo governo/estado, expressos em milhões de Euro
	GERDS_PNP_EUR_HAB	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições privadas sem fins lucrativos, expressos em Euro por habitante
	GERDS_PNP_MIO_EUR	Investimento bruto em Investigação e desenvolvimento (R&D) por instituições privadas sem fins lucrativos, expressos em milhões de Euro

Anexo IV – Análise do modelo com amostra de treino e com validação cruzada

Em baixo estão os resultados de cada modelo para cada variável *target*:

Variável alvo	Modelo regressão	R^2 ajustado	AIC	Pressupostos Verificados
C_HTC_PERC_OF_MANUF	1 – linear	0.323	922.9	Média dos resíduos = 0 Independência dos resíduos
	2 – polinomial	0.389	915.5	Média dos resíduos = 0 Independência dos resíduos
	3 – log da target	0.416	184.1	Média dos resíduos = 0 Independência dos resíduos
C_HTC_CP_MEUR	1 – linear	0.888	-318.7	Média dos resíduos = 0 Independência dos resíduos
	2 – polinomial	0.970	-473.4	Média dos resíduos = 0 Independência dos resíduos
	3 – log da target	0.970	287.6	Média dos resíduos = 0 Independência dos resíduos Distribuição normal dos resíduos
C_HTC_M_PERC_OF_MANUF	1 – linear	0.681	836.9	Média dos resíduos = 0 Homocedasticidade dos resíduos Independência dos resíduos
	2 – polinomial	0.751	814.1	Média dos resíduos = 0 Independência dos resíduos
C_HTC_M_CP_MEUR	1 – linear	0.949	-449.5	Média dos resíduos = 0
	2 – polinomial	0.995	-732.9	Média dos resíduos = 0 Independência dos resíduos
KIS_PERC_OF_SERV	1 – linear	0.639	771.8	Média dos resíduos = 0 Independência dos resíduos
	2 – polinomial	0.751	736.3	Média dos resíduos = 0 Independência dos resíduos
KIS_CP_MEUR	1 – linear	0.958	-369.0	Média dos resíduos = 0 Independência dos resíduos
	2 – polinomial	0.996	-644.4	Média dos resíduos = 0 Independência dos resíduos

Para cada um destes modelos fizemos também uma validação cruzada, usando o método de 10-*fold*. Os resultados estão na seguinte tabela:

Variável target	Modelo	R^2 ajustado	MAE	MSE
C_HTC_PERC_OF_MANUF	1 - Regressão linear	0.17	5.38	77.57
	2 – Regressão polinomial	0.2	5.55	70.3
	3 – log da target	0.34	0.34	0.26
C_HTC_CP_MEUR	1 - Regressão linear	0.74	0.05	0.0
	2 – Regressão polinomial	0.92	0.03	0.0
	3 – log da target	0.54	0.62	0.6
C_HTC_M_PERC_OF_MANUF	1 - Regressão linear	0.64	4.11	39.85
	2 – Regressão polinomial	0.72	3.85	30.74
C_HTC_M_CP_MEUR	1 - Regressão linear	0.61	0.03	0.0
	2 – Regressão polinomial	0.91	0.01	0.0

KIS_PERC_OF_SERV	1 - Regressão linear	0.53	3.83	25.51
	2 – Regressão polinomial	0.58	3.48	23.25
KIS_CP_MEUR	1 - Regressão linear	0.92	0.04	0.0
	2 – Regressão polinomial	0.95	0.02	0.0