Bachelor Thesis
Immanuel Colombo
Student Number: 432880
Supervisor: Bijkerk, S.
First Reader: Emami Namini, J.

Does the Internet Activity within a Zip-Code Area correlate with the difference between Off- and Online Prices of Multi-Channel Retailers in the US?

Abstract:

The internet has changed the retailing sector by created a new channel of retailing other than offline (physical store), namely online (website) retailing. Although they sell the same product the two channels often offer different prices. This empirical study uses cross-sectional data to identify if internet activity that is proxied by the number of Internet Service Providers (ISP) within a zip-code area, correlates with the difference between offline and online prices of multi-channel retailers. Furthermore, the sign of the correlation is assumed to be driven by two opposing forces, namely online dynamic pricing (negative) and higher online competition (positive). The data used originates from the paper: *Are online and offline prices similar? evidence from large multi-channel retailers* (Cavallo, 2017), which contains off- and online prices of goods located in 250 US zip code areas. The observations are predominantly concentrated in the state of Massachusetts. Therefore, the conservative finding of this paper is as follow: in Massachusetts, when the online price of a good is higher than that of its offline price, via higher online competition, increasing internet activity correlates with the decrease of online prices for multi-channel retailers. No evidence of dynamic pricing is found via internet activity.

Table of content:

1. Introduction

In economics, the law of one price is fundamental for how markets work, but prices do vary. One of the first theories that explained why prices vary, is the Hotelling Location Model (1929) which shows that, due to the cost of travel, the price of the same good in one place in the world does not equal the price in another, especially if the locations are distant from one another. Another counter-example to the law of one price, or in this case purchasing power parity, is The Economist's Big Mac Index. The smaller the area of analyses for instance city, district or street, the harder it is to discriminate according to the two above mentioned theories/index. This is because each individual can easily buy the same good, if cheaper, elsewhere at another store at no real cost (no travel cost). Then came the internet. With the introduction of the internet a new dimension (other than location and price) of consumption was created. On the internet, distance can be understood as an online engine search result page, in which a higher placed product is equivalent to a lower physical distance. Thus the physical location of the online product is indecisive. The internet also has another influence on the way consumption is done. Through online consumer's data collection, consumers are exploited by predicting their future consumption habits and accordingly online suppliers can change the price of a good for a certain individual or area. Thus people do no longer pay the same price for the same good at the same location.

The internet has raised many questions about ethical pricing. Net neutrality for instance is one of the main agenda points for the anticipated 2018 mid-term elections in the United States. For clarification "Net neutrality is the principle that all Internet Service Providers (ISP) treat all content equally and not give preference to some digital content providers" (Jacobson, 2017). Hence, the ISP are not allowed to discriminate between content and therefore cannot price the internet freely in the way other deregulated markets function. As touched upon previously, this is not the only discussed discrimination problem in the world of the internet. Price discrimination in the form of dynamic pricing is a fast-growing price strategy used by companies operating online (The Economist, 2016). In the 1990's, the Coca-Cola Company unsuccessfully tried to materialize dynamic prices for their vending machines, which allowed them to vary the price of a coke depending on the outside temperature (The New York Times, 1999 and 2005). Online dynamic pricing involves the tailoring of prices for each individual online-consumer based on their online

profile, which includes: preferences, purchases and search history. Dynamic pricing often occurs in the form of mark-ups, targeted discounts and sales of online goods. A study conducted by the Wall Street Journal in 2012, found that the proximity of the shopper to a Staples (office supply store) rival, negatively affected the online prices on staples.com (Valentino-DeVries et. al., 2012, The Wall Street Journal). To effectively implement dynamic pricing, retailers use software to optimize and customize even difficult market dynamics, such as when to use mark-ups, discounts, substitution and complementary goods (Associated Press 2007). In this empirical study the aim is to investigate how the internet activity correlate with the difference between off- and online prices of retailers. This paper is determined to answer the following central research question: Does the internet activity within a zip-code correlate with the difference between off- and online prices of multi-channel retailers in the US? Additionally, this paper will determine whether increased competition or dynamic pricing, which both arise with the introduction of the e-commerce, might explain this correlation.

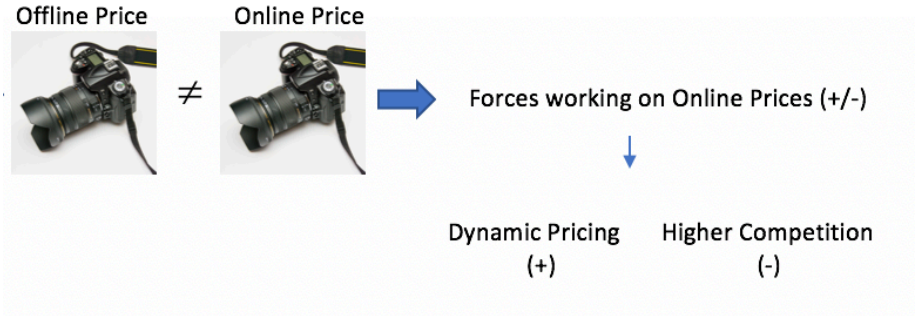
Multi-channel retailing, as defined by BIG-commerce, "is an ecommerce selling strategy that targets customers on various channels" (BIG-Commerce), in this paper being: off- and online, in other words, in the retailer's physical store and on the retailer's website, respectively.

Since this paper aims to understand what the determinants of off- and online price differences are, the assumed drivers of these differences must be addressed, namely dynamic pricing and higher competition. The internet retailers are assumed to function as a competitor to offline retailers (physical stores). Dynamic pricing as previously mentioned, is a process in which the prices are customized depending on certain characteristics attributable to the customer and is a possible reason for a difference since it is able to customize prices. For dynamic pricing to function it requires data. This data can only be obtained if an individual is online, hence the logic: the more an individual is online, the more data there is on that individual, thus it is assumed that the more internet activity the more effective dynamic pricing. Dynamic pricing by tailoring prices should lead to an increase of online prices. This will be elaborated in the theoretical framework. On the other side, there is the increased competition that e-commerce brings to the retail sector and more specifically to the online retail sector. Higher internet activity leads to higher online retail competition, which as economic intuition indicates should lead to a decrease in prices. Thus, two

counteracting forces acting upon online prices. These forces' common driver is internet activity. This leads to the following question: how do we measure online activity? Due to the lack of localized online activity data, this paper uses a proxy. The proxy used in this paper is the "supply" of internet. The reasoning for this is as follows: Internet activity is understood as the demand for internet connection and broadband (internet) providers as the suppliers of internet connection. Since in a perfect market equilibrium it is assumed that supply equals demand, one can assume that there are more competitors of broadband providers in areas with higher internet activity. Hence this paper will assume that the number of broadband providers, proxy internet activity of consumers, believing to be able to find a relationship between the number of broadband providers and the difference between offline and online prices.

As the graph below (1.1) depicts what this paper investigates, namely the reasoning of why online and offline prices of the same good (here exemplified by a camera) differ. The two forces assumed to drive this difference, which is captured by the proxy variable Internet Service Providers (ISP), are Dynamic Pricing and Higher Competition. Below, each force and the direction they affect the online price of the good is indicated (the camera symbolizes a generic good).

Graph 1.1 (Colombo, 2018)



Internet and its influence on the Retail Sector

In 1989 Sir Tim Berners-Lee invented the internet and initiated the third industrial revolution (The Economist, 2012). One of the biggest industries it revolutionized is the retail industry of which e-commerce in 2017 accounted for nine percent. This translates to a total of $ 453.5 Billion in the

US (US Department of Commerce). The growth of e-commerce is a global phenomenon, increasing by 23% in 2017, attaining a total market cap of $2.3 Trillion (International Post Corporation). Thus, the understanding of the effects of the internet on retail sales and the ever-growing percentage of e-commerce in retail is of essence for the understanding of economics in the future. The idea that internet activity is a determinant for prices is essential for grasping how users can influence the price of a good and for understanding how the price systems, currently at hand, are tending towards customized pricing.

To understand the dynamics and the change that the internet has initiated in the retail sector one has to consider the two parties in retail, these being the buyers and suppliers of retail goods. The internet has affected these two parties differently. The internet enables increased price transparency: buyers can choose between more suppliers of substitutes of the same good. It also decreases the search costs leading to a decrease in asymmetric information between buyers and suppliers (Coffinet and Perillaud, 2017). Suppliers, on the other hand, face tougher competition but can also exercise higher market power through price discrimination and dynamic pricing (Goldstein and O'Connor, 2000).

Due to the fact that this paper studies the difference between off- and online prices charged by multi-channel firms between certain zip codes, variables that affect all zip codes equally do not influence this difference, hence it is assumed that aggregate demand or supply shocks in the US can be ignored.

Additionally, this paper tries, as stated in the central research question, to find a correlation and not a causal relationship between internet activity and the difference between off- and online prices, using cross-sectional data. This is due to the limitation that it can only be speculated if the relationship is indeed a causal between the independent and dependent variable. Also, since the paper uses cross-sectional data, it is unable to control for differences in time, as time-series analysis does. Moreover, even if it can be assumed to be a causal relationship, the question remains whether there may be reverse causality, where the dependent variable causes the independent variable and vice versa. Hence this paper, in the conclusion and limitations sections will again touch upon these limitations and consider them when interpreting the results and their implications.

The paper, excluding the appendix, is structured in seven parts. In the next part the literature is presented. This is followed by an explanation of the data used for the research. Part four focuses on the paper's methodology, which elaborates on the methods used to answer the hypothesis. Results are presented in part five, whereas one may expect the results are presented and are tested for robustness. This is followed by the answering of the hypothesis and therefore the central research question in the conclusion (part six). Lastly, in part seven, limitations and recommendations of this paper are presented.

2. Theoretical Framework

For a thorough understanding of the topics touched upon in the introduction an eclectic analysis of existing literature on the various economic topics concerning the research question is done.

In theory, or as put by Andrea Goldstein in her presentation at the OECD Development center in 2001, e-commerce: "makes the whole economic system nationally and internationally more competitive- buyers can shop for the best deal over a wide geographic area – sellers can reach a large group of buyers" (OECD, 2000), as also the paper by Trainer (2016) reaffirmed. In addition, it was added that the internet enables and facilitates dynamic pricing in the form of price discrimination. This is done through the: "use (of) information about consumer buying habits to identify those willing to pay higher prices and take advantage of the fact that higher income consumers, i.e. those with a greater ability to pay higher prices, place a higher value on time" (OECD 2000; reaffirmed by Trainer, 2016)

To establish the groundwork for the theoretical framework this paper must again investigate the reasons for why the price gaps between off- and online prices could vary between zip codes. One of the reasons could, as mentioned, lie in dynamic pricing. Amazon charged different consumers different prices via dynamic pricing (Weiss and Mehrotra, 2001), using characteristics variables such as location and how much a consumer spent on past purchases, thus enabling the price to reflect the consumers' willingness to pay (Weiss and Mehrotra, 2001). Technology has enabled retailers to target consumers in online markets effectively and to measure precisely the results of

their pricing scheme (Grewal et al., 2011). This is possible due to the exponentially increasing amount of data and using more powerful software (Grewal et. al., 2011). If applied correctly, this most importantly increases profit by maximizing the capture of the consumer surplus (Elmaghraby & Keskinocak, 2003 and Sahay, 2007). It must be mentioned that there are limits to its ability to identify consumers' elasticities and demand functions, which can even result in revenue loss due to uncertainty (Besbes, 2009). Dynamic pricing itself has also been empirically challenged by papers such as Cavallo (2017), who claims that at least on the aggregate level the offline and online prices show "no evidence of dynamic price strategies that could potentially cause online-offline differences" (Cavallo, 2017 P. 285). This finding that dynamic pricing does not cause offline and online pricing differences, evidently means that dynamic pricing on an aggregate level seems to be neutralized, but this leaves open if there may be some form of dynamic pricing in a particular place (e.g. zip code area).

Since the introduction of the internet, there has been a continuous debate regarding the influence it has on competition and the difference between on- and offline pricing of goods. The internet increases competition, as laid down by Bakos (1997): the internet will notoriously decrease search costs and ease the match between sellers and buyers. Alternatively, as Brown and Goolsbee (2002) convey in their papers, finding evidence of the decrease of both off- and online prices due to the introduction of the internet for insurance companies. An additional paper found that: "increased product variety made available through electronic markets [...] increased competition significantly enhancing consumer surplus" (Brynjolfsson and Smith, 2003 P. 1580), thus, contradicting the previously mentioned evidence of lowering consumer surplus due to dynamic pricing. When examining Marshallian cross, the two aforementioned counteracting forces act as follows: an increase in consumer surplus must be a result of a decrease of prices and a decrease in consumer surplus due to dynamic pricing increases prices, on aggregate (see below, Figure 1 and 2). Taking into consideration the two counteracting forces that are introduced with e-commerce, the question remains which of the two is the dominant force. This paper will answer this question. Hence, for now, the difference between off- and online prices, not sure which of the two forces dominates, is ambiguously affected by higher/lower internet activity in a zip code area.

Figure 1: Dynamic pricing enables the retailers to identify the demand curve of the consumer. Thus the price is on average increases (can also work the other way around if the algorithm determines that the consumer is only willing to buy the product for a discount). In general, the consumer surplus decreases. Note that this is the aggregate effect of dynamic pricing, the effect is based on each consumer's demand curve being identified and tailoring the offered price (the supply curve). Since the demand curve is identified, the demand curve shifts from D1 to D2 (here assuming that the individual is willing to pay more for the good). The supply curve shifts out for there to be an equilibrium in the market.

Figure 2: depicts the effects of increased competition. Supply curve shifts from S1 to S2 (increased competition) the price decreases and the consumer surplus increases.

Figure 1. (Colombo, 2018)

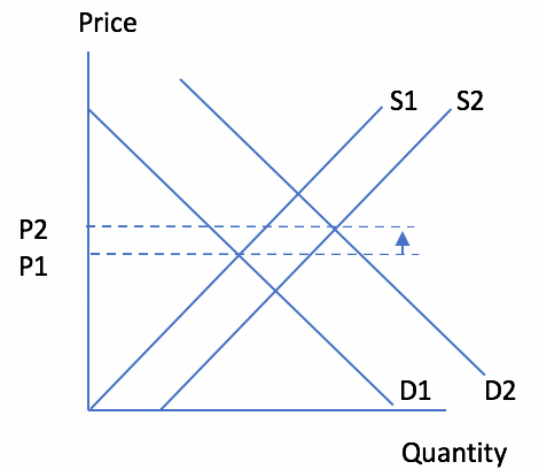
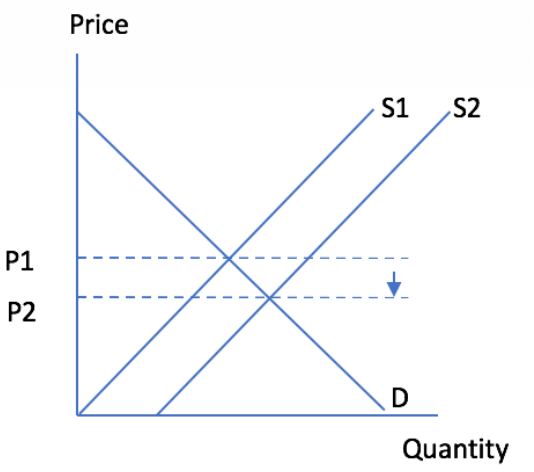


Figure 2. (Colombo, 2018)



As explained in the introduction, internet activity is proxied by ISP. It is assumed that the proxy variable both embodies the force exerted by dynamic pricing and higher competition, thus the sign of the correlation's coefficient is of importance. For the following explanation it is assumed that offline prices are on average higher than those of online prices of a good. A negative coefficient, it indicates that ISP exerts negative pressure on the difference between off- and online prices and therefore the dominant force is dynamic pricing. The contrary holds for higher competition (positive coefficient).

This paper's first hypothesis is:

$H0_a$: The price gap between off- and online goods between zip codes correlate with the number of internet service providers and the amount of internet coverage.

$H1_a$: The price gap between off- and online goods between zip codes does not correlate with the number of internet service providers and amount of internet coverage within a zip code area.

Having laid out the first hypothesis, it is now determined whether these two forces are indeed determinants for the negative (dynamic pricing) and positive (higher competition) differences between off- and online prices. For clarification, positive differences are when online prices are lower than offline prices and vice versa for negative differences. If variables that influence the offline prices are accounted for, then the investigating of positive differences between offline and online prices should correlate with higher competition. Thus, for positive differences, the sign of the ISP competition is expected to be positive, since higher competition increases the price difference between offline and online prices, by exerting downward pressure on online prices. The second hypothesis, where again the internet activity level is proxied by ISP competition within a zip code area, is:

$H0_b$: Positive price differences between offline and online prices positively correlates with the number of internet service providers and the amount of internet coverage in the direction of higher competition.

$H1_b$: Positive price differences do not correlate with the number of internet service providers and the amount of internet coverage in the direction of higher competition.

When online prices are higher than that of offline prices of the same good it is referred to as negative price differences. Here other than for the previous two hypothesis the sign of ISP that indicates dynamic pricing is a negative coefficient. The reasoning is that of the data where the

online price is higher than the price offline, dynamic pricing exerts upward pressure on online prices and therefore the price difference increases (the negative price difference increases).

The third hypothesis:

$H0_c$: Negative price differences between offline and online prices negatively correlate with the number of internet service providers and the amount of internet coverage in the direction of dynamic pricing.

$H1_c$: Negative price differences do not correlate with the number of internet service providers and the amount of internet coverage in the direction of dynamic pricing.

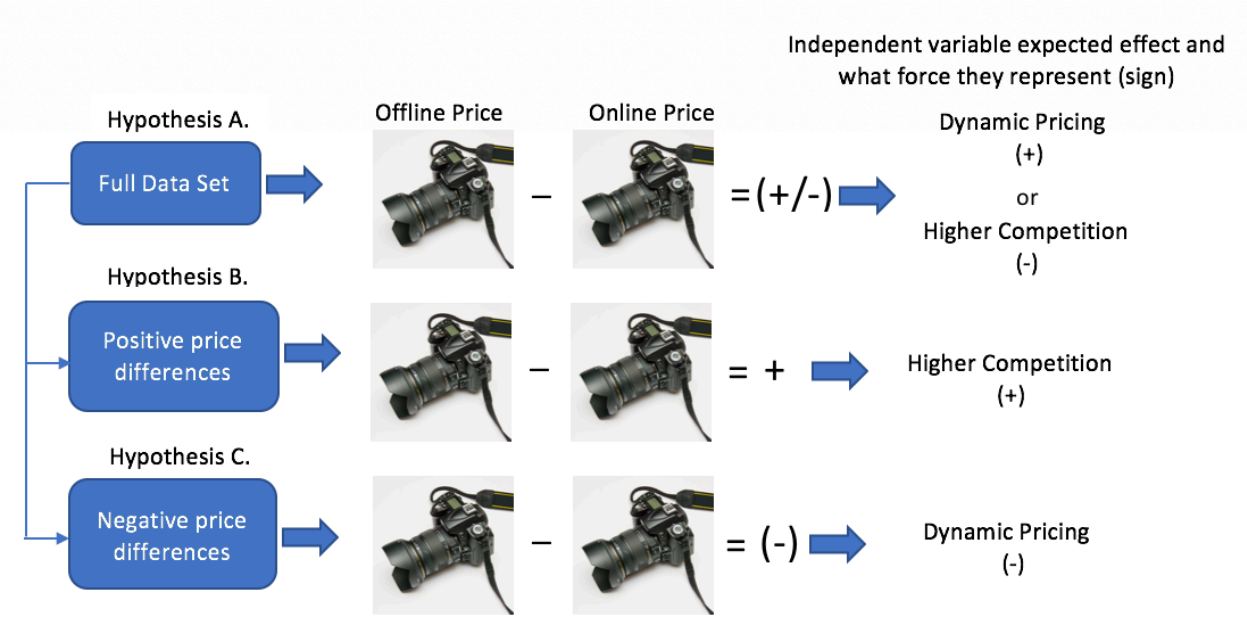
The table below (2.1) explains what the three above mentioned hypotheses investigate and the different coefficients that determine whether to accept or reject the hypothesis.

Table 2.1

Dataset	What ISP competition proxies (independent variable)	Sign of independent variable if force is exerted and the hypothesis is accept if:
Full dataset	Dynamic pricing or Higher competition	(-/+)
Positive Price difference	Higher competition	(+)
Negative Price difference	Dynamic pricing	(-)

Furthermore, a graph (2.2) is created to understand the differences between the hypothesis (the camera symbolizes a generic good).

Graph 2.2 (Colombo, 2018)



Although not part of the empirical research of this paper, this paper will briefly take the time to touch upon the ethics behind dynamic pricing, since it is of utmost importance for the researcher to understand the implications of their results to then give further recommendations. From an economic standpoint, equilibrium is reached when demand equals supply. Thus it makes sense to customize prices to the smallest unit possible, that being a single household or individual, as done by dynamic pricing. The obvious argument against dynamic pricing is the fact that different prices are offered to different consumer for the same good. Research has investigated the sociological effect of this type of discrimination: consumers lose benevolent trust in firms that implement dynamic pricing strategies (Garbarino & Maxwell, 2010). Additionally, dynamic pricing or pricing consumers differently, is perceived as unfair (Haws & Bearden, 2006, and Chapuis. 2012) and in general, there is a negative perception of dynamic pricing by customers as put forward by Weisstein et al. (2013). The idea that in the future every aspect of an individual may affect the price of a good, exhibits an atmosphere of risk and the accompanied risk-averse anxiety over how

behavioral characteristics change prices. It must be noted that societies may benefit from dynamic pricing in the sense of reducing over-consumption. As the paper by Farugui and George (2005) shows, dynamic pricing in the energy market reduces overall consumption, thus also resulting in a positive environmental externality. The consumer sentiment and industries' rationale mentioned here will be taken into consideration in the interpretation of results and concluding remarks of this paper.

The data used in this paper originate from Cavallo (2017), hence relevant findings and differences in this paper's analysis are now presented. Cavallo (2017) notably found that 69% of on- and off-line prices from multi-channel retailers in the US are equal (different prices for the remaining 31% of goods). This paper differs from Cavallo's paper in two ways. Firstly, this research is purely aimed at finding if the differences in prices is determined on a zip code area basis. The second difference is that the analysis is aimed at determining whether internet activity is an explanatory variable for price differences.

The investigation of this paper consists, as said, in the identification of the online price discrimination between zip code areas. Since this paper uses data on various goods that are not substitutes, an online price difference analysis to see which of the two forces is dominant cannot be done simply by looking at the differences in online prices across zip codes. To solve this the paper uses the off- and online price difference of the same good within a zip code area, taking into account factors that can affect the offline price of goods (explained in the following section of this paper). When taking these factors into account there should be no difference in prices of offline retailers between zip codes. Thus, the difference in online prices between different zip code areas can be observed and therefore the impact of dynamic pricing or higher competition. The two types of price discrimination this paper researches are shown below.

Price discrimination 1. Product $_i$ Online (zip code $_i$) - Product $_i$ Online (zip code $_j$)
Price discrimination 2. Product $_i$ Offline (zip code $_i$) - Product $_i$ Online (zip code $_i$)

Price discrimination 1 cannot be identified with the data used, but with the help of the difference in on- and off-line prices (price discrimination 2) this difference can be identified.

3. Data

In order to examine the hypothesis, data is retrieved from various sources. The main source of this paper's data is from the MIT led "Billion Prices Project" and more precisely the data used for the study by Cavallo (2017). Demographic and socio-economic data for the researched zip code areas is retrieved from the United States Census Bureau, the principal agency of the US Federal Statistical System, whose responsibility lies in collecting data on the population and the economy. The residential speed, coverage and number of providers of broadband connection within a zip-code area in the US in 2016 is provided by broadbandnow, an internet platform that collects data on broadband connectivity within a zip code area. The focus in this paper lies in residential internet connection since it represents around 85% of e-commerce in 2015 (Meola, 2017). It is, therefore, reasonable to assume that residential internet activity is the predominant (lone) force in retail e-commerce. Proximity One, another website, is used to find the number of all year-round retailers within a zip code area, representing the offline retail competition for the researched zip codes. All data is of the year 2016, except for the following data: population data (2010), since the national population collection is done on a decennial basis, Proximity One data is from 2017 and Cavallo's (2017) data on the actual goods comprises the years 2014-2016.

This paper assumes that the differences in years for the data of retail goods can be ignored, given the low retail inflation rate of less than < 2 percent (Trading Economics, 2018). This even after taking into account that the inflation rate for e-commerce is on average lower than general CPI, including that of physical stores (offline). Evidence of this is presented by the paper by Goolsbee and Klenow (2018), which shows by comparing online inflation to general CPI that online inflation is 1.3 percent lower per year. Since this difference is very small, this paper ignores the annual inflation rate difference of off- and online prices. Additionally, since in the US there is a positive yearly population growth of 0.7 percent (World Population Review. 2018) and one can assume that the population growth is homogenous across all the investigated zip codes, this paper ignores that the year of the data on population (2010) does not match the rest of the data.

The data from Cavallo (2017) has been made available to the public for academic purposes. It contains an unprecedented total of over forty-thousand product prices of goods from multi-channel retailers from December 2014 to March 2016. The data was collected using the top 20 retailers by market share (Cavallo, 2017). The name of all the retailers used in the US by the paper can be found in Appendix A, Table 3.4. The data on offline prices of goods was obtained via crowdsourcing platforms in which individuals were sent to the physical store to scan a good's price. Online data instead was collected via software given to the individual conducting the offline data collection to then determine the online price of the same product, hence off- and online prices were collected in the same respective zip code area. The data related to goods of various retail sectors including food, clothing, household, drugstores, electronic, office and multiple or a mix of the former sectors mentioned. All off- and online goods' prices used in this paper are in units of US Dollars. It must be said that the results of this paper are limited to the data that has been retrieved from the Cavallo paper.

The preliminary criteria for the chosen data is that there is an off- and online price available for the same product and the availability of a US zip-code for the product. Thus, the remaining dataset consists of 19796 products. This full set of data will later be used for testing for robustness. Of the 19796 products, 7982 (~40%) have different off- and online prices. Of which 6407 originate from the state of Massachusetts, therefore making the analysis biased towards finding results that are representative for the state rather than the whole country. Nevertheless, the paper will continue with the analysis, assuming that the results are representable for the whole of the US. Since this paper wants to identify the determinants of the differences between off- and online prices, the paper will not analyze the remaining 11814 (19796 - 7982) goods. It must also be said of the remaining 7982 observations that there is a large discrepancy between the number of observations per zip code ranging from 4 to 1001 observations, averaging at 99 observations per zip code that later is addressed in the chapter Data Transformation.

Variables Used

To see a summarized version of the origin and description of the variables used in this paper see the appendix (Appendix A, Table 3.1).

The dependent variable for this thesis is the percentage difference between off- and online prices, here the online price of a good is subtracted from the offline price of the same good and then divided by the online price of the good ((offline price – online price) / online price). This is done because the data entails a vast amount of different goods, thus the prices of the goods vastly differ (offline from 0.63 to 1,199.63 US dollars and online from 0.4 to 999.99 US dollars). This makes the absolute difference between offline and online prices of the goods of little use for further interpretation. The percentage difference on the other hand gives a better picture of the aggregate difference of on- and off-line prices. Factors, as mentioned in the introduction, that are assumed to equally affect the prices of goods off- and online are ignored (supply- demand-shocks, inflation etc.). The independent variable is total internet coverage in a zip code. Here the paper distinguishes between two possible determinants for internet coverage to increase the probability of being able to find a correlation to the dependent variable, being either internet service providers (ISP) competition (*numberISPcompetition*) or total ISP coverage (*totalISPcoverage*) within a zip code, explained further on in this section. The following control variables are added to control for omitted variable bias: Education, henceforth as *education* (high school graduate or higher), Median Household income, variable name *MedianHHIncome* (2016), number of all year-round retailers (*allyearretailers*) within a zip code area and demographics (*median age* and *population)*. *Median age* measures the median age of the population in a zip code area and *population* represents the size of the population within a zip code area.

The control variables are chosen based on what is presumed to be explanatory power on how differences in pricing of off- and online goods vary between zip code areas and are based on papers and economic intuition. It has been found that income and education indicate the accessibility of router-based internet connection (Chaudhuri et. al.. 2005) and internet activity (Porter and Donthu, 2006). Additionally, economic intuition suggests that these variables (*education* and *MedianHHIncome*) affect the prices of offline and online goods since in areas of higher income goods are more likely to be more expensive (online: through dynamic pricing) and education is a good indicator for higher income. Furthermore Cheung & Liao (2001) established that internet speed is not decisive for individuals when considering the use of e-commerce, hence the paper does not use internet speed as a variable. Median age (*demographics*) within a zip code area has

an effect on the volume of e-commerce since age is a good indicator for internet activity (Porter and Donthu, 2006), therefore the consumption of offline goods is higher in areas of higher age (assuming consumption is constant over age), thus there is higher demand for offline goods. The variable *Allyearretailers* aims to capture the amount of offline competition within a zip code area, which thus determines the price dynamics of offline retailers. The economic intuition for using the variable *Population* is that it indicates the amount of potential off- and online shoppers and therefore affecting the total amount of competition between the two retail channels and the amount of internet activity in an area.

Descriptive statistics

Descriptive statistics give an initial impression of the variables used. For the main analysis of this paper descriptive statistics can be divided into four parts, Zip code based, full data, ISP and positive and negative (offline minus online) price differences. Since this paper investigates the price discrimination between zip codes, the statistical attributes such as the mean minimum and maximum observation are grouped by Zip Code area and not by observation. Secondly, the complete data of observation, disregarding the location of the observation, is analyzed to understand the dynamics of price differences. Here the other variables are not of interest since they do not vary by good (observation) but by zip code, hence for the descriptive statistic for zip code areas is satisfactory. Lastly, the data is separated into two parts which contain all observations of positive and negative differences between the prices of offline and online goods respectively. This is done to find the underlying strength of the two opposing forces: higher competition and dynamic pricing.

Of the 250 zip-code areas in the Cavallo's (2017) paper, 79 are used in the main part of this paper, selected based on a minimum of 4 differences in observations of off- and online prices. First and foremost, the location of the 79 zip codes in the US is investigated. The 79 are from 22 different states (out of 50 US states). The most originate from the states Massachusetts (17), California (12) and Virginia (8), averaging at ~3.59 zip codes per state, see Appendix A (Table 3.2). This again signifies that the analysis of the data may be most accurate for the three most frequent States and less for the whole US territory. Next, the zip-code areas attributes in terms of socio-economic

characteristics are compared to national statistics. One of these attributes is *education*, measured as the percentage of high school graduates or higher degrees in the zip code area. For the 79-sample zip-code areas education is three percent points higher than the national average (0.9 compared to the 0.87 national average). The median age is lower than the national average by about 1 year at 36.83 (national average 37.7). The most evident difference though lies in the median household income, which lies 35% above the national average at 74,591 US Dollars. These differences are important for the external validity of this paper (being able to argue that the paper's results hold for all US zip codes) and therefore the interpretation of results. A table of all the variables and their statistical attributes can be found in the table below.

Table 3.3 Control variables and their statistical attributes (grouped by Zip Code Area)

Variable	Obs	Mean	Std. Dev.	Min	Max
NumberofISPcompetitors	79	2.66	.79	1	5
TotalISPcoveragee	79	2.41	.48	1.50	4.26
Allyearretailers	79	140.73	90.95	5	472
Education	79	.90	.07	.72	1
Demographics	79	36.84	5.81	21.50	53.6
MedianHHincome	79	74,927.04	34,221.22	30,070	1862,25
Population	79	31,229.11	17,451.33	1,733	94,600

This paper studies prices of retail goods, hence the necessity of understanding their statistical characteristics. As can be seen in Table. 3.4 the mean price of offline prices is slightly higher than that of online prices, ~30.64 and ~29.83 US dollars respectively. Logically it must follow that on average the percentage difference between off- and online and the absolute difference between the two variables is not 0. It is in fact ~.13 percent and ~81 US dollar cents respectively. This indicates that on average, assuming that the only differences in prices is driven by dynamic pricing and higher competition, the prices online are lower than offline. Thus, it seems that competition is the stronger force of the two. The table below gives an overview of all price variables.

Table 3.4 Descriptive statistics of price and dependent variable (by observation)

Variable	Obs	Mean	Std. Dev.	Min	Max
Price Offline	7,982	30.64	53.58	.63	1,199.99
Price Online	7,982	29.83	54.65	.40	999.99
Percentage Difference of Offline and Online price	7,982	.13	.54	-.99	18.95
Offline-Online	7,982	.81	20.22	-315	300

Some information on Internet Service Providers (ISP). The zip codes used in this paper contain a total of 29 different ISP that on average cover a total of ~94 percent of each zip code area. Adding all the coverage of ISP per zip code gives us the total coverage (e.g. Verizon covers 80 percent of the area and AT&T covers 60 percent, total coverage equals to 160 percent or 1.6) and this averages at ~2.4 (240% coverage) per zip code. In the data used the 29 ISP provide an aggregate of 304 internet connection plans for the 79 zip codes (including from the same company), thus an average of ~3.85 different broadband plans per zip code. When controlling for providers that provide more than one type of internet plan (e.g. Verizon FIOS, Verizon High Speed Internet or Verizon) the average is ~2.66 (2.66 competitors), that being the average amount of ISP competition per zip code area. The most frequent internet provider is Verizon with a total of 85 representations in the 79 zip codes, followed by Xfinity and AT&T with 57 and 49 broadband connections respectively. Since Verizon only provides internet services in 50 of the selected data's zip codes, it is evident that it offers more than one broadband connection plan within a zip code area. Below the table with the statistical attributes of the independent variables (zip code based).

Table 3.5 Descriptive statistic Independent variables (grouped by Zip code)

Variable	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
NumberISPcompetitors	79	2.66	.79	1	5
TotalISPcoverage	79	2.41	.48	1.50	4.26

As previously touched upon, more than often there are numerous ISP in one zip code from the same company for example, Verizon FIOS, Verizon High Speed Internet or just Verizon, thus the variable *NumberISPcompetitors* is limited in the sense that it is unable to capture the real amount of internet provided in a zip code area. For instance, an ISP that only provides to 4 percent of the zip code is with this variable equally accounted for as an ISP that provides 97 percent of the zip code area. Hence the variable *TotalISPcoverage* has been created for this paper and consists of the sum of residential internet providers coverage within a zip code, including coverage from the same ISP, and therefore more accurately displays the amount the internet is used within the area given the assumption that internet coverage is the proxy for internet activity.

Positive and negative price differences

As explained in the theoretical framework, there are two opposing forces on online prices, namely increased competition (downward force on online prices) and dynamic prices (upward force on online prices). In the first part, and the main part, of this paper focuses on determining the aggregate effect of internet activity on the percentage difference of off- and online prices. Now the focus is on the two separately. Thus, trying to identify if the two forces are significant forces when the online prices are lower/higher than the offline prices. For this analysis the data is split up in two, one with positive and one with negative differences. In Appendix A the statistical characteristics of percentage difference in off- and online prices for each zip code can be found (Table. 3.6).

Positive differences, where the price online is lower than offline, consist of 5456 observations. This accounts for ~68 percent of all price differences. Online prices here average at a considerably lower than that of the full dataset at ~24.2 versus ~29.8 in the joint dataset. Since also the mean offline price is larger than that of the full dataset, logically the percentage difference of off- and online prices shows a higher mean difference. The statistical attributes of the rest of the price variables can be seen below in Table 3.7.

Table 3.7 Descriptive statistics of price variables for positive differences (by observation)

Variable	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
Price Offline	5,456	31.05	52.18	.86	1,199.99
Price Online	5,456	24.2	42.69	.40	999.99
Offline-Online	5,456	6.85	13.88	.003	300
Percentage difference	5,456	.30	.57	.00004	18.95

Negative differences where the price online is higher than offline contains the remaining 2526 observations (~32 percent) of the data set. As it can be observed below, as anticipated the online price mean is considerably higher than that of the median for the whole dataset (~42 versus ~29.8). The mean price offline is around the same as that of the complete data. Thus the percentage difference is lower at -24 percent in comparison to the entire data average of ~13 percent. Although there are fewer observations of negative price differences the percentage price differences are on average higher for negative observations than that of positive ones. The full price variables statistical attributes for the negative observations are found below in Table 3.8.

Table 3.8 Descriptive statistics of price variables for negative differences (by observation)

Variable	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
Price Offline	2,526	29.77	56.48	.63	949
Price Online	2,526	42	72.7	.99	989.99
Offline-Online	2,526	-12.24	25.05	-315	-.01
Percentage difference	2,526	-.24	.20	-.99	-.00007

Data Transformations

As with all raw data, a number of transformations need to be performed for its harmonization. The first thing that is checked is whether there are no extreme outliers that unbalance the distribution of the variables. To test the distribution of all numerical variables the Skewness and Kurtosis test for normality is used (see Appendix A, Table 3.10). No changes needed to be made, except for the variable *education* (here the natural logarithm is taken) thus all variables have normal distributions (p-value for variable Skewness and Kurtosis: 0.000). Since the distributions for all variables are all normal or have been normalized, OLS analysis can be used. Next, the relationship between the variables is investigated, namely the correlation. As it can be seen below (Table 3.11) there are at first glance no correlations that could result in multicollinearity, except for the two independent variables, which is expected, since they are two different ways of representing the ISP in a zip code area. Furthermore, when computing the variance inflation factor (see Appendix A, Table 3.12), that indicates the extent to which the standard error of the coefficient of interest (% *Offline-Online prices*) is inflated upwards by adding variables to the regression model, there appears to be no multicollinearity between variables in the regression (see Appendix A, Table 3.12, rule of thumb: variable has multicollinearity tendencies at values > 4). Since there is, as previously mentioned, such a large difference in observations between zip code areas (see Appendix A, Table 3.13) it could be advised to weight the observations so that observations of zip code areas with a low observation count are weighted higher than the observations of zip codes with higher observation counts. Here, a tradeoff between equal weighting of zip codes and thus unequal weighting of observations and the equal weighting of observations but unequal weighting of zip code areas, is faced. This paper chooses the latter, since the weighting and thus manipulation of the data could significantly change its results. In checking for robustness this will continue the discussion of weighting the data.

Table 3.11 Correlation of Variables

	Offline-Online Percentage Difference	*Education*	*Demographics*	*MedianHH Income*	*Population*	*Number of competing ISP*	*Totalcoverage ISP*	*Allyear retailers*

Offline-Online Percentage Difference	1							
Education	0.04	1						
Demographics	0.15	0.13	1					
MedianHH Income	0.05	0.53	0.66	1				
Population	-0.04	-0.13	-0.08	-0.08	1			
Number of competing ISP	0.15	0.34	0.32	0.12	0.29	1		
Totalcoverage ISP	0.10	0.05	0.49	0.12	0.20	0.76	1	
Allyear retailers	-0.05	0.07	0.16	0.11	0.54	0.14	0.14	1

Expected sign of variables

The expected sign of the variables on the difference of off- and online prices can be seen in Appendix A (table 3.14). The most important expected signs are that of the number of competing ISP and total coverage by ISP which are both ambiguous because they are both expected to be positive (higher competition) or negative (dynamic pricing) depending which of the forces is dominant. Depending on which of these forces prevails the price difference will be larger (higher competition) or lower (dynamic pricing). Interestingly, the two variables, as can be seen above in Table 3.11, have a positive correlation with the variable *Offline-Online Percentage Difference*, which indicates that the competition effect is larger. Furthermore, the signs of the other variables are in part determined by economic intuition or papers. *Allyearretailers'*, expected sign is negative since more offline competition leads to lowering their prices. *Population*'s expected sign is positive following its positive correlation with Offline-online prices. This research paper adopts the sign that the paper by Porter and Donthu (2006) found for *MedianHH Income* (+), *Demographics* (-) and *education* (+). The expected ambiguous sign does not hold for the independent variable when splitting the dataset into two parts, namely positive and negative

differences. For positive price differences the expected sign of the independent variables is positive since, as mentioned, it is assumed that the positive price differences are driven by higher competition in the online market, thus the positive difference is strengthened by an increase in the independent variable. To clarify: when (offline-online)/(online) = positive and therefore offline prices > online prices, this difference should be increasing with the independent variable. Here the independent variable is expected to be positive due to higher competition dominating the effect of dynamic pricing. Likewise, when (offline-online)/(online) = negative (offline prices < online prices), this difference should be increasing with the independent variable, which in this case is the dominance of dynamic pricing. Thus, the independent variable is expected to be negative.

Note here that for the analysis only the correlation between the independent and control variables with the dependent variable can be found with certainty. Given the data used, finding casual relationships is of speculative nature but can be assumed for variables where it is certain they are not subject to reverse causality.

Data for robustness

Henceforth the full dataset refers to Cavallo (2017) 19796 observations and the paper's dataset to the 7982 observations, respectively.

For the second part of the investigation, the full data of Cavallo (2017) is used, which includes observations of goods that have no price difference. A brief summary of the descriptive statistics and transformations done to harmonize the data, follows. The average competing ISP, within a zip code area, in the full dataset is ~2.76 (papers dataset ~2.65) and the total ISP coverage ~2.36 which is lower than that of the papers main data set (2.4). This signals that although there is on average more competition in Cavallo's full data set, the coverage of the papers dataset is higher. When observing differences between the two datasets the most obvious is the large difference of median house hold income, which in the full dataset is 65,434.09 US Dollars (17% higher than the national average) in comparison to the previous 74,927.03 US Dollars. The average population within the zip codes is higher than that of the paper's dataset at 35,505 (paper's dataset: 31,229). These

differences indicate that the Cavallo's full data has indeed distinct identities. This is formally tested in the methodology section with a difference by difference test. The full descriptive statistics report of variables can be found below in Table 3.15. Data transformations made to this data set mirror the ones of the paper's data set. Hence, the logarithm of *education* is taken. Furthermore, to give more weight to underrepresented zip code areas the function i.weight based on weight of the observation of the corresponding zip code area will be added to the regression, this will be further explained in the methodology section.

Table 3.15 Robustness: Variables and their statistical attributes (grouped by zip code area)

Variable	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
NumberofISPcompetitors	250	2.77	.92	1	6
totalISPcoveragee	250	2.36	.47	1	4.26
Allyearretailers	250	131.22	86.19	3	472
Education	250	.90	.067	.67	1
Demographics	250	37.32	5.74	20.1	58.9
MedianHHincome	250	65,434.09	26,712.26	24,610	186,225
Population	250	34,102	16,822	863	94,600

Having swept through the data and having prepared it for its analysis one can now proceed to the section that explains what statistical procedures that are undertaken to answer the hypotheses and hence the central research question, namely the methodology section.

4. Methodology

The empirical part of this paper consists of two parts. In the first the two forces (higher competition and dynamic pricing) acting in opposite directions are analyzed to determine how much of any price difference, negative or positive, can be explained by the two forces. This aims to answer hypotheses B&C (regressions 1.1-1.3 and 2.1-2.3). This will be followed by the combined analysis of the joint data to determine whether and in what direction price discrimination exists due to internet activity (regression 3.1-3.3). Hence, the regressions below are done three times each with the three different datasets (positive, negative and full). All data analysis is done with OLS regressions using the program Stata.

In order to approach hypothesis: The price gap between off- and online goods between zip codes correlate with the number of internet service providers and the amount of internet coverage, is examined. To capture this effect the following regression is used:

$$(1.1)\ Difference\ of\ On\text{- and offline Prices}_{i\,z} = \beta_0 + \beta_1\, TotalcoverageISP_z + \beta_2\, MedianHHIncome_z + \beta_3\, Population_z + \beta_4\, education_z + \beta_5\, Demographics_z + \beta_6\, AllyearRetailers_z + \varepsilon_0$$

Where $_{i,\,z}$ are the corresponding goods and zip-code area in which the good was observed, respectively. Similar to the first regression the second regression aims to see if the effect of the number of competing ISP differs from that of total coverage ISP.

$$(1.2)\ Difference\ of\ On\text{- and offline Prices}_{i\,z} = \beta_0 + \beta_1\, CompetitionISP_z + \beta_2\, MedianHHIncome_z + \beta_3\, Population_z + \beta_4\, education_z + \beta_5\, Demographics_z + \beta_6\, AllyearRetailers_z + \varepsilon_0$$

To additionally check for omitted variable bias, *CompetitionISP* and *TotalcoverageISP* are both integrated into the third regression:

$$(1.3)\ Difference\ of\ On\text{- and offline Prices}_{i\,z} = \beta_0 + \beta_1\, TotalcoverageISP_z + \beta_2\, CompetitionISP_z + \beta_3\, MedianHHIncome_z + \beta_4\, Population_z + \beta_5\, education_z + \beta_6\, Demographics_z + \beta_7\, AllyearRetailers_z + \varepsilon_0$$

Testing for robustness

For the robustness tests, the complete data set of Cavallo's (2017) is used and the same regressions are performed, thus not only using data in which the price differs but also the 11,814 observations in which the off- and online prices are equal. It is essential to identify whether there is a difference in the additional dataset for each of the variables used in the previous regressions. To solve this

conundrum a difference in difference test is done. As it can be seen below in Table 4.1 all the variables mean between the treatment (there is a difference between off- and online price) and the non-treatment group (no price difference between on- and off-line price) are significantly different, thus further investigation of the additional dataset is reasonable. The second part of the investigation for robustness entails, as touched upon in the Data section that observations can be weighted such that zip code areas with lower counts still have significant representation in the final regression. Therefore the *i.weight* function that enables the weighting of an observation depending on other variables is used. Hence, the *weight* variable is created which equals to 1/ number of observations per zip code area, meaning if there are 100 observations for a zip code the observation will be weighted as 1/100 and therefore each observation as 0.01. In this way, each zip code is equally taken into account in the regressions output.

Table 4.1:
Robustness Treatment and non-treatment groups (difference in difference test)

Variable	*Mean treatment ZIP-CODE*	*Mean non-treatment ZIP-CODE*	*Difference significant (Stata regression, P-value: < .05)*
Median Income	74927.037	60800.63	Yes
Demographics	36.8	37.6	Yes
Population	31229	35505	Yes
Education	0.90	0.89	Yes
Total ISP coverage	2.41	2.34	Yes
# of ISP competitors	2.66	2.82	Yes
All year retailers	140.73	126.58	Yes

To investigate robustness the regressions 1.1 - 1.3 are rerun using the full sample of the data. The regressions for testing robustness will be named 4.1-4.3 and 5.1-5.3 for the weighted robustness test, as one can see below.

$$\textit{(4.1) Difference of On- and offline Prices}_{iz} = \beta_0 + \beta_1 TotalcoverageISP_z + \beta_2 MedianHHIncome_z + \beta_3 Population_z + \beta_4 education_z + \beta_5 Demographics_z + \beta_6 AllyearRetailers_z + \varepsilon_0$$

$$\textit{(4.2) Difference of On- and offline Prices}_{iz} = \beta_0 + \beta_1 CompetitionISP_z + \beta_2 MedianHHIncome_z + \beta_3 Population_z + \beta_4 education_z + \beta_5 Demographics_z + \beta_6 AllyearRetailers_z + \varepsilon_0$$

Again, the third regression includes both *TotalcoverageISP* and *CompetitionISP*:

$$\textit{(4.3) Difference of On- and offline Prices}_{iz} = \beta_0 + \beta_1 TotalcoverageISP_z + \beta_2 CompetitionISP_z + \beta_3 MedianHHIncome_z + \beta_4 Population_z + \beta_5 education_z + \beta_6 Demographics_z + \beta_7 AllyearRetailers_z + \varepsilon_0$$

Since the framework in which the hypothesis will be tested has now been laid down, it is now proceeded to the results part of this paper.

5. Results

The results section aims to present the analysis of the data. First, the positive and negative differences are analyzed. Additionally, since dynamic pricing schemes produce an individual price according to specific variables, with the results one can speculate which variables are used in these pricing schemes. Naturally, the results are presented chronologically starting with regression 1.1. The complete output of the regressions (1.1-3.3) can be seen in Table 5.1 above the results from testing for robustness. All significant coefficients are interpreted, keeping in mind that the dependent variable *percentagedifferenceon-off-lineprice* is in percentages, thus 1 unit increase in a variable's coefficient is equivalent to a 1 percent increase in price difference.

As mentioned in the Data section the results indicate with certainty a correlation of the variable with the independent variable. Causal relationships are possible but cannot be assumed.

Results positive differences (offline prices higher than online) regressions:

For the regressions of observations with a positive difference, all variables with a positive coefficient increase the difference of prices and naturally negative coefficients decrease the difference hence act in the opposite direction of the higher competition force. Regression 1.1 finds that *totalISPcoverage*, as expected, increases the percentage difference between offline and online prices with a coefficient of ~.023 but is insignificant at a 10 percent level thus the coefficient cannot be further interpreted. Control variables that are significant at a 1 percent level include *MedianHHIncome, Demographics* and *Education*. Starting with *MedianHHIncome*, which has a negative coefficient, contrary to the expectations. Although the coefficient is small (-.0000028) median households income averages at 74,927 US Dollars (effect equals .21 percent decrease in the difference of prices) its effect on the difference is thus significant. The rationale behind this finding may be the following: in higher median income zip code areas the effect of dynamic pricing is greater than that of higher competition. This the increases online prices, thus lowers the gap between off and online prices. The variable *demographics* has a positive coefficient, not as expected, of ~.012, which is low (average effect. ~.44 percent increase) but has a higher effect than that of *MedianHHIncome*. The reason for this result is speculated to be that: the higher the median age, the lower the use of internet, thus less data are available on individuals and hence higher differences between offline and online pricing because of inefficient dynamic pricing.

Education here positively influences the difference, as expected, with a coefficient of ~.35. The coefficient at ~.35 needs further clarification considering that the *education* variable is expressed in a percentage and has been transformed to its natural logarithm. Thus an increase of one percentage point of High School Diploma or higher education results in a 0.01 increase in the variable *education*, before the natural logarithm transformation. Furthermore, 1 unit increase of *education* (100%) results in an $e^{(1)}$ (~2.71) increase in the percentage difference of off- and online prices. The median of the variable *education* is ~.9 which is equal to $e^{(.9)}$ which is equal to a ~2.46 percent increase in the difference between off and online prices. The interpretation of this result is straightforward: a higher education results in higher use of the internet which increases online competition and hence increases the positive difference between off- and online prices. All other variables are insignificant at a 10 percent significance level, no further interpretation is necessary.

In Regression 1.2 the independent variable *numberISPcompetitors* is again positive, as expected, but likewise insignificant at a 10 percent level. Here the variables that are significant at a 1 percent level are *MedianHHIncome* and *Demographics*. The coefficients of the two variables mirror the ones of regression 1.1 thus the elaboration of their interpretation is unnecessary. *Allyearretailers* is significant at a 10 percent level and has a coefficient of ~.0003 (negative effect expected). Therefore, *Allyearretailers* is not very significant in terms of absolute value. The interpretation of the variable is nevertheless important. The positive coefficient is against this paper's expectations, and the reasoning can only be assumed to be that online retailers decrease their prices slightly more than the offline retailers do, due to higher competition. Equally, when the regressions include both independent variables (regression 1.3) similar results are obtained. Both independent variables are insignificant. *TotalISPcoverage* is interestingly negative and *numberISPcompetitors* is marginally more positive, this indicates that *TotalISPcoverage* entailed some of *numberISPcompetitors* explanatory power in regression 1.1 and vice versa in regression 1.2. The same variables in regression 1.2 are significant at the same levels and have alike coefficients, hence again there is no need to repeat the interpretation of the variables. The adjusted r-squared suggests that the explanatory power for the three regressions is limited with around 2 percent for all regressions. Notable but not essential is the fact that none of the constants are significant. Finally, reflecting on these findings $H0_b$: Positive price differences between offline and online prices correlate with the number of internet service providers and the amount of internet coverage via higher competition, can be rejected. Once again assuming that a positive difference strictly correlates to an increase in competition and that internet activity can be proxied by *totalISPcoverage* and *numberISPcompetitors.*

Results negative differences (offline price lower than online) regressions:

Here the goal is to investigate whether negative differences correlates with the independent variable in the direction of dynamic pricing, with regressions 2.1 – 2.3. The sample size is significantly smaller than that of positive difference with 2,526 observations rather than the previous 5,456 positive differences. This is important to keep in mind when the joint regression

results are presented, because the results are 2:1 biased towards the results of positive differences. As explained for positive differences, the following regressions containing observations with a negative difference, all variables with a negative coefficient increase the difference of prices and positive coefficients decrease the difference hence act in the opposite direction of dynamic pricing.

Regression 2.1 *totalISPcoverage* is significant at a 1 percent level with a small positive, even when considering the margin of error, coefficient of ~.06. The expected sign is negative since the paper assumes that negative prices correlate with strong dynamic pricing and thus increases the negative difference. Here instead the independent variable *totalISPcoverage* decreases the negative effect implying it is working in the opposite direction of dynamic pricing. The interpretation is therefore difficult. One speculates that the reason is that here higher competition comes into play reducing the price differences. The variable *Population* is also significant at a 1 percent level. Its coefficient is small at 0.000002. Similar to *MedianHHIncome* the mean population size in a zip code of this data set is considerably high at 31,229 (mean impact of ~.07 percent on price difference), but since ~.07 is barely acknowledgeable, the significance in absolute terms of this variable is questionable. The sign of the variable is as expected, thus the reasoning for this finding is that the population size of a zip code increases online competition, hence decreases online prices.

Allyearretailers is also significant at a 1 percent level with a negative, as expected, coefficient of ~.0006. This is a very low effect in the negative price difference between off- and online prices. The mean number of all year retailers lies at ~141, thus the mean effect is ~.08 percent. The sign of the coefficient is as said expected since the number of all year retailers indicates the level of offline competition in the zip code area and economic theory suggests that higher competition decreases the price of goods. Hence a bigger negative difference between off- and online prices. Furthermore, the variable *education* is significant at a 5 percent level. Its coefficient here is positive as anticipated at ~.19. As explained for regression 1.1 this translates to a mean of ~.47 percent increase and thus a decrease of the price difference between off- and online prices. No further control variables are significant.

Regression 2.2 results are the following: *numberISPcompetitors* is significant at a 1 percent level and positive, again not as expected for the independent variable. It's coefficient of ~.046 indicates

that the effect is on average ~.12 percent. The reason for the sign again must be speculated to be that higher competition is the dominant force. Here again the variable *Allyearretailers* is significant at a 1 percent level with an alike coefficient as in regression 2.1, hence no further elaboration is needed. Also, the variable *population* is once more significant but only at a 10 percent level with a lower coefficient of only .0000009.

The regression 2.3 includes both independent variables, but interestingly only *numberISPcompetitors* is significant, indicating that in regression 2.1 the significance of *totalISPcoverage* is merely a proxy for that of *numberISPcompetitors.* Here the *numberISPcompetitors* coefficient is again positive, raising the question if the assumptions about the forces acting upon positive and negative differences are robust. The results indicate that higher competition indeed is the dominant force in the correlation with negative price differences thus closing the gap of off- and online prices. The coefficient here is positive and at .05 slightly higher than in regression 2.2, but unsurprising since the coefficient of the other independent variable *totalISPcoverage* is negative. As in the regression 2.2 *Allyearretailers and population* are significant and their respective coefficients are comparable. Hence no further elaboration is needed. The adjusted r-squared is significantly higher than that of the positive differences at .04, .0519 and .0516, respectively for regression 1.1-1.3. This indicates that the *numberISPcompetitors* has the highest explanatory power and decreases when adding totalISPcoverage, which implies that *numberISPcompetitors* is a better explanatory variable. Also noteworthy is that here the constant term of all three regressions is significant at a 1 percent level and more negative than the positive counterparts were positive in regressions 1.1-1.3.

The results of these regressions indicate that $H0_c$: Negative price differences between offline and online prices corelate with the number of internet service providers and the amount of internet coverage in the direction of dynamic pricing must be rejected, however, negative price differences are decreased by higher competition; the more internet activity, the smaller the negative difference between off- and online prices. Having answered hypothesis B and C it is now proceeded to answering hypothesis $H0_a$, which includes the full dataset.

Full dataset

In the complete data in which both negative and positive price differences are combined the focus lies on which of the two forces (higher competition or dynamic pricing) is dominant. Since the data of the positive and negative differences are combined, the results are expected to correspond to the ones of the regressions 1.1-2.3 with a bias towards the results of positive values, since, as mentioned, the positive dataset is twice the size of the dataset of negative observations. Additionally, from the two previous results sections, it is expected that higher competition is the dominant force (a positive independent variable coefficient).

Regression 3.1 finds that *totalISPcoverage* positively influences the percentage difference between offline and online prices with a coefficient of ~.076 and is significant at a 1 percent level. As mentioned from the results of the previous two sections this result is unsurprising. The coefficient indicates that on aggregate *totalISPcoverage* thus, the internet activity with the online prices, decreases online prices as a result of higher competition. All control variables are significant at a 1 percent level. This is expected since all variables are significant at a 1 percent level in either regressions 1.1 or 2.1 or both, this regression being a combination of the two. All signs of the variables are the same as in the two previous results sections except *Allyearretailer* (negative). This is because regression 1.1 has a less positive and insignificant coefficient (.0003) and regression 2.1 contains highly significant and higher negative in absolute value coefficient (-.0006). The coefficient can be interpreted similarly as in regression 2.1: higher competition decreases the price of offline goods, hence a negative effect on the difference between off- and online prices.

The control variables *MedianHHIncome* and *demographics* coefficients mirror that of regressions 1.1-1.3 which is unsurprising due to the 2:1 bias in number of observations. The same interpretation of these coefficients is valid. Thus again, both signs of the coefficients are against expectations. The rationale for *MedianHHIncome* as mentioned before is that the negative coefficient, here -.000002, is a result of dynamic pricing, in which the online price increases due to an analytic price scheme which adjusts prices upward the higher the median income. The *Demographics* coefficient equals ~.012. Thus the effect, as in regression 1.1-1.3, is only

marginally significant in determining the difference in prices. Here the same logic can be used as in regressions 1.1-1.3. As said the coefficient goes against expectations and again dynamic pricing could explain this result: the higher the median age, the lower the use of internet, therefore the amount of data on the individuals is lower which leads to lower online prices. If the interpretation of the variables is true then *MedianHHIncome* and *Demographics* could be variables used for dynamic pricing algorithms.

The variable *population* coefficient is .0000015, which is smaller than its significant coefficient in regression 2.1. This is due to the fact that regression 1.1 has a small and insignificant negative coefficient. The interpretation is comparable to that of regression 2.1, being that the population size of a zip code area increases online competition, thus decreasing the online price and hence an increase in the price difference. On the other hand, *Education*'s coefficient strongly increases compared to regressions 1.1 and 2.1 it being ~.44. Again, that means that the average effect of *education* is a ~1.08 percentage increase in price increase. The same reasoning as in regression 1.1 and 2.1 holds: high school diploma or higher education increases internet activity, which creates higher online competition, decreasing online prices and thus increases the difference between off- and online prices.

In regression 3.2 in which the independent variable *totalISPcoverage* is switched for *numberISPcompetitors*, the independent variable is significant at a 1 percent level. The coefficient here is ~.067, again signifying the dominant force to be that of higher competition of online retailers decreases the online price and hence increases the gap between off- and online prices. The coefficient of ~.067 is close to the sum of the coefficients of the regressions 1.2 and 2.2 (~.068). Significant at a 1 percent level includes again variables *MedianHHIncome* and *Demographics*, of which both have the same coefficient as in regression 3.1. The variable *Allyearretailers* is now only significant at a 10 percent level. Its coefficient (~ -.00028) is also lower than in previous regressions. In regression 3.3 in which both *totalISPcoverage* and *numberISPcompetitors* are added we see that the independent variable *totalISPcoverage* is not significant as in regression 2.3. This similarity of results to 2.3, although a smaller size compared to 1.3, can be the fact that the negative price differences are on average a lot higher than the positive ones. Here again the same interpretation as in 3.2 can be applied. The significant variables are the exact ones of regression

3.2. Possibly because in regression 3.1 *totalISPcoverage* is not the right independent variable, such that other variables that are insignificant become significant. Hence, there are less significant control variables in regressions 3.2 and 3.3.

The adjusted r-squared or the explanatory power of this presented models is as expected lower than that of regressions 2.1-2.3 and higher than that of regressions 1.1-1.3. This is because the full dataset is the combination of the data sets used for the previous regressions and therefore a combination of their explanatory powers. The exact values are .0299, .0336 and .0325 for regressions 3.1-3.3. These values, although not high, are significant.

Interestingly and as touched upon before, in the joint data set although biased towards the observations of positive differences, the coefficients and their significance seem to be profoundly influenced by the results of negative difference. This is unsurprising as it was displayed in descriptive statistics that negative differences were on average significantly larger than those of positive differences, which is also evident in the difference in sizes of coefficients in the regressions.

From the results of the full data set the hypothesis $H0_a$ (the price gap between on- and offline goods differs between zip codes and correlates with the number internet service providers and the amount of internet coverage) can be answered. Without considering the previous two sections, it could be said that the hypothesis cannot be rejected. There is significant evidence that internet activity, proxied by the number internet service providers and the amount of internet coverage, correlates with the difference between off- and online prices. When the two previous hypotheses are taken into account dynamic pricing being driven by internet activity can be ruled out, higher competition does correlate with the price differences, more specifically with negative price differences. Higher competition may not be a significant driver for positive price difference, but it is for the negative and the joint dataset. This will be further elaborated on in the conclusion. First, the robustness of the paper's results is tested.

Table 5.1: Results from regressions.

Variables	Pos (1.1)	Pos (1.2)	Pos (1.3)	Neg (2.1)	Neg (2.2)	Neg (2.3)	Joint (3.1)	Joint (3.2)	Joint (3.3)
numberISP competitiors		0.0222735	0.0231477		0.0459163***	0.0478623***		0.0666569***	0.0649265***
		(0.0158647)	(0.0266966)		(0.0061208)	(0.0081069)		(0.011746)	(0.0183772)
totalISPcoverage	0.0230839		-0.001853	0.0555146***		-0.0053771	0.0764138***		0.0039928
	(0.0150365)		(0.0302711)	(0.0117722)		(0.0155001)	(0.0135404)		(0.0231002)
Population	0.0000002	-0.0000003	-0.0000003	0.0000022***	0.0000009*	0.0000009*	0.0000015***	0.0000001	0.0000001
	(0.0000006)	(0.0000008)	(0.0000008)	(0.0000004)	(0.0000005)	(0.0000005)	(0.0000006)	(0.0000007)	(0.0000007)
MedianHHIncome	-0.0000028***	-0.0000026***	-0.0000026***	-0.0000004	0	0	-0.0000023***	-0.0000018***	-0.0000018***
	(0.0000003)	(0.0000005)	(0.0000005)	(0.0000003)	(0.0000003)	(0.0000003)	(0.0000003)	(0.0000004)	(0.0000004)
Demographics	0.0119093***	0.0113969***	0.0114241***	0.0012819	0.0005327	0.0006833	0.0122424***	0.0109509***	0.0108793***
	(0.0014465)	(0.0020353)	(0.001791)	(0.0011573)	(0.0010626)	(0.0011608)	(0.0013127)	(0.0016385)	(0.0014908)
Education	0.3520921***	0.2213679	0.2178841	0.1927566**	-0.0738593	-0.0774561	0.4355803***	0.0455942	0.0518749
	(0.1331025)	(0.1953097)	(0.2296761)	(0.0813139)	(0.0952168)	(0.095747)	(0.1065915)	(0.1513695)	(0.1694585)
Allyearretailers	0.0002877	0.0003346*	0.0003354*	-0.0005941***	-0.0004743***	-0.0004730***	-0.0004336***	-0.0002780*	-0.0002796*
	(0.0001791)	(0.0001932)	(0.0001984)	(0.0001022)	(0.0001025)	(0.0001025)	(0.0001471)	(0.0001575)	(0.0001605)
Constant	0.0606266	0.0583974	0.0594668	-0.3683914***	-0.3513427***	-0.3471190***	-0.2481482***	-0.2391391***	-0.2417869***
	(0.0537746)	(0.0533813)	(0.0537371)	(0.0246929)	(0.0200153)	(0.0247297)	(0.0399832)	(0.0381617)	(0.0399971)
Observations	5,456	5,456	5,456	2,526	2,526	2,526	7,982	7,982	7,982
R-squared	0.022	0.022	0.022	0.04	0.054	0.054	0.031	0.033	0.033
Adjusted R-Squared	0.021	0.021	0.021	0.04	0.052	0.052	0.03	0.033	0.033

All cross-sectional regressions include the intercept and the (robust standard errors).

Statistically significant variables are denoted by the following level of significance *** p<0.01, ** p<0.05, * p<0.1

Coefficients in blue are significant independent variables, in yellow significant control variables, in green significant constants

Robustness results:

Testing for robustness is essential for the external validity of the results of the paper and thus gives the empirical results broader applicable economic significance. First, the full data from Cavallo's (2017) paper is taken to test how strong the results of this paper are when taking into account the full dataset. Secondly, the full data is weighted in such a way that rather than observations, zip codes have equal representation in the analysis. Here again, the presentation of the results is done chronologically.

Regression 4.1, including the same variables as in regression 1.1, has the following results. The independent variable *totalISPcoverage* is not significant at a 10 percent level. This though is unsurprising. By adding over 11 thousand observations in which there is no price difference, the independent variable which is believed to drive the difference in prices loses significance. All control variables coefficients signs stay constant compared to the finding of the paper. Three control variables are significant at a 1 percent level, these being *MedianHHIncome*, *demographics*, and *education*, which is the same as in the full dataset regressions. The coefficients are smaller than in the paper. This again is expected since the full dataset waters down the effect that these variables had on the dependent variable in the paper. Regressions 4.2 and 4.3 have similar results to that of regression 4.1. Interestingly, although adding the zero difference observations, the variables, significant or not, all have the same signs. Additionally, the coefficients are lower but proportional when comparing the results to that of the paper's regression. The adjusted r-squares are considerably lower than that of the paper's, at 0.0007 for all three regressions, signifying that the variables used in the regression have little explanatory power. The results of this robustness test are not surprising considering that adding the excluded data with zero difference naturally dilutes the effect of the variables. The complete output of the regressions 4.1-4.3 can be found in Appendix B, Table 5.2.

The second part of the robustness test is to determine whether the number of observations in a zip code influences the outcome of the regression. Therefore, the weighting of the observations is added in such a way that each zip code is accounted for equally regardless of the number of observations. The rationale is that since this paper concentrates on the differences in zip codes, the zip codes with many/few observations are over-/underrepresented, hence through weighting this difference is eliminated. The interpretation of the results of this robustness test can be kept concise since none of the independent or control variables are significant in all three regressions (5.1-5.3). A reason for this is that although the zip codes are now weighted equal, observations are now over/underweighted thus distorting the raw data. The adjusted r-square also indicates that the variables have very little explanatory power with values lower than that of regressions 4.1-4.3 (0.0001). The full results of the regressions 5.1-5.3 are in Appendix B, Table 5.3.

From the results of the robustness test, it can be concluded that the hypothesis A has to be rejected. The conclusions established in this paper do not hold for observations with no differences in off- and online prices. For Cavallo's dataset, the price gap between on- and offline goods between zip codes do neither correlate with the number of internet service providers nor with the total coverage of these internet service providers. This concludes the results section.

6. Conclusion

With the results gathered, a final conclusion with respect to the central research question can be drawn: Does the internet activity within a zip-code correlate with the difference between on- and offline prices of multi-channel retailers in the US for consumers?

From the results of regressions 1.1-3.3, the following answer to the central research question can be presented: Internet activity correlates with the difference in off- and online prices of multi-channel retailers in the US. This correlation is limited to when there is a difference in prices as the robustness test in regressions 4.1-5.3 show. However, since this research intends to investigate the reasons for price differences, the robustness test results are more a question of the external validity of the results. Henceforth, when referring to "price differences" the difference of offline and online prices is meant.

With respect to the forces assumed to drive online prices: higher online competition is the stronger force at least when there are negative price differences. This can be concluded since all the significant independent variables of regressions 1.1-3.3 have a positive coefficient. Meaning that there is a negative correlation of internet activity and online prices. On the other hand, dynamic pricing, channeled through the independent variable, does not seem to be present, which is coherent with Cavallo's (2017) findings.
Furthermore, from the results of the regression, it can also be determined which variables are being driven by either dynamic pricing or higher competition. Although dynamic pricing was not found to be present via independent variables, this is not true for the control variables. *MedianHHIncome* and *Demographics*, seem to be driven by dynamic pricing. This may imply that these variables are

used by computer algorithms to individualize online prices (dynamic pricing). Signifying that the data collected on an individual includes determining their age and income. *Population*'s and *education*'s significant coefficients imply to work in the direction of higher competition. The population size of a zip code area increases competition and thus correlates with the lowering of the online price of a good. Higher education increases the online competition through higher internet use and thus correlates with the decrease of online prices.

Since no real evidence of dynamic pricing is found, at least not via internet activity, as outlined in the theoretical framework, the morality, ethical meaning and implications of dynamic pricing seems to be unnecessary. However, since the control variables *MedianHHIncome* and *Demographics* work in the direction of dynamic pricing, its moral implications should not be completely ignored and multi-channel retailers that use dynamic pricing or similar forms of pricing schemes should be warned of the consumer's negative perception of such behavior.

As the description of the data presented the results are non-representative for the whole of the US. More accurately, due to the number of observations, it is most likely to be only applicable to the state of Massachusetts. Hence the conservative conclusion: this paper finds that the higher internet activity in a zip code area within the state of Massachusetts positively correlate, because of what is assumed the force of higher competition, with the difference of off- and online prices of multi-channel retailers.

The questions that arise from this paper are numerous. First of all, having determined that the number of ISP competitors or internet activity is correlated with the price difference of off- and online prices, this brings up the question whether this, with the growing number of internet users and activity, may lead to even lower internet prices. Furthermore, what determines negative and positive differences between off- and online retail prices. Does the discrepancy of prices increase when adding non-multi-channel retailers data? And most fundamentally whether these results that are correlations of the variables with the dependent variable are just mere coincidences? This paper argues that although it cannot be said that these relationships are causal, it can be said that the variables chosen for this paper could be used as realistic indicators for the determination of price differences of multi-channel retailers.

Another very present matter is dynamic pricing, which although not clearly evident in this paper's findings, will nevertheless be a very challenging issue in the future, especially when it becomes more efficient in identifying consumers demand curves. Also, if online competition increases in the future, will it be able to offset the force of dynamic pricing that potentially increases prices on an aggregate and individual level? The effects of the change in the retailing sector on consumers and suppliers, due to the introduction of the internet, will be ever more critical for economist and policy makers when forming future legislations on the neutrality and freedom of the internet.

Limitations and Recommendations

Limitations of this paper are numerous. This is because what the paper tries to identify requires multiple assumptions to hold. As mentioned, this paper can only speculate if the relationship is indeed a causal- or, more often the case, a correlation relationship between the independent and dependent variable. Thus, using time-series data could help find the causal effect of the independent variable on the dependent variable. One of the most fundamental assumptions used is that internet activity can be proxied by the number of internet service providers within a zip code area. Building on this assumption is that the proxy includes the force of dynamic pricing and higher competition. The assumption of higher competition seems more straightforward. On the other hand, dynamic pricing is best measured by comparing the price of a single good between different buyers. This further limit the interpretation of the findings made on dynamic pricing. Furthermore, the data used limits the paper to the scope of the data. Although the observations are many they are not nationally representative, focused mostly in the states California, New York and especially Massachusetts. Another limitation is that the goods in the dataset are thousands. The analysis would be more robust if the observations were more focused on few goods from the same retail sector. Other limitations include the differences in years of the control variables that in an optimal case would be all the same.

The study of what effect the introduction of e-commerce has on retailing, the retail sector representing 5.9% of US GDP in 2017 (Statistica, 2018), is of essence. Furthermore, if in the US there is a considerable amount of off- and online price difference (in this paper 40% of the data

set), what determines these differences is for economist and consumers alike crucial for the understanding of the future of pricing and retailing. The recommendations for further investigation are the following. In the results section the adjusted r-squared indicates that the explanatory power of the regression is low (max 5%), thus the goal for further research should be in determining other reasons and control variables that would increase the explanatory power of the regression. Another interesting investigation would be in determining what causes positive and negative price differences. Also, further data collection especially representative data, over numerous years of the same goods, not only from multi-channel retailers, and from various states would enable the researcher to make robust economic claims for the determinants of online prices of retailers. As mentioned as a limitation, doing the analysis with more specific goods or within retail sectors can increase the accuracy of the results found. To eliminate the uncertainty of the proxy a further recommendation is collecting data of actual internet activity. Additionally, zip code areas are a very large unit for measuring something that targets individuals such as dynamic pricing. Finding the internet activity from a smaller unit would also increase the accuracy and validity of the results. This concludes the limitation and recommendation section.

Bibliography:

Associated Press (2007), "Inside the World of Price Optimization Software," *Valley News*, April 29, E2.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management science*, *43*(12), 1676-1692.

Besbes, O., & Zeevi, A. (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. Operations Research, 57(6), 1407-1420.

BIGCommerce. What is multichannel retailing. Retrieved from: https://www.bigcommerce.com/ecommerce-answers/what-is-multichannel-retailing/

Brown, J. R., & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry. *Journal of political economy*, *110*(3), 481-507.

Brynjolfsson, E., Hu, Y., & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, *49*(11), 1580-1596.

Cavallo, A. (2017). Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review*, *107*(1), 283-303.

Chapuis, J. (2012). Price fairness versus pricing fairness.

Chaudhuri, A., Flamm, K. S., & Horrigan, J. (2005). An analysis of the determinants of internet access. *Telecommunications Policy*, *29*(9-10), 731-755.

Coffinet J, Perillaud S (2017). Effects of the Internet on Inflation: an overview of the literature and empirical analyse. IMF.

Faruqui, A., & George, S. (2005). Quantifying customer response to dynamic pricing. The Electricity Journal, 18(4), 53-63.

The Economist. 2016. *Flexible figures, A growing number of companies are using "dynamic" pricing.*The Economist. Published on the 12.06.2016. Retrieved on 04.06.2018. Retrieved from: https://www.economist.com/business/2016/01/28/flexible-figures

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. Management science, 49(10), 1287-1309.

(21.04.2012) The third industrial revolution. *The Economist.* Accessed 15.05.2018. Retrieved from: https://www.economist.com/node/21553017

Garbarino, E., & Maxwell, S. (2010). Consumer response to norm-breaking pricing events in e-commerce. Journal of Business Research, 63(9-10), 1066-1072.

Goldstein, A., & O'connor, D. (2000). E-commerce for Development.

Goolsbee, A. D., & Klenow, P. J. (2018). Internet Rising, Prices Falling: Measuring Inflation in a World of E-Commerce. Stanford University working paper, January.

Grewal, D., Ailawadi, K. L., Gauri, D., Hall, K., Kopalle, P., & Robertson, J. R. (2011). Innovations in retail pricing and promotions. *Journal of Retailing*, *87*, S43-S52.

Haws, K. L., & Bearden, W. O. (2006). Dynamic pricing and consumer fairness perceptions. Journal of Consumer Research, 33(3), 304-311.

Hays Constance L.(1999).*Variable-Price Coke Machine Being Tested.* The New York Times. Published on the Oct. 28. 1999. Retrieved on the 08.06.2018. Retrieved from: https://www.nytimes.com/1999/10/28/business/variable-price-coke-machine-being-tested.html

Hotelling H, (1929) *Stability in Competition*, Economic Journal, 39 (153): 41–57

International Post Corporation. (2017). State of e-commerce: global outlook 2016-2021. Accessed on the 09.05.2018. Retrieved from: https://www.ipc.be/en/knowledge-centre/e-commerce/articles/global-ecommerce-figures-2017

Jacobson, L. (2017). *What is net neutrality?* Retrieved from abcnews: https://abcnews.go.com/Technology/net-neutrality/story?id=48596615

Leonhard David. (2005). *Why Variable Pricing Fails at the Vending Machine*. The New York Times. Published on the 27.06.2005. Retrieved on the 08.06.2018. Retrieved from: https://www.nytimes.com/2005/06/27/business/why-variable-pricing-fails-at-the-vending-machine.html

Liao, Z., & Cheung, M. T. (2001). Internet-based e-shopping and consumer attitudes: an empirical study. *Information & management*, *38*(5), 299-306.

Meola Andrew (2017). *The Rise of M-Commerce: Mobile Shopping Stats & Trends.* Business Insider. Retrieved on the 04.06.2018. Retrieved from: http://www.businessinsider.com/mobile-commerce-shopping-trends-stats-2016-10?international=true&r=US&IR=T

Porter, C. E., & Donthu, N. (2006). Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. Journal of business research, 59(9), 999-1007.

Sahay, A. (2007). How to reap higher profits with dynamic pricing. MIT Sloan management review, 48(4), 53-60.

Statistica (2018). *Percentage added to the Gross Domestic Product (GDP) of the United States of America in 2017, by industry (as a percentage of GDP)*. Statistica. Accessed on the 8.07.2018. Retrived from: https://www.statista.com/statistics/248004/percentage-added-to-the-us-gdp-by-industry/

Trading Economics (2018). United States Core Inflation Rate. Accessed on the 15.06.2018. Retrieved from: https://tradingeconomics.com/united-states/core-inflation-rate

Trainer, D. 2016. "How the Internet Economy Killed Inflation." *Forbes*. September 29. Accessed May 9, 2017. https://www.forbes.com/sites/greatspeculations/2016/09/28/how-the-internet-economy-killed-inflation/#52f8b342788b.

Us. Census Bureau News (2018). Quarterly Retail E-Commerce Sales 4th Quarter 2017. Accessed 10.05.2018. Retrieved from: https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

Valentino-DeVries Jennifer, Singer-Vine Jeremy, Soltani Ashkan. 2012. *Websites Vary Prices, Deals Based on Users' information*. The Wall Street Journal. Published on the 24.12.2012. Retrieved on the 04.06.2018. Retrieved from: https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

Weiss, R. M., & Mehrotra, A. K. (2001). Online dynamic pricing: Efficiency, equity and the future of e-commerce. *Va. JL & Tech.*, *6*, 1.

Weisstein, F. L., Monroe, K. B., & Kukar-Kinney, M. (2013). Effects of price framing on consumers' perceptions of online dynamic pricing practices. *Journal of the Academy of Marketing Science*, *41*(5), 501-514.

World Population Review. (2018). *United States Population 2018.* Accessed on the 15.08.2018. Retrieved from: http://worldpopulationreview.com/countries/united-states-population/

Zip-Codes (2010). Accessed on the 05.06.2018. Retrieved from: https://www.zip-codes.com/state/ma.asp

ZIP code Retail Trade Business Patterns (2018) published by ProximityOne information & solutions. Retrieved from: http://proximityone.com/zip-retail-trade.htm

Appendix:

Appendix A: Descriptive statistics

Table 3.1:
Brief description and origin of Variables used in this paper

Variables	*Description*	*Origin*
Difference of off- and online Prices,	Is as the name states the offline price of a good minus its online price, of goods from 2014-2016	Cavallo (2017)
Totalcoverage ISP	Is the sum of all internet coverage within a zip code area including that of ISP of the same company e.g Verizon FIOS and Verizon High Speed Internet (2017)	Broadbandnow.com
Competition ISP	Is the sum of the number of different ISP within a zip code area (2017)	Broadbandnow.com
Median HH Income	Is the median house hold income in 2016 by zip code area	US Bureau of Census
Population	Form the decade population census of 2010 that reflects the population of each zip code area (2010)	US Bureau of Census
Education	Is the percentage of individuals within a zip code area that have obtained a high school degree or higher (2016)	US Bureau of Census
Allyearretailers	Is the number of all year-round retailers within a zipcode area in 2017	Proximityone,com
State	The US state in which the Zip code is found in	Proximityone,com
Demographics	The median age within a zip code area (2016)	US Bureau of Census

Table 3.2: Zip code by State

State	Zip Code Freq.	Percent
AK	1	0.01265823
AZ	1	0.01265823
CA	12	0.15189873
DC	1	0.01265823
FL	5	0.06329114
LA	1	0.01265823
MA	17	0.21518987
MD	3	0.03797468
ME	1	0.01265823
MI	2	0.02531646
MO	3	0.03797468
NC	1	0.01265823
NE	1	0.01265823
NJ	4	0.05063291
NV	1	0.01265823
NY	6	0.07594937
OH	1	0.01265823
OK	1	0.01265823
PA	4	0.05063291
SC	2	0.02531646
TX	3	0.03797468
VA	8	0.10126582
Total	79	1

Retailers in the data as used in the paper Cavallo (2017)

Name of US retailers used in paper	Walmart, Target, Safeway, Stop&Shop, Best Buy, Home Depot, Lowe's, CVS, Macy's, Banana Republic, Forever 21, GAP, Nike, Urban Outfitters, Old Navy, Staples, OfficeMax/Depot.

Table 3.6: The Percentage difference of off- and online prices between zip-code areas

Zipcode	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
1040	14	.1290766	.5426197	-.687625	1.364877
1701	944	.3077874	.4974342	-.9887297	2.988
1760	9	.0336846	.1830364	-.2800112	.3335557
1905	93	.1368498	.226962	-.3941051	.8982036
2110	211	.4429269	.3975297	-.6002401	1.501251
2111	24	.3918217	.4179585	-.3297207	1.927332
2114	52	-.020026	.1402134	-.3464052	.2429266
2115	52	-.020026	.1402134	-.3464052	.2429266
2116	8	.1379533	.272437	-.1430615	.5108165
2125	933	.1807832	.3903339	-.9641317	3.485
2134	435	.1597979	.3097207	-.5709516	1.984925
2138	773	.0773539	.5865098	-.9591669	12.87782
2139	701	.0185771	.3742914	-.7620257	2.338898
2141	709	-.0561759	.3849137	-.7699425	2.215868
2210	16	.083883	.3243199	-.230331	.8322774
2215	1,001	.0882625	.6507603	-.8064907	10.02004
2472	97	.3074952	.5040271	-.9500679	2.625369
2481	235	.0317628	.2074315	-.9620596	1.103286
4106	8	.3681422	.6998052	-.6750338	1.712943
8052	41	.4409522	2.970005	-.553125	18.95
8054	6	.2610475	.4084427	-.245942	1.002004
8057	14	-.0700639	.2953158	-.500417	.500125
8077	44	.1044703	.568375	-.1430615	3.765491
10025	45	.2794385	.4778552	-.5745903	2.604736
10701	30	.1590909	1.068046	-.9905428	5.283019
11101	14	.0001471	.2765474	-.8955588	.2857551
11354	20	.217716	.9504866	-.4000286	4.001
11369	12	.4538054	.6042746	-.2997666	2.079292
12077	99	.151448	.2702398	-.3196347	1.26
12203	14	.2987876	.3875445	-.0425758	1.393333
16148	8	.2875427	.2595649	-.2330769	.6091861
17078	12	.073364	.0751546	-.0374532	.1896733
19107	9	.4562513	.3023278	.0400925	.8418556
19454	15	.5037017	.5336517	-.0828571	1.994
20007	68	-.1080349	.3796173	-.6959548	1.012658
20141	13	.1552629	.4787911	-.60012	1.057958
20175	5	.4771236	.6811165	-.230789	1.5005
21703	7	.0652249	.6883383	-.8401146	1.0002
21804	11	.1851024	.3318928	-.3350084	1.003788
21851	6	-.0359197	.2986444	-.4887526	.3226667

22101	42	-.0357691	.1454245	-.3128911	.6030151
22801	6	.2212463	.3500546	-.1434286	.8675
23233	6	.0730134	.1483723	-.0909918	.251651
23321	5	-.2449194	.2975638	-.6426117	.087108
23602	10	.2482366	.5116972	-.9161074	.9509755
23707	19	-.1209987	.3771719	-.9229515	.2814286
29301	9	-.0113425	.1817836	-.3	.2564103
29607	9	.2846031	.2531932	.0040282	.7576114
32222	13	-.0284422	.189536	-.3501751	.30131
32825	7	.0114245	.4661029	-.6250781	.60012
32901	7	-.1460042	.3608008	-.6667778	.3333444
33071	7	.4820605	.5673102	-.4668	1.000333
34119	8	.0304944	.3826309	-.6669446	.4286939
43219	4	.2852387	.1595093	.0625	.4285714
48083	15	.5916094	.3031869	.13	1.060606
48180	6	-.0647834	.2578845	-.5006258	.2376485
63143	6	-.0470014	.3105408	-.5583039	.3240557
63144	10	.4102863	.9905415	-.2418075	3.169308
64057	7	.3175813	.6240785	-.9646974	.8468642
68046	9	.0530006	.1960723	-.1059783	.500025
70503	10	.0291827	.3941566	-.6781377	.4757379
73018	6	-.0668511	.1263932	-.2056452	.0719424
77429	5	-.1944931	.3908087	-.8853269	.0417362
77521	34	.1402616	.4409289	-.4432071	1.501502
77550	26	.0613883	.3187705	-.6549344	1.002225
85338	7	-.0434886	.2107965	-.3212851	.196
89102	6	-.1253946	.4300931	-.3333889	.7501875
90016	6	-.2081263	.1276844	-.3622414	-.080032
90640	13	.2482442	.8901536	-.4001818	3.003003
90703	23	-.0446666	.3351536	-.6556017	1.0005
91791	105	.1502217	.339106	-.6470824	1.635884
92108	7	1.112846	.3806561	.607717	1.834278
93551	6	.3818167	.7364247	-.4557393	1.50075
94070	360	.1012915	.1933657	-.1078582	2.532391
94501	14	.1816519	.1000619	-.1026958	.3271984
94533	6	-.0442514	.4492901	-.8128544	.4511278
94538	151	.1120119	.6472684	-.713467	5.421687
94539	38	.1118625	.1581232	-.2675585	.596
94560	27	.4704038	.4949812	-.2975644	1.301534
94588	6	-.1010985	.3741515	-.5	.3589834
Average	99.53	.15	.43	-.46	1.73

Table 3.10:
Skewness/Kurtosis tests for Normality

Variable	*Obs*	*Pr(Skewness)*	*Pr(Kurtosis)*	*Prob>chi2*
Percentage Difference of Offline and Online price	7,982	0.0000	0.0000	0.0000
Education	7,982	0.0000	0.7608	0.0000
Median Age	7,982	0.0000	0.0000	0.0000
MedianHHIncome	7,982	0.0000	0.0000	0.0000
NumberofcompetingISP	7,982	0.0000	0.0000	0.0000
TotalISPcoverage	7,982	0.0000	0.0000	0.0000
Allyearretailer	7,982	0.0000	0.0000	0.0000

Table 3.12:
Variance Inflator Factor: checking for multicollinearity

Variable	*VIF*	*1/VIF*
MedianHHIncome	3.45	0.290122
Median Age	3.23	0.309448
TotalISPcoverage	1.90	0.526135
Education	1.80	0.554211
Pupulation	1.20	0.831759
Mean VIF	2.16	

Table 3.13:
Zip-code areas and their frequencies

ZIPCODE	*Freq.*	*Percent*
1040	14	0.18
1701	944	11.83
1760	9	0.11
1905	93	1.17
2110	211	2.64
2111	24	0.30
2114	52	0.65
2115	152	1.90
2116	8	0.10
2125	933	11.69
2134	435	5.45
2138	773	9.68
2139	701	8.78
2141	709	8.88
2210	16	0.20
2215	1,001	12.54
2472	97	1.22
2481	235	2.94
4106	8	0.10
8052	41	0.51
8054	6	0.08
8057	14	0.18
8077	44	0.55
10025	45	0.56
10701	30	0.38
11101	14	0.18
11354	20	0.25
11369	12	0.15
12077	99	1.24
12203	14	0.18
16148	8	0.10
17078	12	0.15
19107	9	0.11
19454	15	0.19
20007	68	0.85
20141	13	0.16
20175	5	0.06
21703	7	0.09
21804	11	0.14
21851	6	0.08
22101	42	0.53
22801	6	0.08
23233	6	0.08
23321	5	0.06
23602	10	0.13
23707	19	0.24
27106	7	0.09

29301	9	0.11
29607	9	0.11
32222	13	0.16
32825	7	0.09
32901	7	0.09
33071	7	0.09
34119	8	0.10
43219	4	0.05
48083	15	0.19
48180	6	0.08
63143	6	0.08
63144	10	0.13
64057	7	0.09
68046	9	0.11
70503	10	0.13
73018	6	0.08
77429	5	0.06
77521	34	0.43
77550	26	0.33
85338	7	0.09
89102	6	0.08
90016	6	0.08
90640	13	0.16
90703	23	0.29
91791	105	1.32
92108	7	0.09
93551	6	0.08
94070	360	4.51
94501	14	0.18
94533	6	0.08
94538	151	1.89
94539	38	0.48
94560	27	0.34
94588	6	0.08
99515	6	0.08
Total	7,982	100.00

Table 3.14:
Expected signs of variables (effect on offline-online prices)

Variables	Number of Competitors	Total Coverage	All year retailers	Education	Median Household income	Population	Demographics
Sign	+/-	+/-	(-)	+	+	+	(-)

Appendix B:

Robustness test results:

Table 5.3: Robustness test weighted Cavallo's full data set

Table 5.2: Robustness test 1. Cavallo's (2017) full data set

Variables	Reg (4.1)	Reg (4.2)	Reg (4.3)
numberISPcompetitiors		0.0100109	0.0148067
		(0.0100192)	(0.0104694)
totalISPcoverage	0.0033355		-0.0117633
	(0.0147145)		(0.0146609)
Population	0.0000007	0.0000004	0.0000005
	(0.0000004)	(0.0000006)	(0.0000006)
MedianHHIncome	-0.0000012***	-0.0000011**	-0.0000011**
	(0.0000004)	(0.0000004)	(0.0000004)
Demographics	0.0048671***	0.0042563***	0.0044927***
	(0.0012256)	(0.0013905)	(0.0014212)
Education	0.3255179***	0.2579179*	0.2428125*
	(0.0977885)	(0.1334077)	(0.1423713)
Allyearretailers	-0.0001001	-0.0000848	-0.0000853
	(0.0001437)	(0.0001425)	(0.0001421)
Constant	0.0007376	-0.0062855	0.0005928
	(0.0205467)	(0.0216471)	(0.020562)
Number of Zip codes	250	250	250
Observations	19,795	19,795	19,795
R-squared	0.0009866	0.0010289	0.0010431
Adjusted R-Squared	0.0007	0.0007	0.0007

All cross-sectional regressions include the intercept and the (robust standard errors).
Statistically significant variables are denoted by the following level of significance *** p<0.01, ** p<0.05, * p<0.1

Coefficients in Blue are significant independent variables, in Yellow significant control variables, in Green significant constants

Table 5.3: Robustness test weighted Cavallo's full data set

Variables	Reg (5.1)	Reg (5.2)	Reg (5.3)
numberofISPcompetitiors		-0.0526151	-0.0771115
		(0.0612267)	(0.0602202)
totalISPcoverage	0.0494887		0.1172151
	(0.1135881)		(0.1084221)
Population	0.0000016	0.0000024	0.0000019
	(0.0000017)	(0.0000022)	(0.0000019)
MedianHHIncome	-0.0000014	-0.0000015	-0.0000016
	(0.0000011)	(0.0000011)	(0.0000012)
Demographics	-0.0039174	-0.0044131	-0.0033681
	(0.0036701)	(0.0036203)	(0.0032851)
Education	0.5663741	0.7229043	0.600428
	(0.6858001)	(0.7946234)	(0.6853775)
Allyearretailers	0.0001398	0.0001993	0.0001622
	(0.0004653)	(0.0004493)	(0.0004742)
Constant	0.2605225	0.5306633	0.2970129
	(0.2676082)	(0.3483737)	(0.2941482)
Number of Zip codes	250	250	250
Observations	19,795	19,795	19,795
R-squared	0.0004501	0.000629	0.0008475
Adjusted R-Squared	0.0001	0.0003	0.0005

All cross-sectional regressions include the intercept and the (robust standard errors).
Statistically significant variables are denoted by the following level of significance *** p<0.01, ** p<0.05, * p<0.1

Coefficients in Blue are significant independent variables, in Yellow significant control variables, in Green significant constants