Proxy CDS Curves for Individual Corporates Globally

Jin-Chuan Duan*

(First Draft: September 8, 2017; This Version: July 5, 2020)

Abstract

CDS with some liquidity are limited to under 500 corporate names globally. We deploy modern data analytics to develop a practical and robust predictive regression linking liquid CDS premiums of different tenors to a set of common factors and obligor/instrument-specific attributes. This model can then be used to generate proxy CDS curves for corporates without liquid or quoted CDS. The key attributes among many potential predictors are (1) the actuarial spread that reflects the actuarial value of a CDS which is available for all exchange-listed firms globally, and (2) credit market environment variables such as aggregate CDS indices.

Keywords: Actuarial spread, distance to default, credit cycle index, investment grade, high yield, big-data, zero-norm penalty, sequential Monte Carlo.

*Duan is with the National University of Singapore (Business School, Risk Management Institute and Department of Economics). E-mail: bizdjc@nus.edu.sg. The author acknowledges CriAt, a FinTech company specializing in deep credit analytics, for making available its proprietary software for selecting regressors subject to a zero-norm penalty, and also thanks Siyuan Huang for her able research assistance. The author appreciates the valuable comments from the participants of many university seminars and academic conferences.

1 Introduction

Corporate credit default swap (CDS) par spread (i.e., standardized premium) is the market price of credit risk posed by a corporate obligor, reflecting probability of default, recovery rate on the reference debt instrument, additional risk premium demanded by risk averse economic agents, liquidity condition of the CDS market, and potential counterparty default. CDS are commonly used for risk benchmarking in credit risk management in general and in accounting practice in particular, where the latter pertains to the new accounting reporting standards on credit exposures (e.g., IFRS-9 for international firms and CECL for US firms). However, CDS with adequate liquidity are hard to come by, and are by any reasonable standard less than 500 corporate names globally. CDS users must either confine their usage to this limited subset of liquid CDS or simply resort to some aggregates derived from the liquid CDS in different industry/rating combinations made available by commercial vendors such as IHS Markit.

We offer in this paper an intuitive, practical and robust predictive regression model, linking liquid USD-denominated CDS par spreads of different tenors to a set of obligor-specific attributes and some common factors. This model can then be used to generate proxy CDS curves for corporates without liquid or quoted CDS. The model is developed as a single universal predictive regression for all corporate CDS over a long time span, and it delivers an R^2 of over 70% out-of-sample and performs robustly for different subgroups of interest. The key to success hinges on the actuarial spread (AS) made available by the Credit Research Initiative (CRI) team of National University of Singapore, which produces daily updated ASes, among other credit risk measures, on all exchange-listed firms globally, and these data are freely accessible. AS as a predictor alone is found to deliver an R^2 of over 35%. Other obligor-specific attributes in the predictive regression include investment vs. speculative grades based on an obligor's credit rating, and some general credit environment variables. Most noteworthy is the set of four CDS indices constructed for four subgroups of corporate names – investment-grade, high-year, financial and non-financial. The CDS Big Bang in April 2009 is another factor of interest, which introduced several key changes to CDS trading including (1) setting the fixed premium rate to either 100 or 500 basis points while using an upfront fee to offset the effect of the fixed premium rate, and (2) removing reorganization as part of the default definition.

Operationally speaking, our predictive regression model, constructed with the 15-year historical record on 405 corporate CDS names, makes it possible to generate daily-updated proxy CDS curves on over 36,000 currently active exchange-listed corporates globally, reflecting the CRI's coverage of virtually all exchange-listed firms in the world today.

CDS and corporate bond pricing is a much researched topic in the literature, and theoretical pricing models abound; for example, Merton (1974), Longstaff and Schwartz (1995), Duffie and Singleton (1999), Duffie and Lando (2000), Das and Sundaram (2000), and Hull and White (2000), to name just a few. By design, these theoretical models mainly focus on the risk premium arising from risk aversion of economic agents, and are typically stylized in a way to avoid practical complications such as multiple risk drivers, liquidity, counterparty default, or supply-demand imbalance. These theoretical models also come with unknown parameters that need to be estimated, and some of the models go further to rely on latent variable(s), for example, unobserved default intensities. In order to have reasonable empirical performance, the unknown parameters and/or latent variable(s) need to be estimated with market prices on some credit-sensitive instruments such as corporate bonds and/or CDS on the obligor in question. Since the model parameters and/or latent variable(s) are obligor-specific, they cannot be easily ported to CDS referencing different obligors. In short, these theoretical models are limited to applications on the pricing of CDS on corporates with traded bonds and/or CDS.

Empirical studies of corporate CDS are even more numerous to cover all. Most studies were designed to focus on whether CDS are priced according to some theory as opposed to addressing how CDS can be practically priced through a predictive relationship developed on other liquid traded CDS. Ericsson, *et al* (2009), for example, studied the CDS premium in relation to three general theoretical predictors – leverage, volatility and riskless interest rate – on a firm-by-firm basis to find an average R^2 in the order of 60%. When dealing with the three predictors on an individual variable basis, the average R^2 drops to less than 15%. Since the regression is run on a firm-by-firm basis, the coefficients developed on one corporate with traded CDS cannot be used for other corporates without CDS even if one is satisfied with the level of R^2 based on the three-variable model. In short, their study confirms the theoretical prediction by addressing the issue of "why" but offers no practical answer to "how" to apply the model. The relationship of CDS premium vs. corporate bond yield spread (risky bond yield minus riskless bond yield) of the same tenor on the same obligor has been the subject of many empirical studies, for example, Blanco, *et al* (2005) and Zhu (2006) confirmed a long-run parity relationship between the two credit risk measures. Kim, *et al* (2017) further investigated the basis (CDS premium vs. corporate bond yield spread) behavior to see whether basis arbitrage is possible. This line of studies again sheds light on whether a theory holds or arbitrage opportunity exists, but offers no practical solution to predicting CDS for corporates without traded bonds of comparable terms.

Our CDS prediction model utilizes the advancement in modern big-data analytics, particularly the zero-norm penalty regression, which means that one chooses an optimal subset of k regressors among all potential predictive variables. When the number of potential regressors increases to, say, several hundred, the number of possible combinations quickly becomes astronomical, making an exhausted search infeasible even with modern high-power computers. In this paper, we apply the zero-norm penalty regression technique developed by Duan (2019), which utilizes sequential Monte Carlo simulations. Modern penalty regression techniques are typically based on the l_1-norm due to computational considerations; for example, the Lasso of Tibshirani (1996), the SCAD of Fan (1997) and Fan and Li (2001), and the adaptive Lasso of Zou (2006). However, selecting regressors based on the zero-norm penalty is conceptually more appealing, because it directly addresses the essence of the variable selection problem. Computing speed aside, it works better because regression coefficients will not be distorted by the penalty term (i.e., shrinkage toward zero even being selected). Also interesting to note is the fact that the regression model fit, measured by R^2, is invariant to linearly transforming a group of regressors but the corresponding l_p ($p > 0$ & $\neq 2$) penalty term is not. Therefore, multicolinearity will interfere with regressor selection based on an l_p ($p > 0$ & $\neq 2$) penalty, but not with the zero-norm regressor selection.[1] Readers are referred to Duan (2019) for a theoretical explanation and simulation evidence of its superior performance.

We consider 42 variables and their interaction terms to have a total of 810 potential regressors in developing this predictive model. The data are USD-denominated CDS par spreads for 405 reference obligors with tenors from 1 to 5 years with a monthly frequency over

[1]Although an l_2-norm penalty regression, i.e., ridge regression, will not encounter this kind of distortion, it is well known to exhibit poor ability in reducing the number of regressors.

the period from August 2001 to February 2017. In addition to US corporates, there are firms from 21 other economies, totaling 141,749 observations in 10 industries by the Bloomberg Industry Classification System. Our zero-norm penalty regression employs a five-fold cross validation on a training sample of two-thirds of the whole dataset, which is randomly chosen with the remainder saved as a holdout sample for an out-of-sample performance study. Our method selects an optimal combination of 20 regressors from the 810 potential variables, and many of them are interaction terms. This predictive regression model delivers an R^2 of over 70% in-sample and out-of-sample, and is found to be fairly stable across different subgroups of interest.

2 Constructing proxy CDS curves

Our approach to constructing a practical proxy CDS curves on five specific tenors (1, 2, 3, 4 and 5 years) is entirely empirical but guided with economic intuition. We first gather a substantial sample of USD-denominated CDS par spreads, spanning over 15 years on a monthly frequency for as many corporate names as we can obtain. Next, we move on to identifying a set of attributes that are concurrently available and intuitively related to the market price of CDS. By considering the interaction terms of these attributes, we obtain a very large set of potential explanatory variables, which in fact equals 810. Finally, we rely on a zero-norm penalty based variable selection technique to conclude that a subset of 20 regressors can robustly predict CDS par spreads.

2.1 The CDS data

Our CDS par spreads are the Bloomberg computed CDS averages with end-of-day set to 6:00pm EST (New York time). We focus on USD-denominated CDS and extract data from Bloomberg on a monthly frequency starting in August 2001 all the way to February 2017. The 405 corporate names in our extracted USD-denominated CDS sample include beyond US firms (309 out of 405) to cover firms from 21 other economies with Canadian firms being the second largest group (20 out of 405). The firms in the sample covers all 10 industries according to the Bloomberg Industry Classification System with Financial being the largest containing 73 firms and Diversified being the smallest having 4 firms. The five CDS tenors are fairly equally distributed where 354 firms with 1-year, 319 with 2-year, 356 with 3-year, 314 with 4-year and 404 with 5-year. Define the post 2008 global financial crisis period as September end of 2008 and afterwards. The post-crisis sample contains 395 firms, whereas the pre-crisis sample has 244 firms. The CDS Big Bang occurred on April 8, 2009, and our post-Big Bang period of monthly frequency naturally starts from April 2009 onward dividing the sample into 374 and 372 firms in the post and pre Big Bang periods, respectively. The CDS sample contains 141,749 firm/month/tenor observations in total with 118,401 being investment-grade and the rest being high-yield. Some descriptive statistics on the CDS data with and without applying a transformation are provided in Table 1.

We construct the proxy CDS model using the transformed value of CDS. If we directly built the prediction model on the original CDS values, the model would often generate negative predictive CDS values, which is obviously undesirable. Our transformation is $\ln(\exp(CDS/100) - 1)$ which requires some explanation. Note that CDS is per usual stated in basis points. This transformation agrees with the natural logarithm when CDS has a small value, but converges to itself (stated in percentage points) when CDS has a large value. Were the natural logarithm the adopted transformation, it would end up distorting large CDS values, implying that a good prediction model for the transformed CDS value may still work poorly for the original CDS value. Our transformation is appealing because it retains a nice property of the natural logarithm for small CDS values while avoiding the distortion for large CDS values.

Our sample also contains 92 CDS data points referencing subordinated debt, and all are 5-year tenor with Shinshei Bank, a Japanese financial institution, as the reference entity. The data on this subordinated debt CDS fall in the period from April 2006 to December 2013. The sample suggests that a great majority of CDS only references senior unsecured debt.

The aforementioned categorical data characteristics (e.g., CDS Big Bang, investment-grade vs. high-yield firms, etc.) will be used along with some other more granular attributes concerning individual corporate names in constructing our proxy CDS model. And some of the categorical features indeed play a prominent role in explaining CDS par spreads, and can help predict CDS values when their market quotes are unavailable.

Table 1: Single-regressor R^2 and summary statistics for CDS and its 42 explanatory variables

	R^2	**Mean**	**Std**	**Max**	**Min**
CDS(bps)		150.098	328.557	9592.201	1.235
Dependent Variable					
T(CDS)		0.620	3.640	95.922	-4.388
Regressor					
SIGMA	0.43800	0.079	0.056	0.949	0.023
T(ASlevel)	0.36034	-2.033	2.163	42.964	-14.850
T(AS)	0.35695	-2.220	2.573	149.907	-16.529
DTDlevel	0.21299	5.522	3.060	20.108	-1.176
isHY	0.21007	0.165	0.371	1.000	0.000
isIG	0.21007	0.835	0.371	1.000	0.000
SIZElevel	0.18403	3.541	1.439	8.138	-2.265
IndustryCCI	0.14844	17.956	7.066	45.310	3.299
NITAlevel	0.13205	0.004	0.006	0.076	-0.060
T(CDSNonFin)	0.12971	-0.174	0.739	1.959	-1.558
T(CDSHY)	0.12867	2.913	1.581	7.655	0.327
T(CDSIG)	0.12800	-0.339	0.704	1.631	-1.772
T(CDSFin)	0.12534	0.671	1.521	4.608	-2.116
CountryCCI	0.11268	21.654	11.801	177.497	0.439
VIX	0.09923	21.693	9.611	59.890	10.420
SIZEtrend	0.07264	-0.004	0.179	1.645	-1.896
TLTA	0.06461	0.669	0.179	2.032	0.121
T(AS)trend	0.02692	-0.187	1.454	120.101	-31.358
DTDtrend	0.01910	0.109	1.356	6.135	-7.047
3mRateUS	0.01886	0.677	1.303	5.124	-0.020
Tenor1y	0.01708	0.180	0.384	1.000	0.000
preCrisis	0.01643	0.212	0.409	1.000	0.000
postCrisis	0.01643	0.788	0.409	1.000	0.000
Tenor5y	0.01332	0.341	0.474	1.000	0.000
isFin	0.00975	0.147	0.354	1.000	0.000
isNonFin	0.00975	0.853	0.354	1.000	0.000
NITAtrend	0.00422	0.000	0.007	0.104	-0.146
3mRateEcon	0.00383	0.958	1.740	23.770	-0.080
SwapSpread5vs1	0.00307	1.108	0.636	2.518	-0.356
Tenor4y	0.00228	0.146	0.353	1.000	0.000
CASHTAlevel	0.00196	0.091	0.109	0.979	0.000
isSub	0.00109	0.001	0.025	1.000	0.000
isSenior	0.00109	0.999	0.025	1.000	0.000
preBigBang	0.00099	0.285	0.451	1.000	0.000
postBigBang	0.00099	0.715	0.451	1.000	0.000
Tenor2y	0.00085	0.144	0.351	1.000	0.000
Tenor3y	0.00080	0.189	0.391	1.000	0.000
CASHTAtrend	0.00068	0.001	0.029	0.483	-0.334
isUS	0.00004	0.858	0.349	1.000	0.000
isNonUS	0.00004	0.142	0.349	1.000	0.000

Note: $T(\cdot)$ denotes a transformation: $\ln(\exp(\cdot/100) - 1)$.

Before proceeding further, one needs to note that several dummy variables are used to indicate the characteristics of individual CDS or the CDS environment; for the CDS Big Bang, for example, we have adopted two dummy variables: preBigBang and postBigBang. A quick reaction is that we have unnecessarily added one extra dummy variable, which should result in an identification problem in the subsequent linear regression analysis. However, it is not the case when interaction terms are included in a variable selection context. This is because a particular variable combination, say, 10 variables, involving one particular dummy may not be replicable by its complementary dummy subject to the constraint of having 10 variables in a combination. The argument also applies to five CDS tenors for which we create five dummy variables instead of four per the usual practice in linear regressions.

2.2 The variables used to predict CDS curves

Variables that capture financial conditions of individual corporate obligors and reflect general credit environment are natural candidates for predicting CDS premiums. With the availability of the Credit Research Initiative (CRI) database at the National University of Singapore, a CDS-like credit risk measure, known as actuarial spread (AS), constructed with physical default probability (PD) term structure is readily available on a daily basis on all exchange-listed firms worldwide. Also available on the CRI database are (1) a suite of daily series of credit cycle indices (CCIs) capable of reflecting the credit environment in general and/or for different industries, and (2) distance-to-default (DTD) estimates for individual firms which is loosely speaking an asset volatility adjusted leverage measure. In the following, we will briefly describe the CRI database, AS, CCI and DTD.

The CRI, launched in 2009 in response to the 2008 global financial crisis, was conceived as a *public good* endeavor to contribute to credit rating reform (see Duan and van Laere, 2012). The CRI has been making its PDs and other credit risk measures freely accessible since day one. The CRI-PDs are computed with the forward-intensity model of Duan, *et al* (2012), which was designed for obtaining the PD term structure while factoring in the censoring effect arising from other corporate exits such as M&As. The CRI coverage includes virtually all exchange-listed firms globally, and its PDs (1 month to 5 years) and ASes (1 year to 5 years) are updated daily on over 36,000 currently active firms. Historical time series are also available on 70,000 plus firms including those delisted corporates due to bankruptcies, M&As and other reasons.[2]

The PD term structure is useful for many applications. A 5-year CDS is, for example, a sort of complex average of PDs over the life of the contract mixed with recovery rate, risk premium demanded by market participants, market liquidity, and potential counterparty default. Duan (2014) showed how AS can be constructed from the PD term structure to mimic CDS of any tenor except for leaving out risk premium, market liquidity and counterparty default. In short, AS is the actuarial value of CDS, which is in principle the closest risk measure to CDS without committing to a specific CDS pricing model. Since the CRI also makes freely available the daily updated ASes (1- to 5-year tenors) on all exchange-listed firms globally, AS becomes our natural candidate for predicting CDS.[3] This choice is clearly supported by the individual R^2 reported later where $\ln(\exp(\text{AS}/100) - 1)$ is shown to have an R^2 of 35.7% on a single-regressor basis in explaining $\ln(\exp(\text{CDS}/100) - 1)$.

In addition, we also consider its transforms – the level and trend of $\ln(\exp(\text{AS}/100) - 1)$. The 12-month moving average of AS plugged in the transform is $\ln(\exp(\text{ASlevel}/100) - 1)$ whereas the difference between $\ln(\exp(\text{AS}/100) - 1)$ and $\ln(\exp(\text{ASlevel}/100) - 1)$ is treated as the trend. These three variables are obviously linearly dependent by design, but we include all three in the set of 42 potential regressors. Choosing a subset of regressors subject to a zero-norm penalty will naturally avoid picking all three, because having all three does not increase explanatory power but only adds to the penalty.

In Duan and Miao (2016), a suite of credit cycle indices (CCIs) were used to describe the credit environment. The country CCI at a particular time point is in our deployment the median AS value for a corresponding tenor where the median is taken over all exchange-listed firms domiciled in that country at that time point. Likewise, industry CCIs are the median ASes for the 10 industries globally according to the Bloomberg Industry Classification System. Our CCIs differ from those of Duan and Miao (2016) in two aspects. First, we use AS in stead of PD because our interest is on CDS where the AS has been constructed with the CDS convention in mind. Second, we use the original median series instead of further

[2]For the technical details on how these PDs and ASes are computed, readers are referred to NUS Credit Research Initiative Technical Report Version: 2020 Update 1 available at http://www.rmicri.org.

[3]The CRI ASes are computed by default with the 40% recovery rate. We adjust those ASes for CDS referencing subordinate debt by lowering the recovery rate to 20%.

subjecting 10 industry CCIs to orthogonalization. In our case, the CCIs are simply used as regressors and the correlations among the CCIs do not affect our regressor selection because the selection technique deployed relies on the zero-norm penalty. Naturally, CDS pricing is expected to reflect the credit environment in general as well as those in different industries, and CCIs are simply used as credit environment indicators.

We have also added four aggregates of CDS spreads to reflect general CDS market conditions in different segments of the CDS market. Specifically, we define CDSindexIG, CDSindexHY, CDSindexFin, and CDSindexNonFin to be the median CDS par spread in each of four CDS categories: investment-grade, high-yield, financial and non-financial reference corporate names. In this construction, we require five as the minimum number of traded CDS with liquidity in each category on the day. Otherwise, it will be tagged as a missing value. The single-variable regression results suggest that each of these four CDS indices provides an R^2 in excess of 12%.

DTD based on the structural credit risk model of Merton (1977) is a commonly adopted measure in credit analysis. Although the concept is standard, its implementation can be challenging due to the fact that the underlying firm asset value in the call option like theoretical setup of Merton (1977) is a latent stochastic process. The Moody's KMV approach has been widely adopted in both academic and commercial applications, which relies on an iterative scheme to estimate the unknown model parameters, the latent firm asset value, and finally the DTD. However, the Moody's KMV approach has statistical shortcomings because it fails to properly account for the Jacobian arising from the call option pricing function, and thus causes some biases. The Moody's KMV approach also specifies a default point formula (100% short-term debt plus 50% long-term debt), which serves as the strike price in the call option analogy. Interestingly, the missing Jacobian also places an implementation limitation whenever the default point formula justifiably needs an expansion to include other liabilities subject to an unknown haircut. Adding to the Moody's KMV default point formula is evidently important for financial institutions where a large portion of corporate liabilities is classified as neither short-term nor long-term; for example, deposits of a bank and policy obligations of an insurance company. The CRI database generates DTDs as per Duan, *et al* (2012) to accommodate other corporate liabilities.[4]

[4]Readers who are interested in technical details are referred to Duan, *et al* (2012) and/or NUS Credit

In addition, we consider common drivers such as interest rate, interest rate term spread and VIX, and individual firm attributes like funding liquidity, leverage, profitability, size, and idiosyncratic equity return volatility (SIGMA). These variables along with the categorical CDS characteristics described earlier are summarized in Table 1. Also reported in Table 1 are the individual R^2 when each of these variables is used as a single regressor along with an intercept term. The results suggest that SIGMA is the best single predictor with an R^2 of 43.8% and closely followed by two AS-based measures at 36% and 35.7%, respectively. Many of these 42 potential regressors have a decent individual R^2. It is worth noting that both the categorical CDS characteristic like isSub (equals 1 if the CDS references a subordinated debt) or the CDS market structural change captured by postBigBang (equals 1 for the period post the April 2009 CDS Big Bang) has a minuscule R^2, but our later results show that postBigBang still plays a meaningful role through interacting with another variable. In the case of isSub, we force it to be a chosen regressor for otherwise its explanatory power will simply be dwarfed due to its relatively minuscule sample size. Table 2 provides correlations among some selected regressors. It is clear from this table that some regressors are highly correlated, for example, $\ln(\exp(\text{AS}/100)-1)$ and $\ln(\exp(\text{ASlevel}/100)-1)$. The CDS indices are also highly correlated among themselves.

2.3 Linear regression subject to a zero-norm penalty

In a general classical linear regression setting, one attempts to relate a dependent variable to k regressors where there are n data points. When there are too many potential regressors vis-a-vis the number of data points, in-sample over-fitting is expected and removing some regressors becomes both conceptually sensible and practically necessary. There has been a long-standing interest in designing theoretically sound and practically implementable methods for selecting regressors. In order to have a more concrete discussion, we state the regressor selection problem as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)'$, and $\boldsymbol{X}$ denotes the n observations of k regressors, i.e., $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)$ with $\boldsymbol{x}_i = (x_1, \cdots, x_n)'$, of which the first vector may represent the intercept

Table 2: Correlations for a subset of regressors

Regressor	T(AS)	3mRate Econ	SwapSpread 5vs1	DTDlevel	SIGMA	T(CDSHY)	T(CDSFin)
T(AS)	1.000	-0.002	0.096	-0.675	0.554	0.279	0.257
T(ASlevel)	0.825	-0.032	0.160	-0.793	0.643	0.228	0.228
T(AS)trend	0.542	0.043	-0.068	-0.014	0.022	0.154	0.116
3mRateEcon	-0.002	1.000	-0.385	0.028	-0.068	-0.217	-0.331
3mRateUS	-0.014	0.749	-0.526	0.096	-0.084	-0.315	-0.453
SwapSpread5vs1	0.096	-0.385	1.000	-0.226	0.213	0.192	0.235
VIX	0.241	-0.154	0.143	-0.251	0.332	0.864	0.845
TLTA	0.268	-0.002	0.000	-0.316	0.188	-0.005	-0.008
CountryCCI	0.429	-0.091	0.320	-0.370	0.410	0.537	0.536
IndustryCCI	0.484	0.009	0.194	-0.428	0.438	0.605	0.570
CASHTAlevel	0.032	-0.037	-0.010	-0.070	0.104	-0.037	-0.026
CASHTAtrend	-0.001	-0.005	0.083	-0.040	0.078	0.025	0.033
NITAlevel	-0.375	0.052	-0.096	0.485	-0.426	-0.054	-0.056
NITAtrend	-0.050	-0.003	0.040	-0.012	0.009	-0.068	-0.059
SIZElevel	-0.317	0.301	-0.047	0.406	-0.429	-0.084	-0.115
SIZEtrend	-0.293	-0.020	0.097	0.015	-0.005	-0.085	-0.057
DTDlevel	-0.675	0.028	-0.226	1.000	-0.594	-0.315	-0.324
DTDtrend	-0.212	-0.079	0.198	-0.052	-0.166	-0.370	-0.297
SIGMA	0.554	-0.068	0.213	-0.594	1.000	0.389	0.394
isUS	0.020	-0.473	0.043	0.028	0.023	0.020	0.038
isNonUS	-0.020	0.473	-0.043	-0.028	-0.023	-0.020	-0.038
isFin	0.183	0.096	-0.016	-0.248	0.068	-0.023	-0.029
isNonFin	-0.183	-0.096	0.016	0.248	-0.068	0.023	0.029
Tenor1y	-0.265	0.020	0.000	0.013	-0.021	0.008	0.022
Tenor2y	-0.047	-0.143	0.060	-0.005	0.012	0.061	0.100
Tenor3y	0.012	0.024	-0.004	0.020	-0.024	-0.002	0.010
Tenor4y	0.070	-0.166	0.064	0.005	0.010	0.057	0.097
Tenor5y	0.188	0.193	-0.089	-0.028	0.021	-0.093	-0.172
isHY	0.313	-0.022	0.012	-0.387	0.420	0.052	0.062
isIG	-0.313	0.022	-0.012	0.387	-0.420	-0.052	-0.062
isSub	0.034	-0.010	-0.009	-0.042	0.019	0.001	0.001
isSenior	-0.034	0.010	0.009	0.042	-0.019	-0.001	-0.001
preCrisis	-0.004	0.664	-0.300	0.073	-0.079	-0.253	-0.443
postCrisis	0.004	-0.664	0.300	-0.073	0.079	0.253	0.443
preBigBang	0.094	0.571	-0.295	0.002	0.069	0.145	-0.047
postBigBang	-0.094	-0.571	0.295	-0.002	-0.069	-0.145	0.047
T(CDSindexHY)	0.279	-0.217	0.192	-0.315	0.389	1.000	0.936
T(CDSindexIG)	0.258	-0.309	0.172	-0.317	0.344	0.944	0.948
T(CDSindexFin)	0.257	-0.331	0.235	-0.324	0.394	0.936	1.000
T(CDSindexNonFin)	0.261	-0.296	0.135	-0.306	0.343	0.939	0.916

Note: $T(\cdot)$ denotes a transformation: $\ln(\exp(\cdot/100) - 1)$.

term. $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)'$ is the k-dimensional regression coefficients, and $\boldsymbol{\epsilon}$ is n-dimensional $i.i.d.$ errors with mean 0 and variance σ^2 that may or may not be normally distributed. The task is to select $k_s \leq k$ regressors meeting some criterion, or alternatively, to set some β's to zero.

The classic way of performing such a task is a greedy-search technique that starts with one regressor with the highest R^2, finds the second regressor that delivers the highest R^2 in explaining the residual from the best one-regressor model, and then repeats the search sequentially until the stopping criterion is reached. The greedy-search technique is known to be suboptimal because a combination of, say, two regressors may deliver a better predictive power while they individually do not produce top explanatory power. In principle, one could exhaust all possible combinations to find the ideal subset of k_s regressors. Practically speaking, however, it is not feasible when the number of potential regressors is large. When k_s is unknown, there will be 2^k potential models, and with 810 potential regressors in this paper, for example, it creates $2^{810} = 6.8 \times 10^{243}$ potential models. As the later result shows, we end up selecting 20 regressors out of 810 potential explanatory variables, which means 4.8×10^{39} possible combinations for a 20-variable model.

A popular modern way of performing regressor selection is through an l_1-norm penalty, commonly known as Lasso, advanced by Tibshirani (1996) and subsequently improved by, for example, SCAD of Fan (1997) and Fan and Li (2001), and adaptive Lasso of Zou (2006). The Lasso and its variants have found great popularity in big-data applications these days due to their simplicity and computational efficiency. However, regressor selection based on the l_1-norm penalty is not most conceptually appealing albeit its practicality. This is because regression coefficients will be distorted by the penalty term (i.e., shrinkage toward zero even being selected). Even though the SCAD and adaptive Lasso do have the oracle property[5], i.e, distortion disappears when the sample size approaches infinity, it is mostly a feature that bears limited practical relevance because in applications the sample size vis-a-vis the number of regressors is unlikely large enough. A more practical concern is perhaps the issue of multicollinearity which analysts inevitably encounter in practice. To understand this point, let us rotate a group of mutually independent regressors to become linearly dependent regressors, knowing that such rotation will not alter the regression model fit, measured by

[5] See Fan and Li (2001) for a formal definition of the oracle property.

R^2. However, the l_1 norm of the regression coefficients is not invariant to a rotation, and hence the rotation will change the model's l_1 penalty, giving rise to a different penalized estimation outcome. In short, multicollinearity may lead to an undesirable variable selection outcome when an l_1-norm based method is deployed. This concern is not a purely theoretical possibility, because the situation repeatedly occurs in many practical applications and the simulation study in Duan (2019) has also clearly demonstrated that the adaptive Lasso greatly over-selects variables.

In principle and probably without contention, a more appealing and direct approach to regressor selection is to pick a fixed number of regressors, say, k_s, where the selection is optimally conducted by minimizing squared residual errors, i.e., the l_2 norm, over all possible combinations. As to k_s, it can be determined, for example, by applying cross validation or the BIC. Such a variable selection approach is known as applying a zero-norm penalty, a term commonly used in scientific computing and big-data analytics. It is viewed as the zero norm because the standard l_p norm approaches this zero-norm when p goes to zero even though such a limiting l_0 "norm" is not a proper norm because of its lack of homogeneity. The penalized regression subject to the zero-norm regularization can be formally stated as

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_{l_2}^2 \\ &s.t. \;\; ||\boldsymbol{\beta}||_{l_0} \leq k_s \leq k \end{aligned} \tag{2}$$

where $||\cdot||_{l_2}$ is the l_2-norm and $||\cdot||_{l_0}$ is the zero-norm, which counts the number of non-zero entries in $\boldsymbol{\beta}$. Also worth noting is the fact that the above minimization problem is equivalent to $\arg\min_{\boldsymbol{\beta}} \left\{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_{l_2}^2 + \lambda||\boldsymbol{\beta}||_{l_0}\right\}$ where the solution is a step function of λ with the jumps corresponding to different values of k_s. This zero-norm penalized regression problem is known to be NP-hard. But the benefit is that this variable selection approach is free of the distortion caused by interference of the l_2-norm objective with the l_1-norm penalty in the presence of multicollinearity. What preventing its adoption in practice is the computational challenge in dealing with extremely large possible combinations that we alluded to earlier. Typical solutions are by approximating the l_0 norm with a penalty function very close to it, for example, Dicker, *et al* (2013). Here we are able to implement the zero-norm regressor selection without approximating the penalty function by deploying the sequential Monte Carlo method developed by Duan (2019).

2.4 The selected proxy CDS model

We apply a proprietary software made available by CriAT, a FinTech company, based on Duan's (2019) sequential Monte Carlo variable selection method. Selection is performed on a randomly selected training subsample of 94,499 CDS observations (two-thirds of the whole sample) along with their attributes mentioned earlier. Once the optimal combination under a k_s is identified, we repeat the selection for different values of k_s and use a five-fold cross-validation to determine the optimal k_s. In the end, we obtain the optimal zero-norm solution as 20 regressors out of 810 potential variables (including all meaningful interaction terms).[6] In order to ensure that the CDS referencing subordinated debt are not overwhelmed in competition with other CDS due to their small sample size (61 in the training sample), the isSub dummy variable is always included in the model, which means we have actually chosen 19 regressors to come up with the 20-variable proxy CDS model. Note that the intercept term is treated as a potential regressor and not chosen in the final 20-variable model.

Table 3 lists the 20 selected variables using the training sample of 94,499 data points via a five-fold cross validation. This model yields an R^2 of 73.98%. We have argued earlier that AS is a variable conceptually closest to its corresponding CDS, and the selection result is consistent with this intuition. Thus, it is natural to see $\ln(\exp(\text{AS}/100) - 1)$ and its close substitute $\ln(\exp(\text{ASlevel}/100) - 1)$ to show up through interaction terms.

Two of the four CDS indices (CDSindexFin, CDSindexNonFin) appear in the selected variables. Intuitively, this can be expected for they characterize the CDS market conditions in a way like stock market indices. The single-variable regression results reported earlier also suggest a high likelihood for them to show up along other variables. Indeed, they form a prominent group of variables in the final model. One can, for example, interpret the selection result as $\ln(\exp(\text{CDSindexFin}/100) - 1)$ influences CDS values through a variable coefficient equal to $(0.5694 \times \text{TLTA} - 0.3138 \times \text{SwapSpred5vs1})$, reflecting different levels of response depending on whether the reference name's book leverage ratio and the market's prevailing

[6] With the 42 explanatory variables, there are 43 potential regressors after including the intercept term. If all interaction terms are considered, the maximum number of potential regressors is increased to 946 $(= 43 \times 44/2)$. However, some of the terms are redundant when the intercept and/or a dummy variable is involved; for example, squaring a dummy variable yields exactly the same dummy variable, and the product of the intercept with the 42 original variables produces the same set of 42 variables. After trimming the redundant regressors, the total count drops to 810.

swap rate spread (5-year vs. 1-year).

The CDS value also responds to SIGMA, which is the corporate equity's idiosyncratic volatility. Again, this is totaly anticipated because the single-regressor result reported earlier in Table 1 suggests it to be the regressor with the highest explanatory power. The impact of SIGMA on CDS values again is through many interaction terms with some being common and others being firm-specific. In addition, accounting ratios such as TLTA (total liabilities over total assets) and NITAlevel (net income over total assets's 12-month moving average) enter the predictive equation. Other firm-specific measures such as SIZE (both level and trend), or common factors such as industry and country CCIs (credit cycle indices), interest rates (the domicile country), swap rate spread and VIX are also relevant to the CDS prediction. The results also suggest that the post-CDS Big Bang dummy can help predict CDS values but the 2008 global financial crisis does not make an additional contribution. The CDS value is apparently also influenced by when the reference corporate is a US firm and whether it is investment grade rated by credit rating agencies.

Worth noting is the fact that tenor5y is able to add significant predictive power to the proxy CDS model albeit the fact that the AS is a tenor-specific CDS-like credit risk measure, suggesting that the premium for a five-year CDS will be relatively higher given other aspects are equal. As stated earlier, the subdebt dummy is forced upon the selected model due to its minuscule sample size of 61 vis-a-vis the total sample size of 94,499. Doing so is to avoid the inevitable outcome that the isSub dummy would be competed away by other variables even though it is a highly significant regressor as suggested by its estimate reported in Table 3.

3 The out-of-sample performance of the proxy CDS model

We now apply the optimally selected subset of regressors on the holdout sample of 47,250 observations to study the proxy CDS model's out-of-sample performance. As reported in Table 4, our 20-variable proxy CDS model has an in-sample R^2 of 73.98% on the training sample of 94,499 observations when the target is the transformed CDS value. However, the real usage is concerned with the CDS value in its original form. After converting back to predict CDS directly, the in-sample R^2 drops slightly to 72.11%. The model's performance, measured by the out-of-sample R^2 on the holdout sample, are 73.79% for the transformed CDS and 71.78% for the original CDS, respectively, which are only slightly lower than their respective in-sample R^2.

To examine whether the proxy CDS model exhibits any bias behavior for different subgroups, we present a set of plots in Figure 1. However, we first plot the in-sample predicted vs. realized values for CDS and transformed CDS (top two plots) based on the training sample of 94,499 observations, which is used to gauge the potential distortion caused by the transformation. As these two plots reveal, the proxy CDS model built on predicting the transformed CDS generates a similar result for the original CDS values. The R^2 displayed on the plot should be understood as the *ex post* best linear relationship between the observed CDS value (vertical axis) and the proxy CDS value (horizontal axis) when the relationship is computed for a particular group. Such an R^2 will naturally be higher than its corresponding R^2 reported in Table 4, because an additional intercept and a flexible slope are introduced. In other words, those R^2 in Table 4 can be understood as setting the intercept to 0 and the slope to 1.

The out-of-sample results for the whole and various subgroups of the holdout sample (47,250 observations for 405 corporates with five tenors over the entire sample period on a monthly frequency) are reported in Table 4 and Figure 1. Among all subcategories, the proxy CDS model has the worst out-of-sample performance for CDS referencing subordinate debt with an R^2 at 61.14%. Despite the fact that the corresponding plot in Figure 1 seems to show good performance at an R^2 of about 80%, one should reference the R^2 in Table 4 for the subdebt category as opposed to the R^2 on the plot, and the reason was given earlier. The in-sample R^2 for the subdebt group equals 62.74%, at about the same as the out-of-sample R^2, suggesting a rather stable performance even for the subdebt category. Note that this subordinate debt group only has 92 observations in total and all for Shinshei Bank over the entire sample period. The training sample contains 61 subdebt data points, and the cross-validation would not have selected the subdebt dummy if we did not force it upon the model. The subdebt dummy interacting with other variables might improve the performance of the model for this subcategory of CDS if some interactive terms were also forced upon the model. Failure to do so, those potential interactive terms would be crowded out by the need to better fit a disproportionably larger number of CDS that reference senior debt. If more CDS referencing subordinate debts are available, we expect to improve the proxy CDS model's performance for this subcategory. Apart from the subdebt group, the results in Table 4 and the plots in Figure 1 suggest some performance difference for CDS of different tenors, but the better out-of-sample result for the 5-year tenor vis-a-vis the 1-year tenor is not particularly large. It is worth to re-emphasize that the proxy CDS model serves as a universal predictor across different tenors and other attributes. The fact that it delivers comparable performance for different tenors reflects the benefit of using AS as a predictor, which by design captures credit quality specific to a reference obligor over different tenors.

Further comparisons (US vs. non-US and financial vs. non-financial reference obligors) are also shown in Table 4 and Figure 1, and their performances do not differ too much, even though the performance favors CDS referencing US and financial obligors. We also compare the data before and after the 2008 global financial crisis with the post-crisis period defined as starting from the end of September 2008. Again, one cannot find a large performance gap pertaining to the potential structural break induced by the global financial crisis. Finally, we find the predictive regression works marginally better out-of-sample post the April 2009 CDS Big Bang.

4 Concluding Remarks

We have developed a proxy CDS model that can robustly predict CDS par spreads for corporate obligors which do not have traded or liquid CDS contracts. This predictive model has many applications for credit risk management in general and accounting practice in particular. Our approach appears to be entirely empirical, but actually utilizes a critical theoretical result in connection with the actuarial spread model of Duan (2014), and it is this variable that makes the predictive model successful. If one can develop a high-quality theoretical pricing model for CDS (incorporating risk premium due to risk aversion and/or factoring in counterparty default) and implement the model solely using equity prices, such a measurement may then serve as a better predictor for CDS. Even with a good CDS pricing model in place, practical usage likely still needs further empirical tuning in a way similar to our approach.

Our empirical model can be viewed as a concrete demonstration of modern big-data analytics in action. Critical to our success is the zero-norm penalty variable selection technique, which enables us to identify 20 regressors among 810 potential variables (including interaction terms) arising from 42 original variables. Although the 42 original variables are chosen for their data availability and based on economic intuition, identifying the optimal set of regressors in light of the astronomical number of possible combinations would not have been possible without such a big-data analytical tool. Obviously, many other financial applications may also adopt a similar approach.

References

[1] Blanco R., S. Brennan and I.W. Marsh, 2005, An Empirical Analysis of the Dynamic Relationship between Investment-Grade Bonds and Credit Default Swaps, *Journal of Finance* 60, 2255-2281.

[2] Das, S.R., and R.K. Sundaram, 2000, A Discrete-Time Approach to Arbitrage-Free Pricing of Credit Derivatives, *Management Science* 46, 46-62.

[3] Dicker, L., B. Huang, and X. Lin, 2013, Variable Selection and Estimation with the Seemless-L_0 Penalty, *Statistica Sinica* 23, 929-962.

[4] Duan, J.C., 2019, Variable Selection with Big Data based on Zero Norm and via Sequential Monte Carlo, National University of Singapore working paper.

[5] Duan, J.C., 2014, Actuarial Par Spread and Empirical Pricing of CDS by Decomposition, *Global Credit Review* 4, 51-65.

[6] Duan, J.C., and W. Miao, 2016, Default Correlations and Large-Portfolio Credit Analysis, *Journal of Business and Economic Statistics* 34(4), 536-546.

[7] Duan, J.C., J. Sun and T. Wang, 2012, Multiperiod Corporate Default Prediction – A Forward Intensity Approach, *Journal of Econometrics* 170(1), 191-209.

[8] Duan, J.C., and E. van Laere, 2012, A Public Good Approach to Credit Ratings – from Concept to Reality, *Journal of Banking and Finance* 36(12), 3239-3247.

[9] Duffie, D., and D. Lando, 2000, Term Structures of Credit Spreads with Incomplete Accounting Information, *Econometrica* 69, 633-664.

[10] Duffie, D., and K.J. Singleton, 1999, Modeling term Structures of Defaultable Bonds, *Review of Financial Studies* 12, 687-720.

[11] Ericsson, J., K. Jacobs, and R. Oviedo, 2009, The Determinants of Credit Default Swap Premia, *Journal of Financial and Quantitative Analysis* 44(1), 109-132.

[12] Fan, J., 1997, Comments on "Wavelets in Statistics: a Review" by A. Antoniadis, *Journal of the Italian Statistical Society* 6(20, 131-138.

[13] Fan, J. and R. Li, 2001, Variable Selectionvia Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, 96, 1348-1360.

[14] Hull, J.C., and A.D. White, 2000, Valuing Credit Default Swaps I: No Counterparty Default Risk, *Journal of Derivatives* 8, 29-40.

[15] Kim, G.H., H. Li, and W. Zhang, 2017, The CDS-Bond Basis Arbitrage and the Cross Section of Corporate Bond Returns, *Journal of Futures Markets* 37(8), 836-861.

[16] Longstaff, F.A., and E.S. Schwartz, 1995, A Simple Approach to Valuing Risky Fixed and Floating Rate Debt, *Journal of Finance* 50, 789-819.

[17] Merton, R., 1974, On the Pricing of Corporate Debt: the Risk Structure of Interest Rates, *Journal of Finance* 29, 449-470.

[18] NUS-RMI Staff, 2020, NUS Credit Research Initiative Technical Report Version: 2020 Update 1, The Credit Research Initiative at the National University of Singapore (http://rmicri.org).

[19] Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Ser. B*, 58(1), 267-288.

[20] Zhu, H., 2006, An Empirical Comparison of Credit Spreads between the Bond Market and the Credit Default Swap Market, *Journal of Financial Services Research* 29, 211-235.

[21] Zou, H., 2006, The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, 101(476), 1418-1429.

Table 3: Selection of 20 regressors out of 810 potential variables (interaction terms included) using a training sample of 94,499 observations via a five-fold cross validation

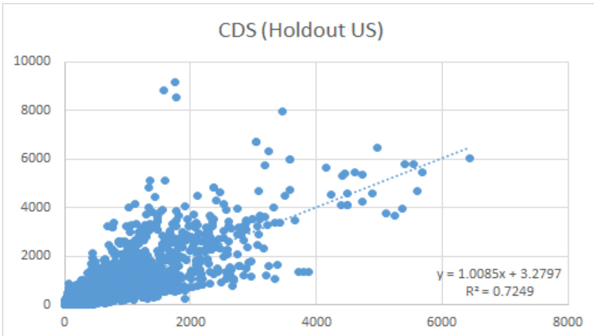
	Training Sample		
Regressor	**Estimate**	**Std. Error**	***t*-Stat**
T(AS)*VIX	0.012532	0.000159	78.62
T(ASlevel)*isHY	0.371620	0.007843	47.38
T(AS)trend*SIZEtrend	0.304932	0.005852	52.11
T(AS)trend*isIG	-0.260963	0.007547	-34.58
IndustryCCI*isHY	0.073927	0.001166	63.41
IndustryCCI*postBigBang	0.037993	0.000739	51.41
CountryCCI*3mRateEcon	0.007591	0.000140	54.05
T(CDSindexFin)*TLTA	0.569360	0.014543	39.15
T(CDSindexFin)*SwapSpread5vs1	-0.313825	0.006405	-49.00
SIGMA*T(CDSindexNonFin)	7.841172	0.133864	58.58
SIGMA*TLTA	26.563192	0.196056	135.49
SIGMA*SIZElevel	-2.825277	0.059294	-47.65
SIGMA*SIZEtrend	-17.351329	0.212121	-81.80
SIZElevel*VIX	0.009153	0.000256	35.79
SIZElevel*TLTA	-0.462154	0.007116	-64.94
SIZEtrend*NITAlevel	121.801512	4.122555	29.55
SIZEtrend^2	2.609726	0.069613	37.49
isUS*isIG	-0.392691	0.014613	-26.87
Tenor5y	0.473201	0.013663	34.63
isSub	2.390254	0.241578	9.89
R^2	73.98%		
Sample Size	94,499		

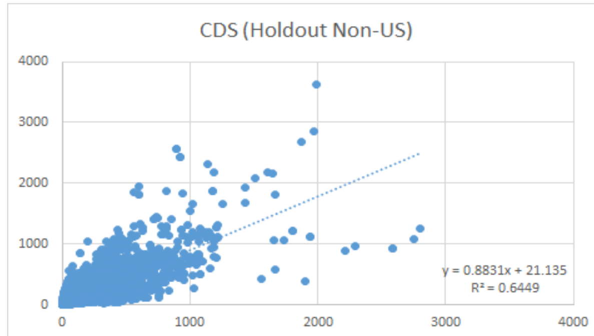
(1) $T(\cdot)$ denotes a transformation: $\ln(\exp(\cdot/100) - 1)$;
(2) isSub is always included because its marginal R^2 contribution would be too small to be chosen due to the minuscule sample size of 61.

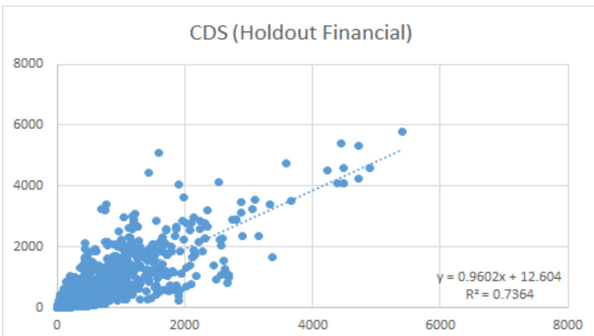
Table 4: R^2 of the proxy CDS model for the training and holdout samples as well as for various subcategories

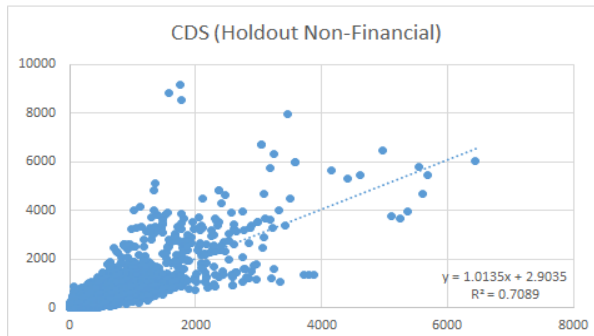
	Training Sample			**Holdout Sample**		
	Transformed CDS	**CDS**	**Obs**	**Transformed CDS**	**CDS**	**Obs**
Total	73.98%	72.11%	94499	73.79%	71.78%	47250
Subdebt	64.76%	62.74%	61	61.98%	61.14%	31
1-Year	69.24%	67.06%	17030	72.18%	70.56%	8518
5-Year	72.50%	70.93%	32271	74.78%	73.53%	16008
Investment Grade	69.73%	68.77%	78937	68.80%	67.46%	39464
High Yield	65.93%	65.53%	15562	65.22%	64.79%	7786
US	74.14%	72.31%	81159	74.36%	72.47%	40515
Non-US	72.22%	69.76%	13340	67.84%	63.31%	6735
Financial	75.13%	73.50%	13829	75.29%	73.50%	6949
Non-Financial	73.17%	71.17%	80670	72.99%	70.86%	40301
Pre-Crisis	68.66%	63.85%	19954	72.04%	68.94%	10089
Post-Crisis	73.99%	72.38%	74545	73.46%	71.68%	37161
Pre-Big Bang	74.72%	73.03%	26869	72.61%	70.44%	13491
Post-Big Bang	73.16%	70.92%	67630	74.92%	73.13%	33759

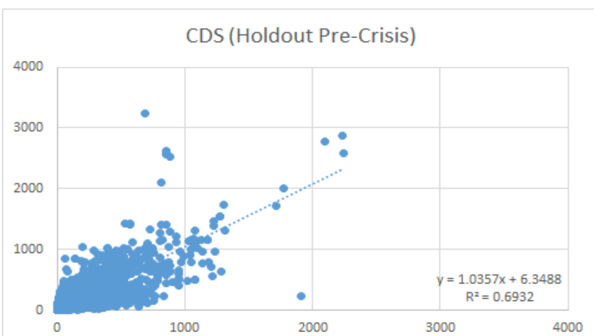
Figure 1: Performance of the proxy CDS model in predicting CDS for the training and holdout samples as well as various subcategories

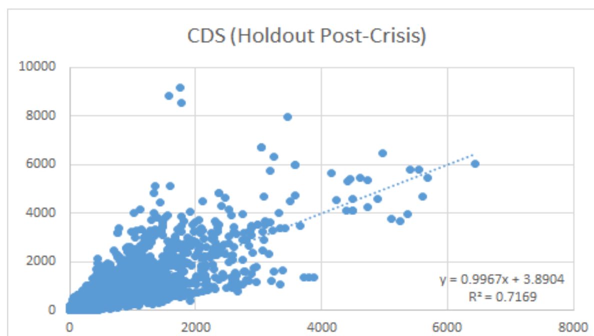

CDS (Holdout US)
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
y = 1.0085x + 3.2797
R² = 0.7249

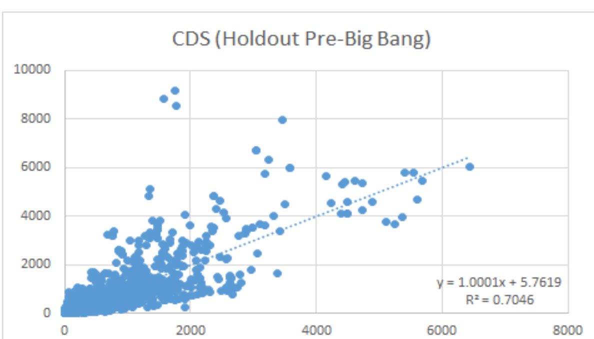

CDS (Holdout Non-US)
4000
3000
2000
1000
0
0
1000
2000
3000
4000
y = 0.8831x + 21.135
R² = 0.6449

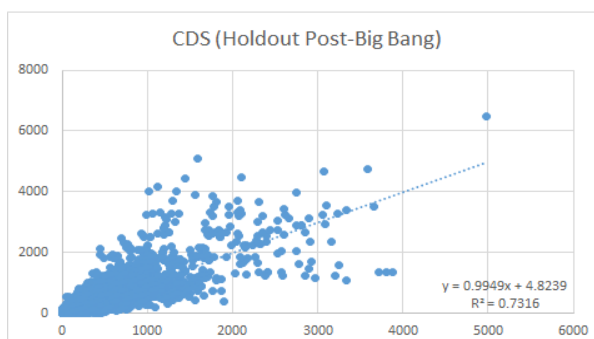

CDS (Holdout Financial)
8000
6000
4000
2000
0
0
2000
4000
6000
8000
y = 0.9602x + 12.604
R² = 0.7364


CDS (Holdout Non-Financial)
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
y = 1.0135x + 2.9035
R² = 0.7089


CDS (Holdout Pre-Crisis)
4000
3000
2000
1000
0
0
1000
2000
3000
4000
y = 1.0357x + 6.3488
R² = 0.6932


CDS (Holdout Post-Crisis)
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
y = 0.9967x + 3.8904
R² = 0.7169


CDS (Holdout Pre-Big Bang)
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
y = 1.0001x + 5.7619
R² = 0.7046


CDS (Holdout Post-Big Bang)
8000
6000
4000
2000
0
0
1000
2000
3000
4000
5000
6000
y = 0.9949x + 4.8239
R² = 0.7316