FORECASTING OPTION-IMPLIED VOLATILITY USING CREDIT RISK

Laleh Samarbakhsh*, Madhu Kalimipalli

Current Draft:
October 31, 2018

Affiliations: Samarbakhsh is from Ted Rogers School of Business Management at Ryerson University, Toronto, Canada. Kalimipalli is from Lazaridis School of Business and Economics, Wilfrid Laurier University, Waterloo, Canada. *Contact author e-mail address: lsamarbakhsh@ryerson.ca
Abstract

We study the time series properties of credit default swap spread in enhancing the forecasting power of future implied volatility and hence future option prices. We run a series of contemporaneous and forward-looking models as encompassing regressions and test whether we can improve the predictability power of implied volatility by extending the forecast model that includes historical volatility (HV) and current implied volatility (IV) to enhance credit default swap. Both in-sample and out-of-sample estimation errors show that CDS can be a significant factor in improving forecasting performance of implied volatility.

Keywords: CDS Spread, Implied Volatility, Encompassing Regressions, Out-of-sample performance
1 Introduction

The need for enabling users to manage credit risk has made credit derivatives one of the most important financial innovations of the last quarter of a century. The volume of the credit default swap market skyrocketed in the last ten years after the first CDS contract was issued by JP Morgan in mid 1990s, making it one of the fastest growing over the counter (OTC) derivatives markets during this period. The CDS contract is a mean of protection against bond default. The seller of the contract is committed to remunerate the buyer with the face value of the bond upon default of the debtor. The buyer in turn has to pay the certain amount as an insurance premium to the seller. This insurance premium is determined as a percentage of the bond’s face value and usually named as CDS spread.

Similar to put options, CDS is used for downside risk protection. From the hedging point of view, the insurance and risk protection similarities of the put options and CDS is further shown by the comovement of CDS spread and option implied volatility (See Figure 1). This study will contribute to the existing literature by investigating the time-series properties of the CDS when augmented as an enhancement to the future implied volatility forecast models.

\[^1\text{See Greenspan [2004]}\]
Forecasting option Implied Volatility (IV) is of interest to option market participants, who routinely formulate volatility and option price forecasts for trading and hedging purposes. Given that the IV is a re-parameterization of the market option price, forecasting IV falls within the vast literature on the predictability of asset prices. In addition, to the extent IV yields a forward-looking measure of firm’s total risk it has applications predicting firm’s beta, credit risk, etc.

The extant literature have explored better estimations of future implied volatility using encompassing regressions on index options. In this paper, we have the advantage of testing forecasting power of on the cross-section of extensive set of equity options. We combine the approach of Chan, Jha, Kalimipallli (2009) and Cao, Yu, and Zhong (2010) by (1) testing on an extensive panel of cross-sectional firms and options data, (2) forecasting the future implied volatility, and (3) finding how incorporating for CDS in the time series model can improve the forecasting performance of the implied volatility of the underlying asset, particularly out-of-sample.

In other words, we use credit risk to forecast out-of-sample option prices. Credit risk matters for option pricing as options are valued on firms which are not free from default. So we ask that if credit risk matters for option prices, can we develop better out-of-sample forecasts for IV using lagged credit-risk measures?
Previous literature is mostly focused on predicting IVs at aggregate or index level. We, however, provide a characterization of IVs forecasts at individual firm level and use a robust measure of credit risk i.e. CDS spreads of the underlying firm written for senior sub-ordinated debt.

We provide a first comprehensive study of role of credit risk on IV forecasting using an exhaustive sample of cross-sectional firms.

2 Literature Review

Previous literature on forecasting IV falls broadly under four main categories: (1) Modeling IV evolution at the index level (SPX: options on S&P 500 index), where a time series model for the changes in IV is used. (2) Modeling moments of risk neutral distribution (RND), where variance, skewness and kurtosis of the RNDs are modeled through time. (3) Encompassing regressions, where past historical volatility, and IVs are used to forecast realized volatility (RV) or option IV. (4) Implied volatility functions, where cross-sectional IV surface is defined as a polynomial function of moneyness and maturity to capture the pricing biases and option prices are obtained using numerical methods.

2.1 Modeling IV Evolution at the Index Level

Modeling volatility indices is an important part of IV predictability studies. Konstantinidi, Skiadopoulos, and Tzagkaraki (2008) investigate whether the
evolution of implied volatility can be forecasted by focusing on a number of European and US implied volatility indices. They use an economic model that estimates the change in lagged IV based on a number of lagged default risk and firm level variables. They test for both point and interval forecasts and study the statistical and economic significance of these forecasts. They further use principal component, ARIMA, ARFIMA, and VAR models and assess the performance by trading strategies in the volatility futures markets. Although they find statistical support for predictability, they show that the findings are not economically significant.

In our paper, we focus on cross section of the firms and are interested to be able to forecast IV for each option ID. But how do we test the forecast performance of the results?

Goncalves and Guidolin (2006) provide a clear framework for calculating estimation for options with the specification in mind that any averaging should be performed on count of traded options only. We use their approach in designing MSE, RMSE, and MAE.\footnote{MSE stands for Mean Square Error, RMSE for Root Mean Square Error, and MAE for Mean Absolute Error. The definitions of these error terms will be discussed in the regression section.} Motivated by empirical evidence on lack of stability of the parameters characterizing the implied volatility surface (IVS) in option prices, Goncalves and Guidolin (2006) investigate the predictability patterns of IVS. They employ a two-stage model that first estimates the surface along cross-sectional moneyness and maturity dimensions. Then in the second stage they model the dynamics of the first-stage coeffi-
cient. They find that the movements of the S&P 500 IVS are in fact very predictable. As a result, a set of profitable delta-hedged trading rules can be created, however the spreads will disappear due to higher transaction costs and on a wider sample of IVs.

Although performed at the index level, Goncalves and Guidolin’s paper is particularly interesting as the performance measures introduced in the paper are able to identify the predictability power of IV forecasts across different models. We employ the same measurement errors in comparing forecast performance of our models and show impact of CDS inclusion on the wide cross section of options.

2.2 Modeling Moments of Risk-neutral Distribution (RND)

A parametric approach for unveiling implied volatility is to model the moments of the underlying RND. This can be done both on the cross-section and the index level. By focusing on data extracted from the market prices of Standard & Poor’s (S&P) 500 index options, Neumann and Skiadopoulos (2013) investigate whether there are predictable patterns in the dynamics of higher-order risk-neutral moments (RNMs). They conduct a horse race among alternative forecasting models within an out-of-sample context over various forecasting horizons (daily, weekly, monthly). They find that higher RNMs can be statistically forecasted. However, only the 1-day-ahead skew-
ness forecasts can be economically exploited since the economic significance vanishes once the transaction costs is incorporated.

Comparing the index options vs. individual equity options, and in order to explain the risk-neutral skewness implied from option prices, the findings of Dennis and Mayhew (2002) empirically establish a link between the risk neutral skewness and the systematic risk of the underlying stock. They explain the structural difference in distributions by investigating the relative importance of several firm characteristics such as implied volatility, firm size, trading volume, leverage, and beta. They show that risk-neutral skewness tends to be more negative for stocks with larger betas. This is an evidence for the importance of market risk in option pricing. Their findings show that the index options have a more pronounced volatility smile/smirk than individual equity options.

If the moments are successfully modeled, this means the dynamics of implied volatility surfaces can also benefit from the modeling.

In the same line of parametric approach to index level forecast, Panigirtzoglou and Skiadopoulos (2004) try to model the dynamics of implied distributions by obtaining a parsimonious description of the dynamics of the S&P 500 implied cumulative distribution functions by applying PCA\(^3\). After identifying the factors, they employ arbitrage-free Monte Carlo simulation methods that model the evolution of the whole distribution as a diffusion process. Traditionally, modeling only the first two moments as diffusion

---

\(^3\)Principal Component Analysis
processes is deemed to be sufficient to span the whole IV. Their findings have important implications for “smile-consistent” option pricing and for risk management, and they also test their model out-of-sample.

2.3 Encompassing Regressions

Encompassing regressions are an additional line of studies with the goal of better modeling of IV and RV. In their papers Christensen and Prabhala (1998) model the relationship between IV and RV. Chan, Jha, and Kalimipalli (2009) and Becker, Clements, and White (2007), both use the S&P 500 implied volatility index (VIX) when running encompassing regressions. VIX modeling through encompassing regressions shows better out-of-sample performance. Our paper’s methods are closely related to Chan, Jha, and Kalimipalli (2009), in that we also use the encompassing regressions to model IV.

However, an important contribution of this paper is that we use a wide cross section of individual options, as opposed to limiting the tests to the options index.

2.4 Modeling through Implied Volatility Functions and IV Surface

Implied volatility functions are deterministic approach to modeling IV. Derman and Kani (1994), Dupire (1994), and Rubinstein (1994) hypothesize that
asset return volatility is a deterministic function of asset price and time, and develop a deterministic volatility function (DVF) option valuation model that has the potential of fitting the observed cross section of option prices exactly. Dumas, Fleming, and Whaley (1998) use S&P 500 options from to examine the predictive and hedging performance of the DVF option valuation model and find it that is no better than an ad hoc procedure that merely smooths Black–Scholes (1973) implied volatilities across exercise prices and times to expiration.

Although deterministic volatility models allow for more flexible volatility surfaces, these models refrain from introducing additional risk factors. This means we need stochastic models to introduce additional risk factors, and options are then needed for spanning of the pricing kernel. Buraschi and Jackwerth (2001) develop a statistical test based on this difference in spanning. Again, they use index level options data and show that both in- and out-of-the-money options are needed for spanning which supports stochastic volatility, interest rates, or jumps models.

Throughout the four different lines of literature discussed, it is shown that most forecasting attempts of IV have been empirically tested on index level data or a limited selection of underlying firms, and not the cross section of options. In our paper, we test for this by testing on an extensive collection of options data. All tests will be run in a time series pattern and then averaged cross-sectionally.
3 Data and Summary Statistics

To perform this analysis, we employ cross-section and time series observations from the following databases: Option Metrics, Markit data (CDS), and CRSP.

The credit default swap data is available from January 2002. As such, we collect matching data from January 2002 to December 2009, in order to cover the matching time period with the CDS data.

Markit database reports the CDS spread on the 1-year, 5-year and 10-year contracts. For the purpose of our research and in order to be dealing with the highest liquidity, we limit the data to 5-year spreads only.

Table 1 displays the Summary statistics of the complete dataset used for the forecasting. The three variables of interest are historical volatility, CDS, and IV since we are focused on time-series forecast of implied volatility. Panels A and B show the characteristics of these three variables across moneyness and maturity bins as well as through time. Based on the distribution of observations, we can see that across all moneyness groups, most liquid bins lie under short maturity, and across all maturity groups, most frequent observations belong to at-the-money and out-of-the money put options.
Panel C of Table 1 shows that during crisis and distress periods, as expected, IV increases to high of 51% on average with the median of 45%.

Are lagged variables significant cross-sectionally? Given the purpose of this paper is the time series analysis, we shall soon start building and performing tests on a time series basis. However, in order to highlight the significance of the variables chosen for the model, we run one set of univariate pooled panel regressions on the three key lagged variables. Table 2 shows the result of univariate regression of lagged variables on implied volatility. As indicated all three lagged variables of $CDS_{t-1}$, $HV_{t-1}$, and $IV_{t-1}$ are significant variables for the cross section of implied volatility. Note that the R-squared of the regressions are not very high due to simplicity of the model at this stage. Future research will focus on improving the liquidity as well as variables.

In the next section we design and test the time series forecast models.

4 Methodology

In this section we explain the methodology used, mainly the design of the forecasting model of future implied volatility.

---

4 Lagged CDS is the value of CDS for one period before for the same firm, i.e. last trading day’s value. Note that this results in many missing values that have to be dropped when the time series regressions run. No interpolation is performed for the lagged variables. Same logic applies to lagged IV value. See Cao and Yu (2010) for a similar data approach for time series forecast.
4.1 Volatility Measures Background

By definition volatility aims to capture the strength of the (unexpected) return variation over a given period of time. However, two distinct features importantly differentiate the construction of all (reasonable) volatility measures. First, given a set of actual return observations, how is the realized volatility computed? This means that the emphasis is explicitly on “ex-post” measurement of the volatility. Second, decision making processes often require forecasts of future return volatility. The focus this way is on “ex-ante” expected volatility.

The ex-ante concept requires a model that can effectively map the current information set into a volatility forecast. In contrast, the (ex-post) realized volatility may be computed (or approximated) without reference to any specific model, thus rendering the task of volatility measurement essentially a nonparametric procedure. An example is a rolling historical volatility measure.

In addition to these model classes, the implied volatility approaches are also significantly covered in the literature. The implied volatilities are based on a parametric volatility model for the returns, as defined above, along with an asset pricing model and an augmented information set consisting of options prices and/or term structure variables.
Various design of IV and HV have been used in conjunction with default risk studies. Campbell and Taksler (2003) used 180-day rolling standard deviation of equity return as an explanatory variable for the credit spread. Collin-Dufresne, Goldstein and Martin (2001) suggest the stock option implied volatility to be one of the important determinants of the changes in CDS spreads. We chose to use the 120 days rolling historical volatility because it corresponds to the maturity of the put options observed in our sample.

In our research we retrieve IV from the Optionmetrics database, and hence it is important to explain the estimation method employed. The calculated interpolated implied volatility for each option on each day, uses a methodology based on a kernel smoothing algorithm. The data is first organized by the log of days to expiration and by “call-equivalent delta” (delta for a call, one plus delta for a put). A kernel smoother is then used to generate a smoothed volatility value at each of the specified interpolation grid points. At each grid point on the volatility surface, the smoothed volatility is calculated as a weighted sum of option implied volatilities.\(^5\)

The implied volatility can also be seen as the volatility prognosis given the information available at the current state and hence it will be state-

\(^5\)See OptionMetrics Manual for additional details. Also, limited additional details provided in Appendix.
sensitive. Nonetheless the estimation of the implied volatility by CDS can be complicated due to the lack of data on CDS trades on a consistent daily basis. We do deal with missing variables which is due to the illiquidity of CDS compared to IV observations. Currently, interpolation is done for a single day missing (with trade data available prior and after). Future research can benefit from the augmented dataset (time period) and advanced interpolation (two or more consecutive days missing) in order to deal with this issue.

4.2 Out-of-sample Evaluation background

What is the purpose of out-of-sample forecasts? Cross-validation is the process of assessing how the results of a statistical analysis will generalize to an independent data set. Note that IV is a reparamaterized value. If the model has been estimated over some, but not all, of the available data, then the model using the estimated parameters can be used to predict the held-back data. We test the performance by measuring the out-of-sample mean squared error (MSE), also known as the mean squared prediction error, in order to test the strength of the model in forecasting the desired variable. In addition, we measure the mean absolute error (MAE) and compare these statistics for each forecast model.

With the data and measures description above, we continue to build the baseline regression model of forecasting implied volatility in the next section.
5 Regression Model Specification and Empirical Results

Based on the descriptions above, we can now specify the regression model to test our hypotheses. We split the analysis into two sets of contemporaneous and lagged models, as well as out-of-sample performance models.

Because the average maturity for the options used in our implied volatility estimation is about four months, both historical volatility and future realized volatility are computed over 120 trading days in this exercise in order to match the horizon of option-implied volatility. We use daily data for the regression with the Newey and West (1987) correction to the standard errors for autocorrelation and heteroscedasticity.

We report two tests for each of contemporaneous and forward looking in-sample models, one “with” and one “without” inclusion of the CDS measure. The results can be found in Tables 3 and 4. We first explain the model and then elaborate on the findings of these two tables.

The contemporaneous regression specification, hence, can be explained per below:

\[ IV_t = \beta_0 + \beta_1 HV_t + \epsilon_t \]  

(1)
Table 3 shows the results of our findings. The specification above is run on each option ID and then averaged through all option IDs and reported. HV tend to be an insignificant factor in determining the contemporaneous implied volatility. CDS, on the other hand, shows to be statistically and economically significant, and the inclusion improves the $R^2$ of the forecast model to 27.6%.

**On Forecasting Levels vs. Changes:** We perform level regressions and not changes. This is mostly because the irregular intervals or missing data for CDS observations limits the design of any change regressions. From a statistical perspective, first differencing is appropriate if the dependent variable and regressors are integrated, but this is difficult to determine for our irregularly spaced data. The potential extension of the data set and additional interpolation techniques will allow us to explore the change forecast in future studies.

We then move on to in-sample forecasts with 1-day ahead. Table 4 shows the results for this forecast.

$$IV_t = \beta_0 + \beta_1 HV_{t-1} + \beta_2 IV_{t-1} + \epsilon_t$$  \hspace{1cm} (3)
where $IV_t$ is the daily volatility on day $t$, $HV_t$ is the historical volatility on day $t$, and $CDS_t$ is the 5-year observed CDS spread on day $t$.

We estimate each of the above models for the sample period. (To report the results, we run this time series regression on each cross-sectional firm and then report the average betas, the cross-sectional mean of the regression coefficients, and the Newey-West adjusted t-stats, as well average R-squared values.)

To test multiple horizons, we generate multistep forecasts on a given day for each option. We examine 1-day-ahead (accounting for daily forecast), and 5-day-ahead (accounting for weekly forecast). The results show that CDS is a significant estimator of the forecasted $IV$.

In the next section we investigate out-of-sample performance of the designed forecast models.

### 5.1 Out-of-sample Tests

The reasoning for looking at the out-of-sample forecasting performance in addition to the in-sample fit comes from the objective of the analysis. In forecasting it is not necessarily the model that provides the best in-sample fit that produces the best out-of-sample volatility forecast, which is the main objective of the analysis\footnote{Shamiri and Isa (2009).}. Hence it is common to use the out-of-sample forecast to aid the selection of which model is best suited for the series under
study. The out-of-sample forecast refers to that the data used in the fitting of the model is different from the data used for evaluating the forecasts. Typically the data in divided into two subsets, one in which the model parameters are fitted (estimation subsample), and another subset used to evaluate the forecasting performance of the models (forecasting subsample).

In Tables 3 and 4 we tested in-sample performance. Tables 5 and 6 investigate out-of-sample performance for two different step sizes: Table 5 shows the result for \( k = 1 \), which corresponds to daily forecast and Table 6 shows the results for weekly forecast, i.e. \( k = 5 \).

To calculate the out-of-sample results we estimate each model based on a rolling window on the sample period, and based on the number of days to maturity, then we generate multistep out-of-sample forecasts for each day. To feed sufficient data points, out-of-sample is not performed on the first year of the data. In summary, we can explain the process as below:

- Consider an option with a specific option ID (that belongs to a moneyness and maturity group). We derive the observed IV for each day from OptionMetrics from the in-sample period.

- We use the underlying firm’s 5 year CDS spread, and the recorded historical volatility for each day. (Note that lagged CDS only exists if there exists a trade observation for the previous period. Otherwise, need to drop the observation)

---

• We split the lifetime of the observed option into two periods: estimation subsample (time= 0 to \( t \)) and forecasting subsample (\( t \) to \( T \)).

• Using the fitted values until time \( t \), we then forecast IV for \( T - t \).

• The reported errors measure performance on the forecasting subsample only and not the fitted values in the estimation subsample.

For both models we calculate three error values: MSE, RMSE, and MAE values across all option IDs. Based on comparison of all three measures across bins, we can see that: First, the model with CDS enhancement outperforms the rest of the forecast models since the error values are lowest. Observing across each of the bins, we can see that overall the forecasting error drops when the model is enhanced by CDS (i.e. last row). In addition, for the complete sample, (including all maturity and moneyness groups in one sample), the RMSE drops by 4 basis points between the first and last models.

Second, we can also conclude that Long option bins in general have the best out-of-sample performance of IV forecast. The forecasting error for these bins (Long ITM, Long ATM, and Long OTM), are consistently measured lowest in their groups.

\(^8\) (i) The root mean squared prediction error in implied volatilities (RMSE) is the square root of the average squared deviations of BS implied volatilities from the model’s forecast implied volatilities, averaged over the number of options traded. (ii) The mean absolute prediction error in implied volatilities (MAE) is the average of the absolute differences between the BS implied volatility and the model’s forecast implied volatility across traded options.
Overall, inclusion of CDS can decrease the out-of-sample forecasting measurement error. Next, we investigate whether the predictability power of CDS acts differently during crisis periods and across industry.

5.2 Impact of Crisis and Industry

We further like to compare the out-of-sample performance of IV forecasting with the enhanced CDS, on the key sub-samples. Table 7 investigates impact of financial crisis on model’s out-of-sample performance. The results show that the estimation error is consistently lower for non-crisis periods. The test of differences in mean residual of the two subsamples are also statistically significant. This shows that although we can improve the out-of-sample performance of IV forecast, our progress is somewhat restricted during crisis periods.

Finally, we like to see whether financial firms do better in out-of-sample forecasting or not. Following the same approach as above we can test this and see that non-financial and financial firms do not show any “economically” significant difference in out-of-sample performance of their IVs, despite the statistically significant results in test of difference in mean. Table 8 shows that the values are very similar and the difference is economically minor.
6 Conclusion

Extant literature show that credit risk impacts option pricing. This association is an interesting component that can further be explored in the time series context. Specifically, we can use the forward-looking information content of CDS in forecasting option Implied Volatility.

In addition, forecasting option Implied Volatility (IV) is of interest to option market participants, who routinely formulate volatility and option price forecasts for trading and hedging purposes.

Previous literature is mostly about predicting IVs at aggregate or index level. In this paper, however, we provide a characterization of IVs forecasts at individual firm level. We use a robust measure of credit risk i.e. CDS spreads of the underlying firm written for senior sub-ordinated debt and therefore provide a first comprehensive study of role of credit risk on IV forecasting using an exhaustive sample of 550 firms for the 2002-2009 period.

Inclusion of CDS improves the precision of implied volatility forecast, specifically for out-of-sample. We further show that the forecasting error of implied volatility will vary across moneyness and maturity indicating deep OTM and deep ITM options can be subject to higher pricing error. Longer maturity options have lowest estimation errors and show that
7 References


Konstantinidi, Eirini, George Skiadopoulos, and Emilia Tzagkaraki. “Can


Tang, Dragon Yongjun, and Hong Yan. “What moves CDS spreads.” Working paper, University of Hong Kong, University of South Carolina and Shanghai Advanced Institute of Finance, 2011.

8 Appendices

8.1 Appendix A: Complete Variables Glossary

In this section, the detailed definitions for each of the variables used in the baseline regression(s) are provided.

• **CDS spread**: 5-year maturity CDS spread, daily observations from Markit. When data is missing for one day in between two observations, it gets interpolated. Any consecutive missing for more than one day, is treated as missing value and automatically is dropped from the panel regressions.

• **Option-implied volatility (IV)**: Put option IV, daily observations, directly from Option Metrics. ⁹

• **CDS Liquidity Proxy**: equals to \((\text{No.ofContributors} - \text{AvgNo.ofContributors}) \times CDS\text{spread}\), or “Demeaned number of contributors times CDS spread”. The number of contributors are retrieved directly from Markit.

• **Firm Return (%)**: For each underlying firm daily price is retrieved from CRSP database. \(\text{FirmRet} = \frac{P_1 - P_0}{P_0}\).

• **Historical Volatility (%)**: Standard deviation of a fund’s daily returns over a rolling 6-month period.

• **Market to Book Value**: Tobin’s Q calculated with data from Compustat quarterly files.

• **Debt ratio**: Total Debt/Total Asset; data from Compustat quarterly files; defined as total liabilities divided by the sum of total liabilities and total assets.

---

⁹IVs in OptionMetrics are retrieved from the Volatility Surface data file. The calculated interpolated implied volatility for each option on each day, uses a methodology based on a kernel smoothing algorithm. The data is first organized by the log of days to expiration and by “call-equivalent delta” (delta for a call, one plus delta for a put). A kernel smoother is then used to generate a smoothed volatility value at each of the specified interpolation grid points. At each grid point on the volatility surface, the smoothed volatility is calculated as a weighted sum of option implied volatilities. (See OptionMetrics Manual for additional details.)
market capitalization.

- **VIX**: Daily market volatility, \( VOLATILITY_{S&P500}^{(VIX)} \), retrieved from CRSP, and available on OptionMetrics.

- **TED Spread**: The difference between the interest rates on interbank loans and on short-term U.S. government debt (T-bills). It is calculated as the difference between the three-month LIBOR and the three-month T-bill interest rate. Data files for each variable retrieved from Datastream.

- **IV Skew**: As formulated in the paper, IV skew is defined for each observation as “Avg OTM implied volatility – Avg ATM implied volatility”:

\[
IVSkew_{i,t} = \text{AVERAGE}_{OTM}(IV_{i,t}) - \text{AVERAGE}_{ATM}(IV_{i,t})
\]
8.2 Merging CDS Markit and Option Metrics

There are total of 1833 Ticker/Comp Name combination in the Markit Database. Out of these 1833, 34% are private or subsidiaries (See appendix A on this cleanup activity). Removing these, would leave a remainder of 1203 public to work with. The next step is to limit data to firms with Options. Matching the 1203 with Options Metric data, we are down to 1063 which both have CDS written on the underlying bond, and options written on their underlying company. The sample that we work with would have 1063 unique firm IDs. (see chart below)

8.3 Cleaning up Option Metrics Data

I downloaded all Option Metrics data from 1996 to 2011 inclusive. The Option Prices file (under Option Metrics Data files\Options) gets downloaded with all possible variables (32 variables in total) in monthly files format from 1996 until 2010. (The average volume for every month is 1-2GB)

There are 32 listed variables downloaded, per below:

**FILTERS:**
1) DROP UNNECESSARY VARS \( (ssflag,indexflag,exchanged,issuetype) \)
   keep if \( cpflag == "C" \) \| \( cpflag == "P" \) keep if \( ssflag == 0 \) (121656 observations deleted)
   drop if \( exchanged == 32768 \) (0 observations deleted) keep if \( issuetyper == "0" \) (1681856 observations deleted)
2)prior to saving the final file, I ”sort secid optionid date” to organize the data; it makes it much easier in the data matching, & give the compress command one more time before saving the datafile.
3)LITERATURE FILTERS: Many in our sample do not have actively traded options. The choice of non-zero open interest emphasizes the information content of options that are currently in use by market participants. We also drop all zero volume options. I also exclude all options that violate Put-Call parity. For the first round of the analysis, I keep only put options.
8.4 Compustat Fundamentals Data Collection

I collect and compute the quarterly Company Variables (i.e. Size, Book/Market, and Leverage) for all of the in the sample. I also winsorize data and store statistics both before and after winsorization.

For merging with Option Metrics and CDS, I keep the pre-winsorization file.

8.5 Report on CDS Markit Data Clean-up for TICKERS & COMP NAMES

Markit Look-Up Dimensions: TICKER, Company Name, Date Period(s)

Issues:
1. Tickers may get recycled and as such multiple matches available when merged with other databases
2. Company names are in short form or miss spaces/dots in Markit so cannot automatically be matched. Need manual review.

Solution:
Review the complete Ticker/CompName list from Markit and verify their identification (i.e. match with the unique PERMNO from CRSP).
- Use Google/Yahoo Finance search engines.
- Confirm Private/Public Companies
- Find alternative names and tickers

Manual Correction Steps
1. Look up the ticker in CRSP full database
2. If the ticker and company name match fully, there should be one unique PERMNO. Record any alternative names.
3. If there are alternatives, provide correct match for the time period provided in Markit.
4. If there are no matches, use Google and Yahoo finance searches. – Is the company a Private company?

Results Statistics

There are total of 1833 unique Ticker/Comp Name combination in the
CDS Markit Database. The final correction and inclusion provided final match of 1203 companies in total.

From the remaining, the provided correction notes indicate that:

a. 15 firms are subsidiaries with parent company not found before, so parent ticker is provided.

b. Total of 100 unmatched tickers are Private firms.

c. Total of 340 unmatched tickers are Subsidiaries with their parents present in the database.

d. Fewer than 10 firms are result of mergers or acquisitions that are still valid for inclusion. Details provided.

This is the result of individual scrubbing of each firm/ticker.
Figure 1. Time-series comovement of CDS spread and implied volatility

This chart shows the IV and CDS dynamics for Lehman Brothers (LEH) and Merrill Lynch (MER) over the Jan2002-Jan2009 sample. Consistent with the hypotheses of the paper on the forecasting power of CDS, we can see the comovement of CDS and implied volatility. Our goal here is to trace the time-series relationship of CDS in estimating future implied volatility and use it in the IV forecast model.
Table 1. Summary Statistics
Panel A reports the overall summary statistics of the time-series means of 550 sample firms for each of the key variables. Panel B reports the summary statistics across bins. CDS Spread is the five-year composite credit default swap spread; historical volatility is the 252-day historical volatility; implied volatility is the volatility inferred from put options with nonzero open interests; The sample period extends from January 2002 through 2009.

### Panel A: All data

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS Spread</td>
<td>1.10</td>
<td>0.24</td>
<td>0.47</td>
<td>1.09</td>
<td>2.17</td>
</tr>
<tr>
<td>Historical Volatility</td>
<td>0.38</td>
<td>0.03</td>
<td>0.25</td>
<td>0.58</td>
<td>0.05</td>
</tr>
<tr>
<td>Implied Volatility</td>
<td>0.36</td>
<td>0.26</td>
<td>0.33</td>
<td>0.43</td>
<td>0.17</td>
</tr>
</tbody>
</table>

| Nobs                     | 205441 |

### Panel B: Across Bins

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS Spread</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical Volatility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implied Volatility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Obs.</td>
<td>11.45%</td>
<td>7.00%</td>
<td>5.51%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ITM</th>
<th>ATM</th>
<th>OTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS Spread</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical Volatility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implied Volatility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Obs.</td>
<td>18.75%</td>
<td>10.12%</td>
<td>6.83%</td>
</tr>
</tbody>
</table>

|                          |       |       |       |
| CDS Spread               |       |       |       |
| Historical Volatility    |       |       |       |
| Implied Volatility       |       |       |       |
| Percent Obs.              | 17.55%| 12.96%| 9.83% |

### Panel C: Implied volatility across years

<table>
<thead>
<tr>
<th></th>
<th>IV Mean</th>
<th>IV Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0.4707201</td>
<td>0.428546</td>
</tr>
<tr>
<td>2003</td>
<td>0.3727789</td>
<td>0.348667</td>
</tr>
<tr>
<td>2004</td>
<td>0.3205156</td>
<td>0.3027075</td>
</tr>
<tr>
<td>2005</td>
<td>0.2985788</td>
<td>0.282648</td>
</tr>
<tr>
<td>2006</td>
<td>0.3101929</td>
<td>0.297454</td>
</tr>
<tr>
<td>2007</td>
<td>0.3316317</td>
<td>0.312099</td>
</tr>
<tr>
<td>2008</td>
<td>0.5067201</td>
<td>0.450139</td>
</tr>
</tbody>
</table>
Table 2. Pooled Regression of Implied volatility

The table presents univariate pooled regression of Implied volatility on the set of key lagged variables to be employed in the forecasting model. The pooled regression is aimed to highlight the comovement of each with the implied volatility.

*Lagged CDS = previous observation day’s CDS for the issuer firm.*  
*Lagged IV = previous observation day’s IV for the same Option ID.*  
*Lagged HV = previous observation day’s HV for the issuer firm.*

<table>
<thead>
<tr>
<th>Dependent Var: IV</th>
<th>(1) Lagged IV</th>
<th>(2) Lagged HV</th>
<th>(3) Lagged CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>0.0118***</td>
<td>0.0640***</td>
<td>0.354***</td>
</tr>
<tr>
<td>Lagged CDS</td>
<td>(21.08)</td>
<td>(7.729)</td>
<td>(75.40)</td>
</tr>
<tr>
<td>Lagged HV</td>
<td>0.327***</td>
<td>0.325***</td>
<td>0.232***</td>
</tr>
<tr>
<td>Lagged IV</td>
<td>(585.9)</td>
<td>(510.0)</td>
<td>(142.4)</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.327***</td>
<td>0.325***</td>
<td>0.232***</td>
</tr>
<tr>
<td></td>
<td>(585.9)</td>
<td>(510.0)</td>
<td>(142.4)</td>
</tr>
<tr>
<td>Observations</td>
<td>76,258</td>
<td>117,519</td>
<td>105,602</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.010</td>
<td>0.002</td>
<td>0.089</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1  
t-statistics in parentheses
Table 3. Contemporaneous regression of implied volatility (run on each option)

This table estimates the following three models of variations with and without CDS:

\[ IV_t = \beta_0 + \beta_1 \times HV_t \]

\[ IV_t = \beta_0 + \beta_2 \times CDS_t \]

\[ IV_t = \beta_0 + \beta_1 \times HV_t + \beta_2 \times CDS_t \]

Where CDS is the observed credit default swap spread, HV is the rolling historical volatility, and IV is the option implied volatility.

We run this time series regression on each cross-sectional firm and then report the average betas. Both cross-sectional mean of the regression coefficients, the (Newey-West adjusted) t-stats, as well as standard errors and average R-squared are presented.

(*Italic values show the average t-stat of cross-sectional values.)

<table>
<thead>
<tr>
<th></th>
<th>HV without CDS</th>
<th>CDS without HV</th>
<th>HV and CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Beta_0 (80)</td>
<td>0.3347***</td>
<td>8.1225***</td>
<td>0.4293***</td>
</tr>
<tr>
<td></td>
<td>30.27</td>
<td>29.09</td>
<td>14.09</td>
</tr>
<tr>
<td>Average Beta_1 (81)</td>
<td>0.0449258</td>
<td></td>
<td>2.577973</td>
</tr>
<tr>
<td></td>
<td>0.51</td>
<td></td>
<td>.6215886</td>
</tr>
<tr>
<td>Average Beta_2 (82)</td>
<td></td>
<td>3.5876***</td>
<td>4.8797***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.43</td>
<td>2.69</td>
</tr>
</tbody>
</table>

| N                    | 13429          | 33081          | 9921       |
| Average R-squared    | 8.51%          | 18.28%         | 27.62%     |
Table 4. Time-series regression of implied volatility (run on each option)

This table estimates the forecasting power of the CDS spreads when augmented to the IV forecasting model. Both regression coefficients and Newey-West adjusted t-stats are presented.

This table estimates the following two models with and without CDS:

\[ IV_t = \beta_0 + \beta_1 \times HV_{t-1} + \beta_2 \times IV_{t-1} \]

\[ IV_t = \beta_0 + \beta_1 \times HV_{t-1} + \beta_2 \times IV_{t-1} + \beta_3 \times CDS_{t-1} \]

Where HV is the rolling historical volatility, and IV is the option implied volatility. CDS is the 5-year maturity credit default swap for the underlying firm. We run this time series regression on each cross-sectional firm and then report the average betas. Both cross-sectional mean of the regression coefficients, the (Newey-West adjusted) t-stats, as well as average R-squared are presented.

<table>
<thead>
<tr>
<th></th>
<th>Without CDS</th>
<th>With CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Beta_0 (β0)</strong></td>
<td>-0.0256***</td>
<td>-0.0873***</td>
</tr>
<tr>
<td></td>
<td>-4.20</td>
<td>-5.25</td>
</tr>
<tr>
<td><strong>Average Beta_1 (β1)</strong></td>
<td>0.1092***</td>
<td>0.0561***</td>
</tr>
<tr>
<td></td>
<td>12.29</td>
<td>11.50</td>
</tr>
<tr>
<td><strong>Average Beta_2 (β2)</strong></td>
<td>0.3289***</td>
<td>0.0750***</td>
</tr>
<tr>
<td></td>
<td>14.38</td>
<td>10.55</td>
</tr>
<tr>
<td><strong>Average Beta_3 (β3)</strong></td>
<td></td>
<td>0.0027***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.30</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>8379</td>
<td>6284</td>
</tr>
<tr>
<td><strong>Average R2</strong></td>
<td>0.21</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Table 5. Daily Out-of-sample performance of the model specifications for each of the maturity and moneyness bins (k= 1 day)
Out-of-sample performance of the model specifications for each one of the implied volatility forecast. We do this estimation for all three models, and calculate the errors for nine different bins as well as all sample. The mean square error (MSE), the root mean squared prediction error (RMSE), and the mean absolute prediction error (MAE) are provided.

(i) The root mean squared prediction error in implied volatilities (RMSE) is the square root of the average squared deviations of BS implied volatilities from the model’s forecast implied volatilities, averaged over the number of options traded. (ii) The mean absolute prediction error in implied volatilities (MAE) is the average of the absolute differences between the BS implied volatility and the model’s forecast implied volatility across traded options.

Table 5. Out-of-sample performance of IV prediction: k= 1 day

<table>
<thead>
<tr>
<th></th>
<th>ITM</th>
<th>ATM</th>
<th>OTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All sample</td>
<td>Short</td>
<td>Medium</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0182</td>
<td>0.0316</td>
<td>0.0202</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1349</td>
<td>0.1779</td>
<td>0.1420</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0182</td>
<td>0.0316</td>
<td>0.0202</td>
</tr>
</tbody>
</table>

Panel A: HV\(_{t-1}\)

MSE 0.0201 0.0364 0.0215 0.0122 0.0200 0.0162 0.0117 0.0285 0.0163 0.0099
RMSE 0.1419 0.1907 0.1467 0.1105 0.1414 0.1274 0.1081 0.1688 0.1277 0.0995
MAE 0.0972 0.1277 0.0999 0.0821 0.0994 0.0913 0.0778 0.1174 0.0871 0.0710

Panel B: CDS\(_{t-1}\)

MSE 0.0137 0.0250 0.0166 0.0076 0.0140 0.0116 0.0062 0.0200 0.0122 0.0065
RMSE 0.1171 0.1582 0.1288 0.0870 0.1185 0.1077 0.0789 0.1415 0.1103 0.0803
MAE 0.0810 0.1103 0.0878 0.0653 0.0841 0.0758 0.0606 0.1011 0.0742 0.0589

Panel C: CDS\(_{t-1}\) and HV\(_{t-1}\)

The root mean squared prediction error in implied volatilities (RMSE) is the square root of the average squared deviations of BS implied volatilities from the model’s forecast implied volatilities, averaged over the number of options traded. (ii) The mean absolute prediction error in implied volatilities (MAE) is the average of the absolute differences between the BS implied volatility and the model’s forecast implied volatility across traded options.
Table 6. Weekly Out-of-sample performance of the model specifications for each of the maturity and moneyness bins (k=5 days)

Out-of-sample performance of the model specifications for each one of the implied volatility forecast. We do this estimation for all three models, and calculate the errors for nine different bins as well as all sample. The mean square error (MSE), the root mean squared prediction error (RMSE), and the mean absolute prediction error (MAE) are provided.

(i) The root mean squared prediction error in implied volatilities (RMSE) is the square root of the average squared deviations of BS implied volatilities from the model’s forecast implied volatilities, averaged over the number of options traded. (ii) The mean absolute prediction error in implied volatilities (MAE) is the average of the absolute differences between the BS implied volatility and the model’s forecast implied volatility across traded options.

Table 6. Out-of-sample performance of IV prediction: k= 5 days

<table>
<thead>
<tr>
<th></th>
<th>ITM</th>
<th></th>
<th></th>
<th></th>
<th>ATM</th>
<th></th>
<th></th>
<th></th>
<th>OTM</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All sample</td>
<td>Short</td>
<td>Medium</td>
<td>Long</td>
<td>Short</td>
<td>Medium</td>
<td>Long</td>
<td>Short</td>
<td>Medium</td>
<td>Long</td>
<td>Short</td>
<td>Medium</td>
</tr>
<tr>
<td><strong>Panel A: HV_{t-1}</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>0.0122</td>
<td>0.0235</td>
<td>0.0195</td>
<td>0.0054</td>
<td>0.0208</td>
<td>0.0154</td>
<td>0.0045</td>
<td>0.0234</td>
<td>0.0135</td>
<td>0.0044</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1104</td>
<td>0.1532</td>
<td>0.1398</td>
<td>0.0732</td>
<td>0.1443</td>
<td>0.1241</td>
<td>0.0669</td>
<td>0.1529</td>
<td>0.1160</td>
<td>0.0662</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>0.0809</td>
<td>0.1215</td>
<td>0.1005</td>
<td>0.0615</td>
<td>0.1009</td>
<td>0.0914</td>
<td>0.0619</td>
<td>0.1160</td>
<td>0.0761</td>
<td>0.0600</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Panel B: CDS_{t-1}** | | | | | | | | | | | | | |
| MSE              | 0.0141 | 0.0242 | 0.0238 | 0.0073 | 0.0153 | 0.0148 | 0.0058 | 0.0204 | 0.0153 | 0.0052 | | | |
| RMSE             | 0.1187 | 0.1555 | 0.1543 | 0.0857 | 0.1237 | 0.1215 | 0.0762 | 0.1428 | 0.1238 | 0.0721 | | | |
| MAE              | 0.0822 | 0.1165 | 0.1019 | 0.0638 | 0.0905 | 0.0853 | 0.0605 | 0.1009 | 0.0806 | 0.0539 | | | |

| **Panel C: CDS_{t-1} and HV_{t-1}** | | | | | | | | | | | | | |
| MSE              | 0.0141 | 0.0255 | 0.0209 | 0.0064 | 0.0278 | 0.0171 | 0.0060 | 0.0258 | 0.0138 | 0.0062 | | | |
| RMSE             | 0.1186 | 0.1596 | 0.1445 | 0.0797 | 0.1668 | 0.1306 | 0.0778 | 0.1607 | 0.1176 | 0.0785 | | | |
| MAE              | 0.0752 | 0.1196 | 0.0935 | 0.0558 | 0.0971 | 0.0831 | 0.0527 | 0.1163 | 0.0745 | 0.0511 | | | |
Table 7. Impact of Financial Crisis: Out-of-sample performance during crisis vs. non-crisis periods

The table presents results of out-of-sample performance of Implied volatility forecast for two subperiods of Crisis (2007-2009), and Non-crisis (2002-2006) periods, on the set of key lagged variables. We show the **MAE (Mean Absolute Error)** for each of these regressions and also test for the difference between the two set of residuals.

The results show that during the crisis period the out of sample performance power consistently and significantly drops, as measured by a larger value of MAE.

Also, across all four models, the model with lagged CDS, lagged IV and lagged HV, possesses the smallest level of MAE.

*Lagged CDS = previous observation day’s CDS for the issuer firm.*
*Lagged IV = previous observation day’s IV for the same Option ID.*
*Lagged HV = previous observation day’s HV for the issuer firm.*

<table>
<thead>
<tr>
<th>Forecasted Var: IV</th>
<th>(1)</th>
<th>(2)</th>
<th>Test of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: $IV_t$ on ${IV_{t-1}}$</td>
<td>0.0356</td>
<td>0.0508</td>
<td>0.0152***</td>
</tr>
<tr>
<td>Model 2: $IV_t$ on ${IV_{t-1}, HV_{t-1}}$</td>
<td>0.0300</td>
<td>0.0422</td>
<td>0.0121***</td>
</tr>
<tr>
<td>Model 3: $IV_t$ on ${IV_{t-1}, CDS_{t-1}}$</td>
<td>0.0328</td>
<td>0.0487</td>
<td>0.0159***</td>
</tr>
<tr>
<td>Model 4: $IV_t$ on ${IV_{t-1}, HV_{t-1}, CDS_{t-1}}$</td>
<td>0.0286</td>
<td>0.0414</td>
<td>0.0129***</td>
</tr>
</tbody>
</table>

* t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The table presents results of out-of-sample performance of Implied volatility forecast for two subsections of financial and non-financial firms, on the set of key lagged variables. We show the MAE (Mean Absolute Error) for each of these regressions and also test for the difference between the two set of residuals.

The results show that the difference between MAEs are statistically significant for 3 out of 4 models; However, it is interesting to note that the economic magnitude of the difference is minor between financial and non-financial firms.

Consistent with prior results, across all four models, the forecasting model that includes lagged CDS, lagged IV and lagged HV, possesses the smallest level of MAE and outperforms the rest of the models.

Lagged CDS = previous observation day’s CDS for the issuer firm.
Lagged IV = previous observation day’s IV for the same Option ID.
Lagged HV = previous observation day’s HV for the issuer firm.

<table>
<thead>
<tr>
<th>Forecasted Var: IV</th>
<th>Non-Financial</th>
<th>Financial</th>
<th>Test of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: $IV_t$ on ${IV_{t-1}}$</td>
<td>0.0408</td>
<td>0.0460</td>
<td>0.0052***</td>
</tr>
<tr>
<td>Model 2: $IV_t$ on ${IV_{t-1}, HV_{t-1}}$</td>
<td>0.0348</td>
<td>0.0359</td>
<td>0.0009</td>
</tr>
<tr>
<td>Model 3: $IV_t$ on ${IV_{t-1}, CDS_{t-1}}$</td>
<td>0.0386</td>
<td>0.0431</td>
<td>0.0045***</td>
</tr>
<tr>
<td>Model 4: $IV_t$ on ${IV_{t-1}, HV_{t-1}, CDS_{t-1}}$</td>
<td>0.0335</td>
<td>0.0358</td>
<td>0.0022***</td>
</tr>
</tbody>
</table>

*t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1